INTELLIGENT TECHNIQUES FOR PREDICTION OF CYTOSINE-PHOSPHATE-GUANINE (CPG) SITES ASSOCIATED WITH LUNG CARCINOMA

Abdel-Badeeh Salem, Mohamed Gawish, Mohammed Maher, Yasmin Amr, Amany Hussein, Keryakous Zarif, Basant Mohamed, Hebatullah Mohamed

Abstract: *DNA Methylation is a process by which cell assure the regulation of gene expression. Improved methods of detecting methylated sites are needed, instead of experimental methods which are very expensive and time consuming. In this paper, metaheuristic techniques namely; Genetic Algorithm, Artificial Immune System, and Hybrid Immune Genetic Algorithm are implemented to solve the problem of feature selection and select the susceptible CpG sites from dataset. Reducing the dimensionality of the dataset by applying previous algorithms resulting the following sets. After running the three algorithms many numbers of iterations, the average number of CpG sites determined by each algorithm is found to be less than 10% of the original dataset size. A new signature set was created by gathering all the common CpG sites from the three generated sets. Its size is equal to 0.1% of the original dataset size. Then it is used to generate the proteins regulatory network.*

Keywords: *CpG sites, Genetic Algorithm, Artificial Immune System, Clonal selection Algorithm, Hybrid Algorithms, Kruskal Wallis Test, Bioinformatics.*

ACM Classification Keywords: *I.5.2 Computing Methodologies - Pattern Recognition - Design Methodology-Feature evaluation and selection; I.2.8 Computing Methodologies – Artificial Intelligent - Problem Solving, Control Methods, and Search - Heuristic methods*

Introduction

DNA is a code of life; it is a polymer of four simple nucleic acids units called nucleotides. It consists of four nucleotide bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). It is responsible of passing of genetic data between generations. CpG sites are where a Cytosine nucleotide exists next to a Guanine nucleotide separated by a phosphate group. DNA Methylation is one of the main factors causing gene silencing leading to an epigenetic change. Epigenetics involve inactivation of tumor suppressors and activation of oncogenes (Aine et al., 2015; Waterland & Michels, 2007). DNA Methylation involves addition of a methyl (CH3) group covalently at carbon 5 in the pyrimidine ring of a cytosine base, DNA methylation generally occurs in the context (5'-CG-3')dinucleotides, the

methylation pattern differs from one cell to another (depending on the functions which the cell needs to be active) and from a disease to another(Ahn & Wang, 2013; Meng, Murrelle, & Li, 2008).

CpG islands are present on the promoter of genes, found un-methylated in a normal state but methylated in Cancer with presence that's five times larger than the normal state. Distinguishing the methylated Cytosine (5mC) is based on principles like Bisulfite conversion which differentiates between methylated and un-methylated Cytosine by treating the DNA with sodium Bisulfite while un-methylated Cytosine turn into Uracil , methylated Cytosine are not affected, the changes resulting in the DNA sequence can be detected through PCR amplification proceeded by DNA sequencing as (Guo et al., 2015).

A number of CpG sites in a high throughput methylation arrays are irrelevant and don't provide information to distinguish the normal cells of cells with Cancer. In this paper efficient computational intelligence techniques such as Genetic Algorithm, Artificial Immune System, and Hybrid Immune Genetic Algorithm have been utilized to reduce the number of CpG sites resulted from a high throughput methylation array resulting in CpG sites most likely causing Cancer. Although wrapper methods generally outperform filter methods, they are computationally intensive and may become inefficient in practice for large datasets (Guyon, Weston, Barnhill, & Vapnik, 2002; Li, Zhang, & Ogihara, 2004; Zhang et al., 2006).

This paper is organized in the following manner. First we show existing work in predicting CpG sites. Then we present the three techniques used on our work for predicting CpG sites. Experimental design, results, and statistical analysis are presented afterwards. Followed by a section that provides discussion. Finally, conclusions is provided.

Related Work

There are two types of feature selection methods. The first type is filter-based methods that assess the relevance of features by looking only at the intrinsic properties of the date. Filter-based methods are quite popular because they are more efficient, more scalable, and independent of the classification algorithm. On the other hand, they have limitations and the classification accuracy of the selected genes is less accurate. The other type of feature selection methods is the wrapper methods, which employ classifiers to determine feature selection based on the predictive accuracy of the classifier(Guyon et al., 2002; Li et al., 2004; Zhang et al., 2006).

One work that is worth mentioning here is the work done by Model, Adorjan, Olek, and Piepenbrock (2001). In that work, the simple Fisher criterion was used as a feature selection strategy combined with SVM in order to discriminate between acute lymphoblastic leukemia and acute myeloid leukemia using methylation pattern data.

Another work is done by Meng, Murrelle, and Li (2008). In that work a two-stage feature selection method was developed to select a small optimal subset of DNA methylation feature to distinguish lung cancer tissue samples from normal lung tissue samples using DNA methylation data.

Current work

In this work, the problem of predicting CpG sites has been tackled by three metaheuristic techniques (Genetic Algorithm, Artificial Immune System, and Hybrid Immune Genetic Algorithm).As shown in Figure 1, the three techniques are applied on the same dataset to get three different sets of CpG sites sets(Kim, Park, & Kon, 2013; Xu & Zhang, 2005). After that, a new set called signature set was created by gathering the common CpG sites of the three generated sets in order to find the most related CpG sites to lung carcinoma. Finally, the signature set was used to produce the proteins regulatory network.

