Simultaneous Gene Selection and Cancer Classification using Chemical Reaction Optimization

Jitesh Doshi, Mahesh Chindhe, Yogesh Kharche, Shameek Ghosh, Jayaraman Valadi*

Abstract **- Microarray gene expression datasets comprise of a large number of genes in contrast to a small number of samples, thus having a high dimension of variables. Analysis of microarray data can lead us to many useful conclusions. In many microarray data analyses, selecting a small subset of genes which are of significance for a particular type of disease is an important issue but selection of such genes become difficult due to many irrelevant genes and noisy genes. The process of gene selection helps to extract the most informative genes, which consequently aid to build a robust prediction model using those genes. In this study, we employ a hybrid Chemical Reaction Optimization (CRO) based filter-wrapper methodology, which uses an information gain gene ranking heuristic to simultaneously extract informative gene subsets and build robust cancer classification models. The performance of the proposed method was tested on three benchmark gene expression datasets obtained from the Kent Ridge Biomedical datasets collection and the LIBSVM data repository. CRO results demonstrate its capability to select relevant genes with high confidence in comparison to the results reported earlier.**

Index terms **- Cancer Classification, Chemical Reaction Optimization, Gene Ranking Heuristics, Gene Selection.**

I. Introduction

MICROARRAY gene expression experiments provide the expression levels of thousands of genes. Computational analyses of such datasets present exciting opportunities for studying cancer and various genetic diseases. Gene expression datasets have a complex structure owing to lesser number of samples in comparison to an extremely large attribute/feature space having high noise levels, irrelevant attributes, missing values, several outliers and sample variance. Computational analysis of such data can help derive informative genes which may be responsible for a certain disease. The problem of gene selection thus concerns itself with, identifying a subset of relevant genes (or input variables), that can help in building a robust classification model for the target disease [1].

Several methods have been developed for the purpose of gene selection. Two important categories of gene selection methods are wrappers and filters [1]. Wrappers employ a learning algorithm to score the quality of gene subsets based on their predictive power (by internally calling a classification function). Bio-inspired algorithms like the Ant Colony Optimization and Genetic Algorithm in conjunction with a classifier like Support Vector Machines (SVM) may fall into this category [1-5]. On the other hand, filters rank genes based on their statistical properties with reference to the dataset. Methods like statistical tests and mutual information come under this category.

Using these two kinds of methodologies in a combined fashion should give us a better way of analyzing data and finding the solutions to our problem more accurately.

This study proposes the use of a hybrid gene-selection methodology employing a Chemical Reaction Optimization (CRO) based filter-wrapper approach. As part of a wrapper approach, CRO explores the gene-expression profile search space to iteratively obtain relevant and informative gene subsets, powered by a gene ranking heuristic, which consequently helps in the construction of robust classification models. CRO is recently proposed optimization technique based on behavior of molecules in a chemical reaction [6]. In present study, CRO has been used along with Support Vector Machine (SVM) and information gain attributes evaluation to extract a set of informative genes from a set of large number of genes [8, 15].

Modified CRO algorithm for analysis of gene expression data thus presents a useful tool to identify informative genes from a large pool of genes. This algorithm has been implemented in Java and a Java-based application has been written which will be available soon.

II. Materials and Methods

A. Chemical Reaction Optimization (CRO)

Chemical Reaction Optimization is a recently proposed meta-heuristic for optimization purposes which has been inspired by processes of molecular interactions in chemical reactions [6]. In a chemical reaction, the reactants form final products after passing through a number of intermediate stages. Throughout the process, the reactants transition to various forms in elementary chemical changes. The stable final products are achieved at the lowest potential energy. Thus, every reacting system seeks to achieve the minimum of free energy.

Manuscript received March 06, 2014; revised March 29, 2014. This work was supported in part by the Department of Science & Technology (DST), New Delhi, India

J. Doshi was with Bioinformatics Centre, University of Pune, Pune, India. He is now with Bioengineering dept., University of Illinois at Chicago, Chicago, IL, USA 60607 (email: jdoshi5@uic.edu).

M. Chindhe and Y. Kharche are with Dept. of Computer Science, University of Pune, Pune, Maharashtra 411007 India (emails: mahesh.chindhe@gmail.com, yogeshshreeram@gmail.com).

S. Ghosh. was with the Evolutionary Computing & Image Processing Group, Centre for Development of Advanced Computing, University of Pune, Pune, Maharashtra 411007 India. He is now with Advanced Analytics Institute, University of Technology, Sydney, Broadway NSW 2007 Australia (email: shameek09@gmail.com)

*J. Valadi is with the Evolutionary Computing & Image Processing Group, Centre for Development of Advanced Computing, University of Pune, Pune, Maharashtra 411007 India and with the Center for Informatics, Shiv Nadar University, Chithera, Dadri, Gautam Budh Nagar, Uttar Pradesh 203207 India. (phone: 91-120-2663801 extn. 117; e-mail: jayaraman.valadi@snu.edu.in).

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

CRO attempts to mimic this process where molecules may be modeled as solutions and can thus move through a certain number of intermediate transitions. The complete framework of CRO has been described extensively in [6-8]. The salient properties of a chemically reacting system are described briefly here. A molecule in CRO has some properties. These are the molecular structure (solution), potential energy (objective function measure), kinetic energy (measure of tolerance for accepting an inferior solution), number of hits (present total number of moves), minimum structure (present optimal solution), minimum value (present optimal function value), and minimum hit number (number of moves when the current optimal solution is found). If ω is denoted as the molecule and *f*, the objective function, then

$$PE_{\omega} = f(\omega)$$

Typically, a molecular transition is accepted if, $PE_{\omega} \geq PE_{\omega'}$ where ω may transform to ω'. If the given condition is not satisfied, then a change is allowed with $PE_{\omega} + KE_{\omega} \geq PE_{\omega'}$. KE_{ω} is the kinetic energy of the molecule, which is employed to facilitate the transition. KE is used as a tolerance measure to allow the transition of an existing molecule to a less favorable molecule. In this way, CRO allows the system to accept solutions, which can aid in escaping a local minimum. Typically, a central energy buffer is maintained in this context which stores an initial amount of energy. CRO may therefore allow a less favorable transition to occur by removing the required KE from the energy buffer. Thus throughout the process of CRO, molecules typically attempt to achieve a low *PE* . The various stages of CRO are described further.

In a chemical reaction, molecules may collide with each other or container walls. In this context, there may be four types of elementary chemical reactions molecules may go through, which are, *on-wall ineffective collision, decomposition, inter-molecular ineffective collision, and synthesis*.

On-Wall Effective Collision

This may happen when a molecule hits the wall and bounces back. An ineffective collision indicates that changes in molecular structure are minimal. Thus the molecule representing a solution would change in its neighborhood, based on equation (1).

$$PE_{\omega} + KE_{\omega} \geq PE_{\omega'} \quad (1)$$

Decomposition

A decomposition occurs when a molecule encounters a collision with the wall and decomposes into two pieces. The resultant molecules are expected to be very different from the original molecule. The decomposition is possible if condition (2) is satisfied.

$$PE_{\omega} + KE_{\omega} \geq PE_{\omega 1} + PE_{\omega 2} \quad (2)$$

Here $\omega 1$ and $\omega 2$ are the resultant molecules. Generally, this case becomes unlikely since ω , $\omega 1$ and $\omega 2$ can have similar PE values. Thus, a central energy buffer is accessed for extracting KE to facilitate the decomposition.

Intermolecular Ineffective Collision

This process involves the collision of two molecules, which have bounced back. The collision results in states, where a condition needs to satisfy equation (3).

$$PE_{\omega 1} + PE_{\omega 2} + KE_{\omega 1} + KE_{\omega 2} \geq PE_{\omega 1'} + PE_{\omega 2'} \quad (3)$$

Synthesis

In a synthesis, two molecules vigorously combine to create a resultant molecule if equation (4) is satisfied.

$$PE_{\omega 1} + PE_{\omega 2} + KE_{\omega 1} + KE_{\omega 2} \geq PE_{\omega'} \quad (4)$$

B. CRO based Gene Selection (CRO-GS)

As stated initially, a simplistic chemical reaction optimization (CRO) model has been adopted for simultaneous gene selection and cancer classification, in this study. Gene selection involves extracting a subset of informative genes from the given samples available as part of our datasets [1]. This section illustrates the use of a CRO based technique in the form of a filter-wrapper algorithm in conjunction with Support Vector Machines (SVM) [9] for extracting informative gene subsets.