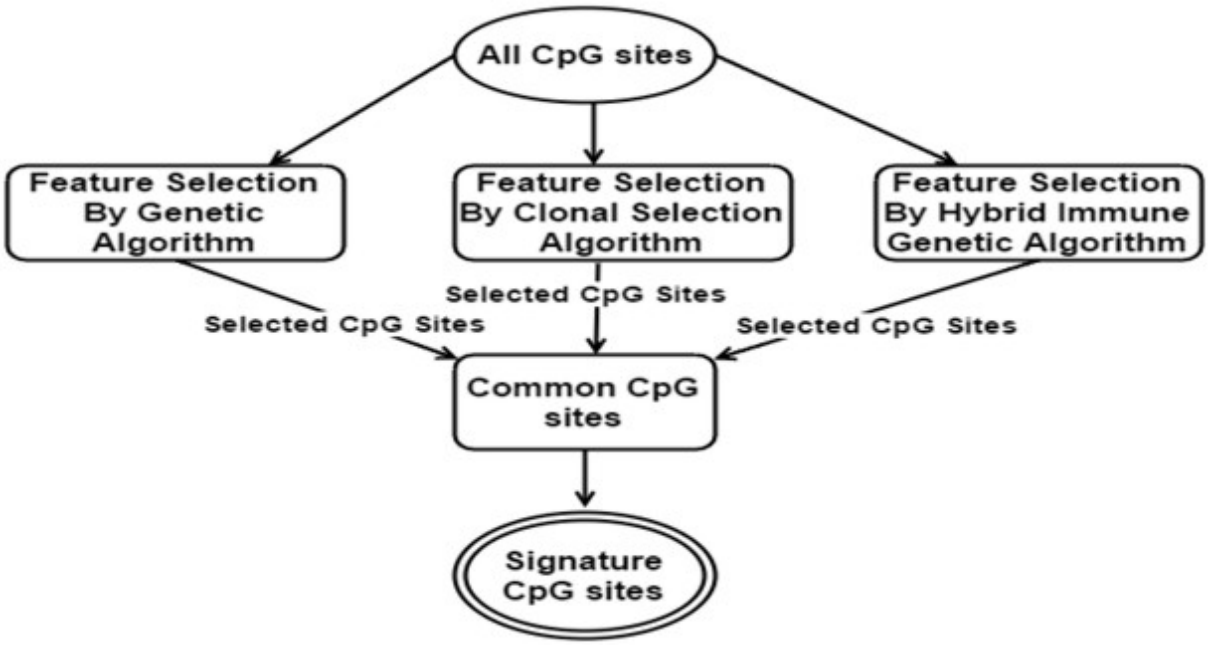


Figure 1.Generating the signature CpG sites

Solving CpG sites selection by Genetic Algorithms

Overview

In the early 1970s, John Holland introduced the concept of genetic algorithm (GA) as a class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by Darwin's classical theory of natural evolution, such as inheritance, mutation, selection, and crossover.

GA steps

Figure 2 presents the steps of GAs (Jourdan, Dhaenens, & Talbi, 2001; Vafaie & De Jong, 1992),

- Step 1: initialize population of chromosomes randomly: The population has a set of binary chromosomes that are of fixed size. The size equals the number of features in the dataset, Fill chromosomes randomly 0 or 1; 0 means that feature is deactivated and 1 means that feature is activated.
- Step 2: evaluate fitness for each chromosome by getting all active feature (1's index) in this chromosome, then create subset dataset by getting expressed data for these features from the original dataset then pass it to support vector machine to calculate validation error. The fitness of this chromosome is 1/validation error (Guyon et al., 2002; Kim et al., 2013).
- Step 3: select two chromosomes for reproduction by using roulette wheel (Banzhaf & others, 1999).
- Step 4: Crossover between the two selected chromosomes, single point crossover is applied to get offspring as illustrated in Figure 3.
- Step 5: Bit-Flip mutation for offspring as illustrated in Figure 4.
- Step 6: Replacement by using elitist strategy by evaluating fitness for each offspring; if its fitness greater than its parents, replace.
- Step 7: Repeat Steps 2 through 6 while the validation error is minimum.

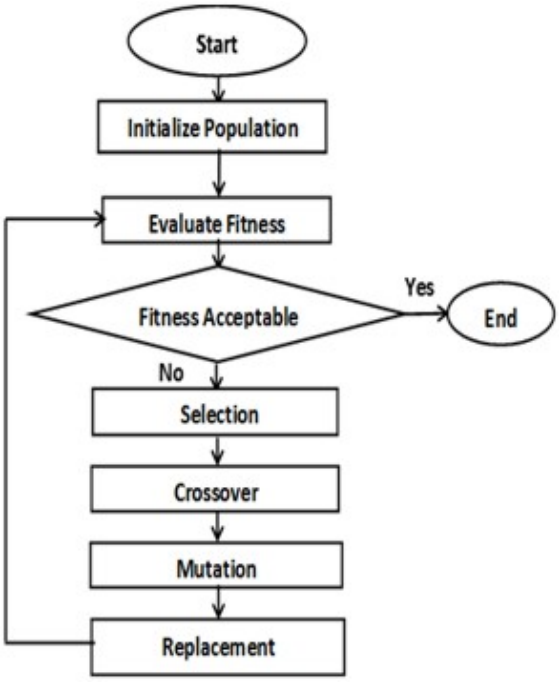


Figure 2. GA flowchart

Crossover

Applying one point crossover on selected chromosomes to get offsprings by generating numbers R1 and R2 randomly then compare if R2 is smaller than or equal to probability of crossover (0.5) do crossover between parents to get offsprings, otherwise, offsprings equal parents.

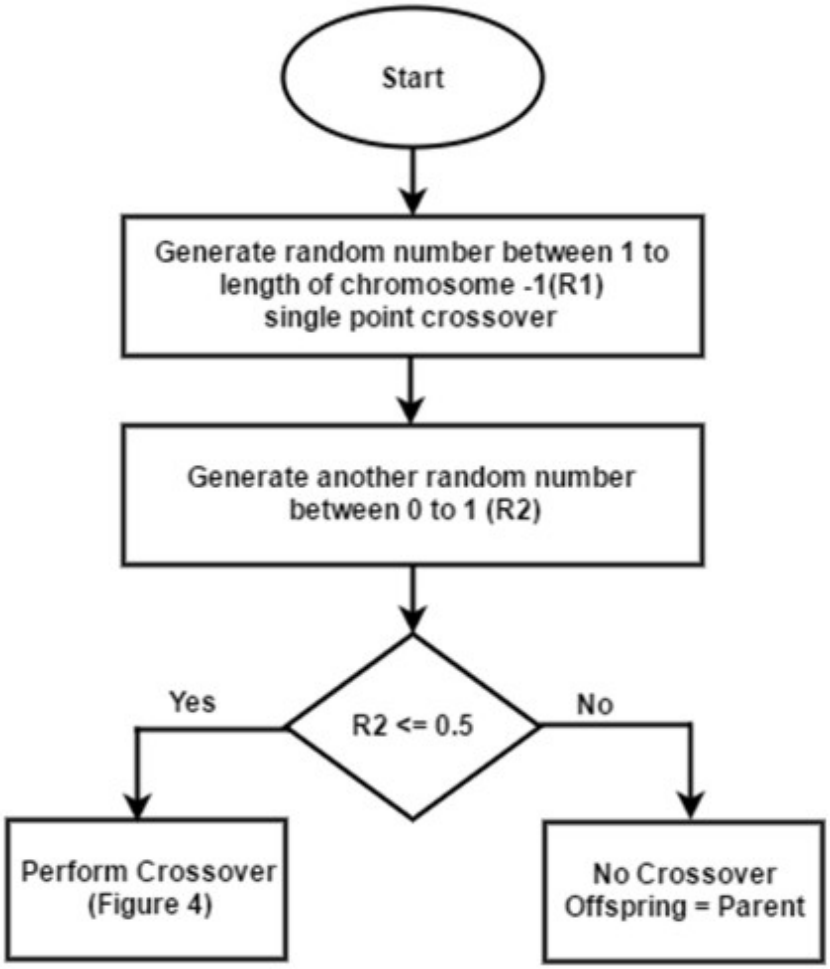


Figure 3. Crossover flowchart

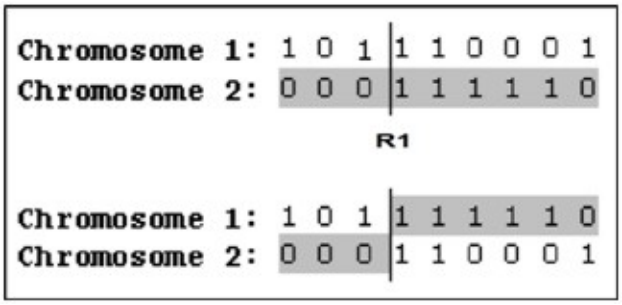


Figure 4. One-Point Crossover

Mutation

Two different probabilities are applied one for mutating to one (small value 0.05) and the other for mutating to zero (large value 0.5) ;to reduce the amount of 1s to get the final result faster as illustrated in Figure 5 .