CRO-GS involves initializing a number of molecules (or solutions), where each molecule is encoded as a feature vector having a predefined size. The size of the vector may be decided by the user. For example, for a total of 1000 features and a predefined subset size of 10, a feature vector (or a molecule in this case) will have 10 entries. Each entry can have a feature index (selected from 1...1000) as its value. Thus, if X is a molecule, then a possible configuration of X may correspond to {1, 10, 13, 17, 67, 78, 89, 90, 391, 992}, where each element indicates the corresponding feature in the dataset. Initially, a population of molecules is generated randomly. The potential energy (PE) of each molecule is computed using relation (5).

$$PE_{mol} = 100 - CVA \quad (5)$$

In (5), CVA is the 10 fold SVM cross validation classification accuracy of the concerned molecule (based on the feature subset it encodes). PE_{mol} corresponds to the objective function value (as discussed before) for a molecule. In the gene selection problem, CRO attempts to minimize PE_{mol}.

Additionally, before a chemical reaction is initiated the central energy buffer is initialized with a value (considered as a CRO parameter). The chemical reaction iteration stage in CRO-GS is controlled by a decision against the CRO collision rate parameter. The flow of the CRO control may thus move into unimolecular collisions or inter-molecular interactions, depending on the condition. For a unimolecular collision, a molecule may be randomly selected from the current set. A neighboring molecule is then obtained by selecting a random position in the feature vector (the molecule) and replacing the corresponding feature index

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

value by probabilistically selecting a feature index from the entire set of features based on a gene ranking heuristic (described in *section C*). Next, the potential energy (PE) of the new molecule is computed using equation (5). Using equation (1), we compare the PEs of the original and the neighboring molecules, following which either an *on-wall ineffective collision* or a *decomposition* may be initiated.

For intermolecular collisions, two molecules may randomly be selected and the neighboring molecules for each of them are generated using a similar procedure as described before. If the sum of the PEs of newer molecules is less than the earlier ones (refer equation 3), then an *intermolecular ineffective collision* is initiated and the related changes are carried out. Otherwise, a synthesis of the two molecules takes place where a new molecule is generated by combining the two molecules. This is done by probabilistically selecting a feature index from either of the two molecules, while iterating through each position of the new molecule. Thus the new molecule is a combination of the two existing molecules, by selecting a feature index from either of the two at the corresponding feature vector position.

The above stages constitute a single iteration of a chemical reaction. This process may continue for a maximum number of iterations. At the end of the final iteration, the molecule with the least PE (or cross validated error rate) is selected as the most optimal gene subset.

A representative version of the CRO-GS gene selection algorithm is stated in Fig. 1.

```
Algorithm: CRO-GS for Gene selection
Initialize CRO Parameters
while (termination criteria not reached)
do {
    if(rand > collisionRate){
       generate single neighbouring molecule
       if(oldmoleculePE > newmoleculePE){
           do on wall ineffective collision
       }
       else{
           do decomposition using KE from buffer
       }
    }
    else{
       generate two neighbouring molecules
       if(sum(oldmoleculesPE) > sum(newmoleculesPE)){
           do intermolecular ineffective collision
       }
       else{
           do synthesis using KE from buffer
       }
    }
       update energy buffer
}
End while
Return best molecule with most optimal gene subset
```

Figure 1: CRO-GS Algorithm for Gene Selection.

C. Heuristic Gene Ranking

Due to the massive search space of possible gene subsets, we make use of the Information gain (IG) filter to provide a ranking of all genes [10]. The infogain gene ranking is subsequently used to create a neighboring molecule. Information Gain (IG) is an entropy-based measure, which selects the gene that has the best capability to differentiate the samples into separate classes. A gene with a higher IG is considered more relevant. A neighboring molecule is generated by replacing a feature at a random position in an original molecule. To do this, we may probabilistically select a good feature based on non-zero infogain values provided by the gene ranking. A probabilistic selection of a gene using the infogain gene ranking is illustrated in Fig. 2. As shown, the 17th feature is replaced by the 100th feature (accessed through the gene ranking) based on a threshold decision.