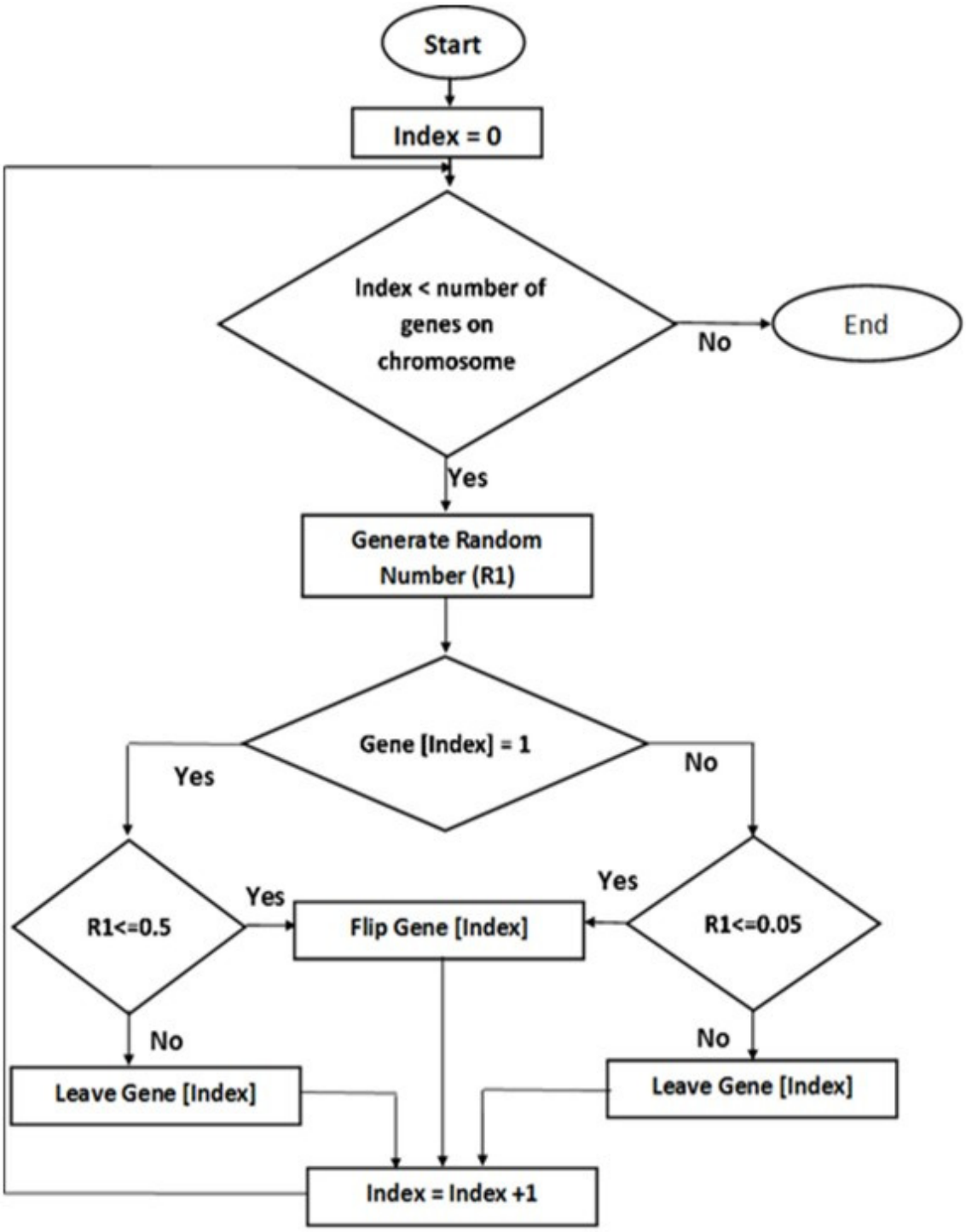


Figure 5. Mutation flowchart

Replacement

The fitness of the offspring is calculated, and then if it's higher than the fitness of the parents, the parents will be replaced by the offspring to get the new population which has a higher fitness as in genetic algorithm the fittest only can survive.

Solving CpG sites selection by Artificial Immune System

Overview

Artificial Immune System (AIS) is a metaheuristic search inspired by the theoretical immunology and observed immune functions principles and models, which are applied to complex problem domains. Its flowchart is presented in Figure 6.

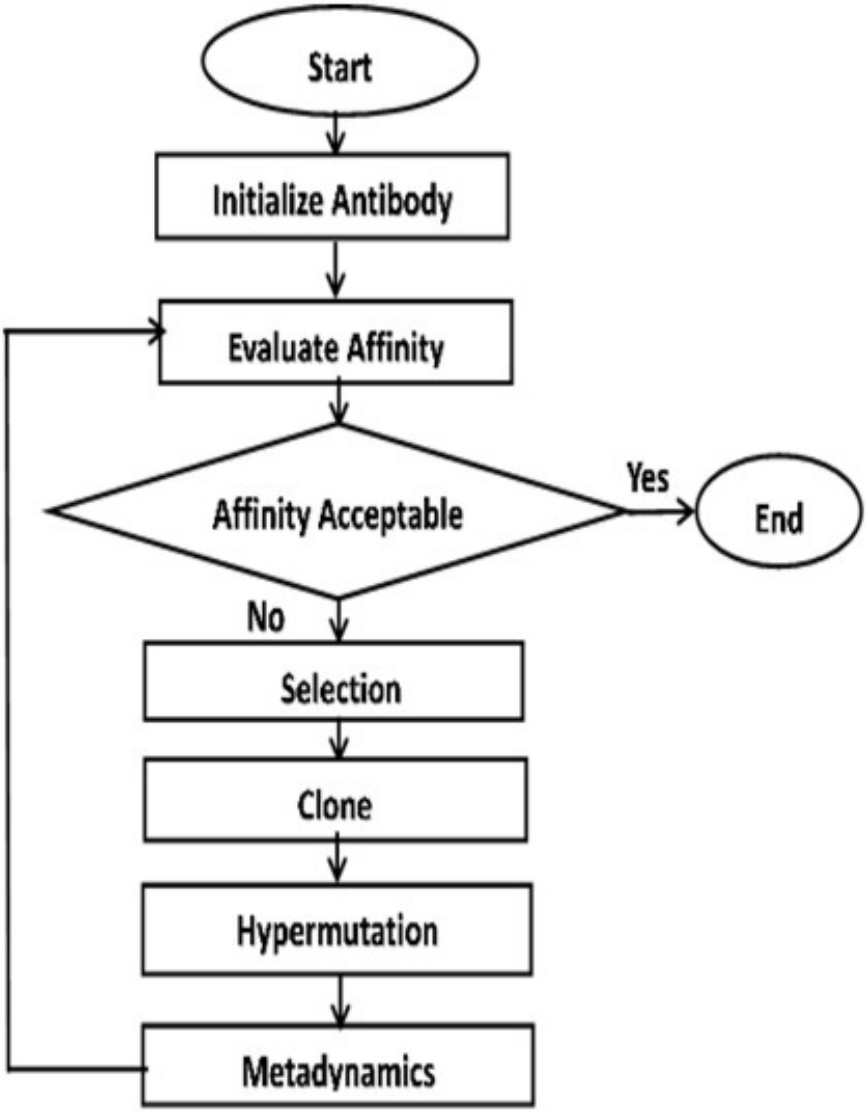


Figure 6. AIS flowchart

AIS Steps

- Step 1: initialize population of antibodies randomly: The population has a set of binary antibodies that are of fixed size. The size equal to the number of features in the dataset. Fill antibodies randomly 0 or 1; 0 means the feature is not active and 1 means the feature is active.
- Step 2: evaluate affinity for each antibody by getting all active feature (1's index) in this antibody then create subset dataset by getting expressed data for these features from the original dataset then pass it to Support vector machine to calculate validation error, the affinity of the antibody is 1/validation error(Guyon et al., 2002; Kim et al., 2013).
- Step 3: select populations for Lowest 20 percentage of Affinity that is better matching between antibody and antigen
- Step 4: Copy lowest 20 percentage to replace it in the next lowest 20 percentage of population
- Step 5: Hybermutation for selected and cloned (result on step 3 and 4) 40 percentage; by generating a random number; if the random number is less than probability of mutation flip bit.
- Step 6: Randomize lowest 60 percentage of population that's called Metadynamics
- Step 7: Repeat Steps 2 through 6 while the validation error is minimum.

Solving CpG sites selection by Hybrid Immune Genetic Algorithm

For increasing the exploration of a search space, Hybrid Immune Genetic Algorithm (HIGA) is proposed by Nabil, Badr, and Farag (2009), which is a result of hybridizing AIS with the GA's crossover operator. Its flowchart is presented in Figure 7.

HIGA steps

- Step 1: initialize population of antibodies randomly: The population has a set of binary antibodies that are of fixed size. The size equal to the number of features in the dataset. Fill antibodies randomly 0 or 1; 0 means the feature is not active and 1 means the feature is active.
- Step 2: evaluate affinity for each antibody by getting all active feature (1's index) in this antibody then create subset dataset by getting expressed data for these features from the original dataset then pass it to Support vector machine to calculate validation error, the affinity of the antibody is 1/validation error(Guyon et al., 2002; Kim et al., 2013).
- Step 3 select populations for Lowest 20 percentage of Affinity that is better matching between antibody and antigen
- Step 4: Copy lowest 20 percentage to replace it in the next lowest 20 percentage of population

- Step 5: Hybermutation for selected and cloned (result on step 3 and 4) 40 percentage; by generating a random number; if the random number is less than probability of mutation, flip bit.
- Step 6: Do Crossover on selected population after Hybermutation not cloned population and replace the selected population by the new offspring.
- Step 7: Randomize lowest 60 percentage of population that's called Metadynamics
- Step 8: Repeat Steps 2 through 7 while the validation error is minimum.
-