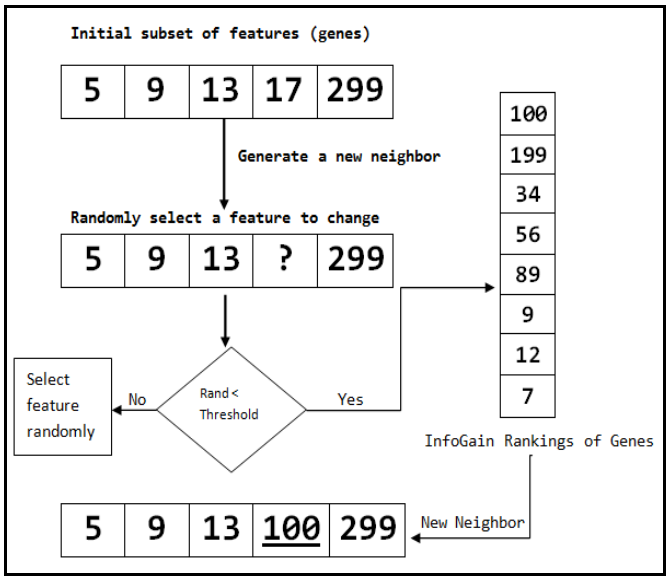

Figure 2: Neighbor generation using Infogain gene ranking

D. Support Vector Machines

Support Vector Machines (SVMs) [8] help construct a classification model by employing a maximum margin linear hyper-plane to solve binary linear classification problems. For non-linear problems, SVM transforms the input data to higher dimensional features and then attempts to apply a linear hyper-plane. SVM also employs appropriate kernel operations allowing computations in the input space to deal with intractability. For our purposes, we employ the LIBSVM [11] software suite for evaluation of the molecules (feature vectors) in each iteration.

In CRO, a population of molecules tries to go through four possible transitions mentioned before based on their potential energies in an iterative manner to reach to the final optimal set of genes. The CRO based data flow architecture is illustrated in Figure 3.

III. Results and Discussions

In order to test the performance of CRO, we conducted extensive simulations of the hybrid CRO filter-wrapper algorithm for three benchmark cancer gene expression datasets, obtained from the Kent Ridge Biomedical datasets repository [12] and the LIBSVM repository [11] (made available from various other sources). The dimensions of the datasets are tabulated in Table I.

Based on our simulations, one can say that comparable results for all three datasets were observed, while considering a maximum of 100 initial molecules and 10000 iterations. Generally, at the end of 100 generations, the fitness values of

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the gene subsets would converge and not show much improvement. Parameter tuning was also carried out extensively for CRO and SVM to arrive at optimal results. The algorithm parameters for CRO - SVM are as shown in Table II.

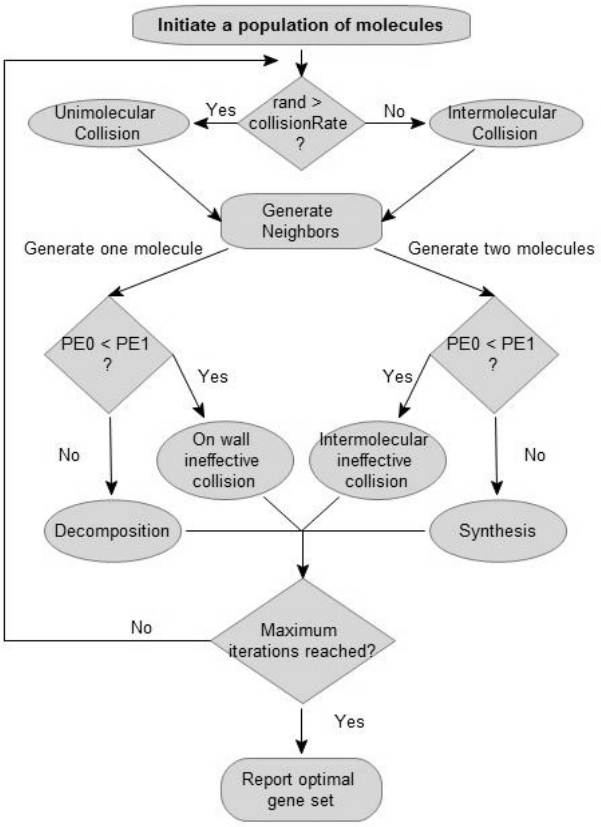

Figure 3. CRO-GS Data Flow Architecture

TABLE I
DATA DIMENSIONS

Cancer Dataset	*No. of Genes*	*No. of classes*	*No. of Samples for I & II*
Colon	2000	2	62 (40 & 22)
Breast	7129	2	44 (22 & 22)
Leukemia	7129	2	72 (25 & 47)

TABLE II
CRO-GS PARAMETERS

CRO Parameters	*Values*
Collision Rate	0.1
Decomposition threshold	1500
Synthesis threshold	10
Step Size	0.2
Initial KE in buffer	10000
SVM kernel Type	Radial basis function (rbf)
SVM gamma -g	0.02
SVM cost -c	50