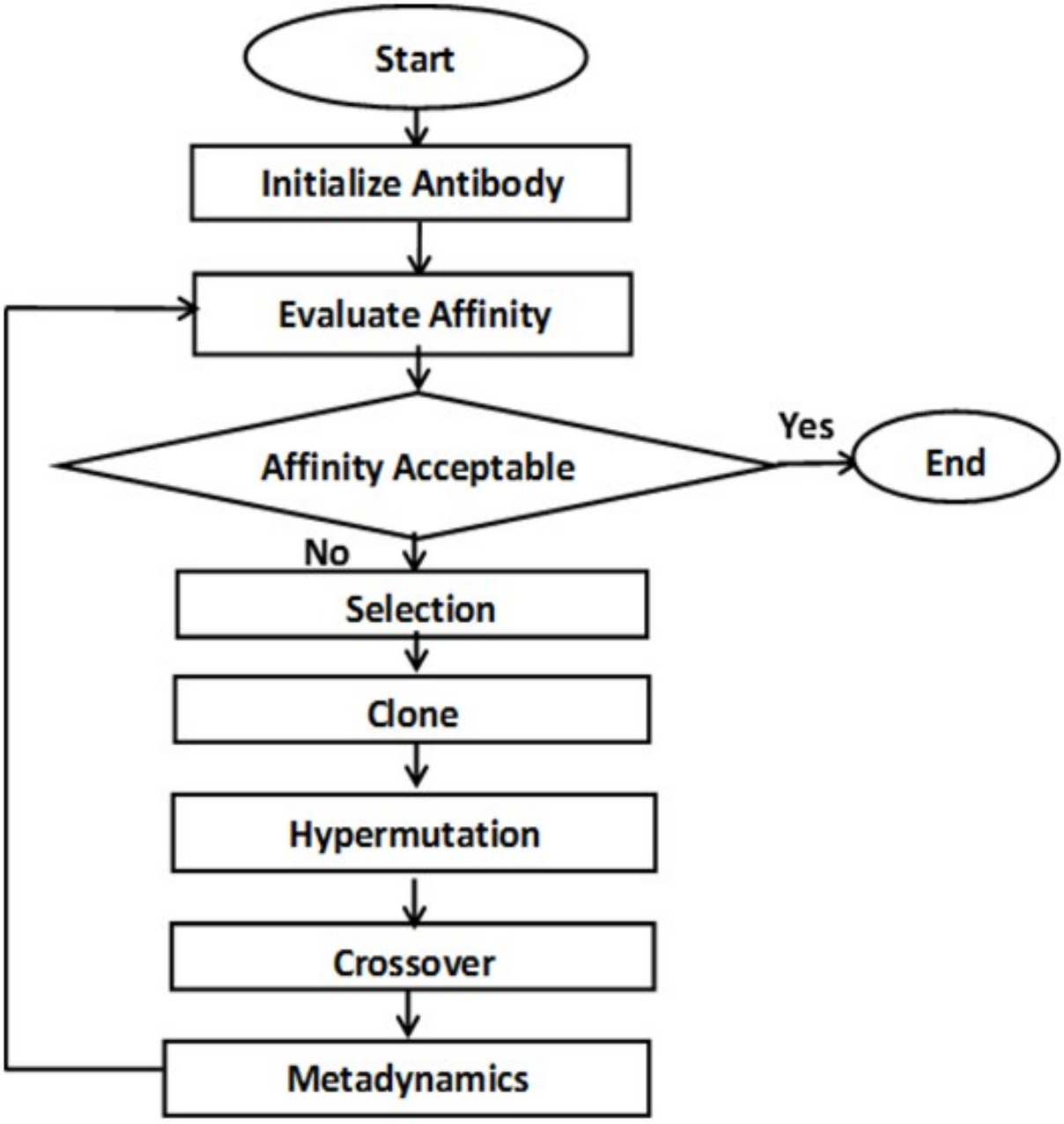


Figure 7. HIGA flowchart

The evaluation function

The same evaluation function was used in the three metaheuristic techniques in order to calculate the fitness or affinity value. It takes the dataset and the binary (chromosome/antibody) as inputs. The binary (chromosome/antibody) are used to select a subset from the original dataset by choosing only features that has a corresponding active value in the (chromosome /antibody).

The SVM that is one of state-of-the-art classification method was chosen to calculate the test error. It has been widely used in microarray data analysis (Guyon et al., 2002).

Leave-one-out cross-validation was employed to evaluate the classification performance of each subset. Each sample was excluded from the training set, one at a time, and then classified based on the SVM trained from the remaining samples. This procedure was repeated, in turn, for all samples, and the cross-validation error was defined as the sum of misclassifications. Finally the function returns the cross-validation error as its output.

Finding the Signature CpG sites

After applying the previous three metaheuristic techniques on the same dataset, three different sets of CpG sites are generated. In order to find the smallest and most important CpG sites set, a new set (signature) was created that has only the common CpG sites of the three generated sets as depicted in Figure1.

Generating the regulatory network

After generating the signature CpG sites, the dataset has been used in order to return the related genes (Network & others, 2014). A set of proteins that relate to the genes are entered into string-db to produce the proteins regulatory network as shown in Figure 8(Szklarczyk et al., 2010).