Consequently, we obtained the gene subsets that reported the least 10 fold CV error rate. To obtain consistent CRO parameter estimates as given in Table II, we carried out 30 runs for each dataset. According to results in Table III, CRO-GS performs well in comparison to previously reported algorithms for all the three datasets. The CRO-GS based gene subset sizes selected were 20 for Colon, 6 for Breast and 20 for Leukemia. For colon cancer, CRO-GS reported 95.16% (10 fold CVA) which compares well against ACO-RF (95.47%), ACO-SVM (96.77%), and BBO-SVM (98.39%) reported earlier [3, 5]. For the duke breast cancer data, with a 10 fold CVA of 97.36% CRO-GS has performed well in contrast to some of the more powerful models based on Bagging (92%), BBO-SVM (99.56%) and Ensemble (94%) techniques [13, 5]. CRO-GS with leukemia reported a 10 fold CVA of 100%, which was compared with a baseline SVM model (97.06%), BBO-SVM (99.60%) and ACO-AM (96%), reported earlier [14, 5].

TABLE III
CRO-GS RESULTS

Cancer Dataset	*Colon*	*Duke Breast*	*Leukemia*
10 fold CVA	95.16%	97.36%	100%

Comparison of these results with other methods certainly entail CRO-GS as a promising methods to identify informative genes from cancer microarray data and use these informative genes in turn to make a robust classification model for differentiating between cancer types. Main contribution of this methodology would be to identify a set of genes, which might be playing an important role in a certain kind of diseases and to build a robust classification model to classify different classes of a disease.

IV. CONCLUSION

The hybrid CRO-GS demonstrates good results consistently on comparison with the highest accuracies for colon cancer, breast cancer and leukemia cancer datasets. In general, CRO is robust and flexible for discrete optimization. One can significantly speedup the algorithm by possible parallel implementations where the classification accuracies for individual candidate solutions (or molecules) may be computed in parallel.

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," The Journal of Machine Learning Research, vol. 3, pp.1157–1182, 2003.

[2] D. Patil, R. Raj, P. Shingade, B. Kulkarni, and V. K. Jayaraman, "Feature selection and classification employing hybrid ant colony optimization/random forest methodology," Combinatorial chemistry & high throughput screening, vol. 12, no. 5, pp. 507–513, 2009.

[3] S. Sharma, S. Ghosh, N. Anantharaman, and V. K. Jayaraman, "Simultaneous informative gene extraction and cancer classification using aco-antminer and aco-random forests," in Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. Springer, 2012, pp. 755–761.

[4] A. Gupta, V. K. Jayaraman, and B. D. Kulkarni, "Feature selection for cancer classification using ant colony optimization and support vector machines." 2007.

[5] S. Nikumbh, S. Ghosh, and V. K. Jayaraman, "Biogeography-based informative gene selection and cancer classification using svm and random forests," in Evolutionary Computation (CEC), 2012 IEEE Congress on. IEEE, 2012, pp. 1–6.

[6] A. Y. Lam and V. O. Li, "Chemical-reaction-inspired metaheuristic for optimization," Evolutionary Computation, IEEE Transactions on, vol. 14, no. 3, pp. 381–399, 2010.
[7] J. Xu, A. Y. Lam, and V. O. Li, "Chemical reaction optimization for task scheduling in grid computing," Parallel and Distributed Systems, IEEE Transactions on, vol. 22, no. 10, pp. 1624–1631, 2011.
[8] A. Y. Lam and V. O. Li, "Chemical reaction optimization: A tutorial," Memetic Computing, vol. 4, no. 1, pp. 3–17, 2012.
[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory. ACM, 1992, pp. 144–152.
[10] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques. Morgan kaufmann, 2006.
[11] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, p. 27, 2011.
[12] J. Li and H. Liu, "Kent ridge bio-medical data set repository," Institute for Infocomm Research. http://sdmc. lit. org. sg/GEDatasets/Datasets.html, 2002.
[13] A. Blanco, M. Mart´n-Merino, and J. De Las Rivas, "Combining dissimilarity based classifiers for cancer prediction using gene expression profiles," BMC Bioinformatics, vol. 8, no. Suppl 8, p. S3, 2007.
[14] G. Cong, K.-L. Tan, A. K. Tung, and X. Xu, "Mining top-k covering rule groups for gene expression data," in Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005, pp. 670–681.
[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009.

ISBN: 978-988-19252-7-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)