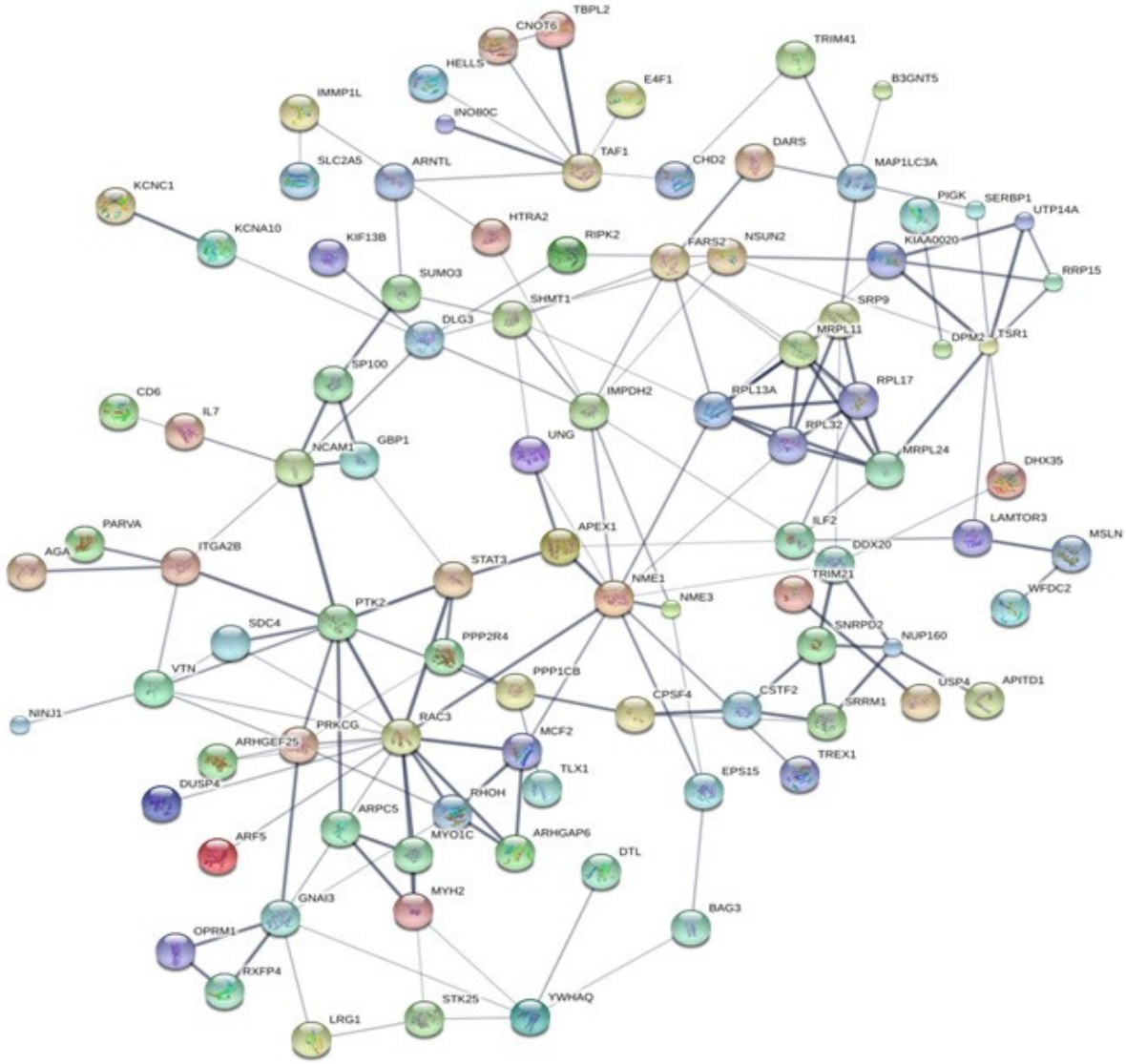


Figure 8. The proteins regulatory network

Results and Discussion

Dataset

The dataset used in this work is collected through the analysis of tumor and matched normal material from previously untreated lung adenocarcinoma patients (Network & others, 2014). And it is extracted using an Illumina array platform which is based on Bisulfite conversion, with an accuracy of 99.9% which can detect a 17% difference in DNA methylation and 2.5% DNA methylation as (Müller, Assenov, & Lutsik, 2015).It consists of twenty persons; each person has twenty seven thousand (27,000) CpG sites, and labeled by the predicted output, as the CpG sites of the highest fitness only can survive.

Array based genome wide Methylation analysis methods based on Bisulfite Conversion generates high throughput of about 27000 assays per sample, also known for their ability to derive information at single nucleotide resolution which make these methods very specific .

Parameter settings

In the used metaheuristic techniques several parameters have to be assigned. The maximum number of generations is set to 500 for GA, AIS, and HIGA. For all techniques, the population's size is set to 10.Both GA's crossover probability and HIGA's crossover probability are set to 0.5. The [1 to 0] mutation probability is set to 0.5. The [0 to 1] mutation probability is set to 0.05 as illustrated in table 1.Each technique has been run 100 times.

Table 1. Parameter Settings

Parameter name	GA	AIS	HIGA
Population size	10	10	10
Max No. of generations	500	500	500
Cross over probability	0.5	-	0.5
[1->0] mutation probability	0.5	0.5	0.5
[0->1] mutation probability	0.05	0.05	0.05
No. of run	100	100	100

Experimental Results

In the experiment, the performance of the used techniques has been evaluated over the dataset using SVM classifier. The comparison between the different techniques as shown in Figure8is in term of the best accuracy classification returned by SVM. It can be seen from Figure 9 that the HIGA technique obtains the best accuracy classification compared to the other techniques. The GA technique obtains the second best accuracy classification. Finally the AIS technique comes at the end.

After running the three algorithms for 100 iterations, the average number of CpG sites determined by the Genetic algorithm found to be from 2050 ~2100 , And Clonal selection algorithm from 2100~ 2200 and Immune genetic hybrid algorithm from 2000~2065 , The Combination between 3 algorithms resulted 25CpG site.

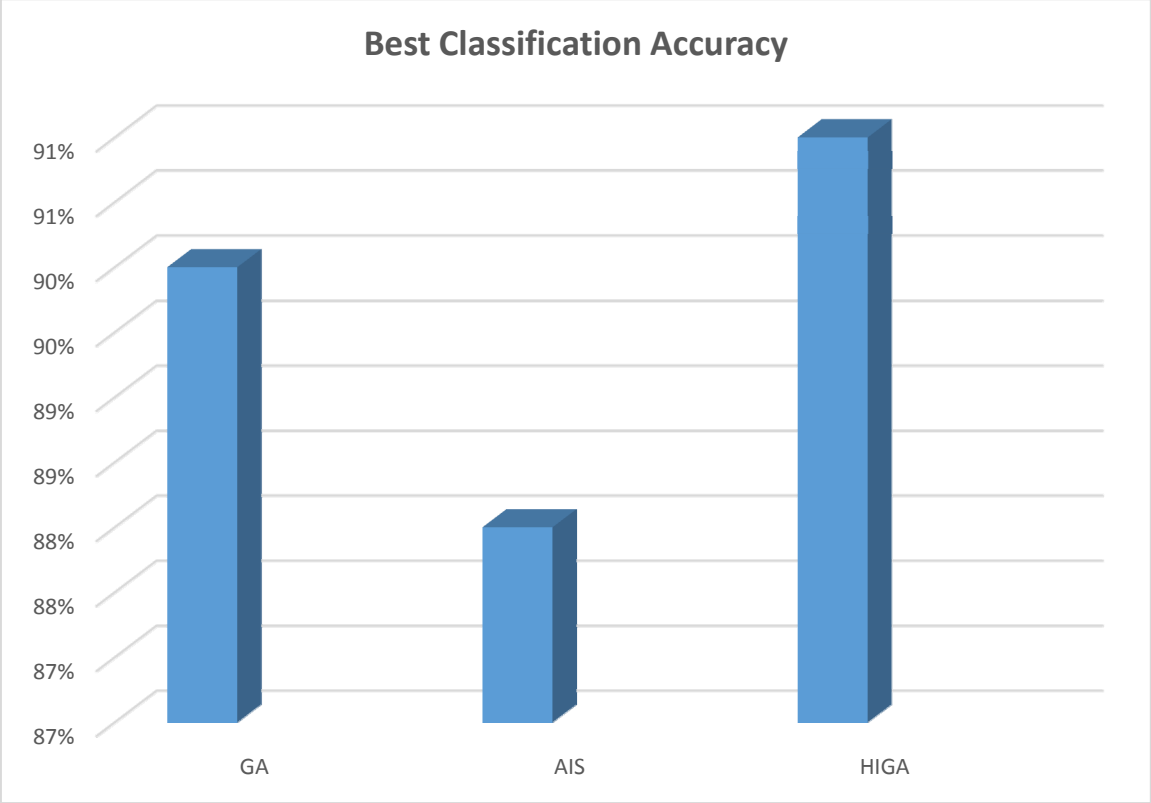


Figure 9. The best classification accuracy

Statistical analysis

In order to illustrate that the experimental results are statistically significant, the Kruskal-Wallis test has been performed on the results. The Kruskal-Wallis test is a rank-based nonparametric test. It can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable (McKight & Najab, 2010).

Conclusion and Future Work

Features selection is necessary in order to get a reliable predication results by using a small size training datasets. So, this work has tackled the problem of features selection by using three different techniques (GA, AIS, and HIGA).The three techniques have produced three different subsets each one contains different features. In term of best classification accuracy, the HIGA algorithm has outperformed both GA and AIS.

In order to get the most relevant features a new subset (signature) was created by gathering the common features from the three subsets. The proteins regulatory network that relevant to the signature subset was produced using the string-db.

Protein ligand interaction can also be investigated in conjunction with protein-protein interaction. Likewise, other metaheuristics can be investigated such as cuckoo search bat algorithm, firefly algorithm, differential evolution and others.

This study can be extended to the Four hundred Fifty thousand CpGs once a superior hardware is available. This incurs to the use of deep learning neural networks such as Convolutional neural network, deep belief networks and restricted Boltzmann machines, this is because of the huge amount of features (450,000) that will necessitate the use of deep learning.

Acknowledgement

"The paper is published with partial support by the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua)".

Bibliography

[Ahn & Wang ,2013]. Ahn S. & Wang T., „A powerful statistical method for identifying differentially methylated markers in complex diseases'' , *Pacific Symposium on Biocomputing , p.69.*

[Aine, Sjödahl, Eriksson, Veerla, Lindgren, Ringnér, Höglund, 2015], Aine M., Sjödahl G., Eriksson P., Veerla S., Lindgren D., Ringnér M. & Höglund M., „Integrative epigenomic analysis of differential DNA methylation in urothelial carcinoma'', Genome mid, 7(1), 23.

[Banzhaf & others,1999] *Foundations of genetic algorithms';* Morgan Kaufmann Publishers Inc.

[Guo, Yan, Xu, Bao, Zhu, Wang, … others, 2015] . Guo, S.,Yan, F., Xu, J., Bao, Y., Zhu, J., Wang, X., … others, „ Identification and validation of the methylation biomarkers of non-small cell lung cancer (NSCLC)'', *Clinical Epigenetics*, *7*(1), 1–10.

[Guyon, Weston, Barnhill & Vapnik, 2002]. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V., „ Gene selection for cancer classification using support vector machines'' *Machine Learning*, *46*(1-3), 389–422.

[Jourdan, Dhaenens & Talbi, 2001]. Jourdan, L., Dhaenens, C. & Talbi, E.-G, „ A genetic algorithm for feature selection in data-mining for genetics'', *Proceedings of the 4th Metaheuristics International ConferencePorto,(MIC'2001)*, 29–34.

[Kim, Park & Kon, 2013], Kim, S., Park, T. & Kon, M. A, „Computational methods for cancer survival classification using intermediate information'' , *IWBBIO* (pp. 517–525).

[Li, Zhang & Ogihara, , 2004]. Li, T., Zhang, C. & Ogihara, M. , „A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression.'', *Bioinformatics*, *20*(15), 2429–2437.

[McKight & Najab, 2010]. McKight, P. E. & Najab, J. „ Kruskal-Wallis Test''. *Corsini Encyclopedia of psychology.*

[Meng, Murrelle & Li, 2008]. Meng, H., Murrelle, E. L. & Li, G. , „Identification of a small optimal subset of CpG sites as bio-markers from high-throughput DNA methylation profiles.'', *BMC Bioinformatics*, *9*(1), 457.

[Model, Adorjan, Olek & Piepenbrock, 2001]. Model, F., Adorjan, P., Olek, A. & Piepenbrock, C. , „Feature selection for DNA methylation based cancer classification.'' ,*Bioinformatics*, *17*(suppl 1), S157–S164.

[Müller, Assenov & Lutsik, 2015]. Müller, F., Assenov, Y. & Lutsik, P.„RnBeads-Comprehesive Analysis of DNA methylation Data.''

[Nabil, Badr & Farag, 2009]. Nabil, E., Badr, A., & Farag, I.„ An immuno-genetic hybrid algorithm'',Int J Comput Commun Control, 4(4), 374-385.

[Network & others, 2014]. Network, C. G. A. R. & others „Comprehensive molecular profiling of lung adenocarcinoma''. *Nature*, *511*(7511), 543–550.

[Szklarczyk, Franceschini, Kuhn, Simonovic, Roth, Minguez, 2010]. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., … others. , „ The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored'', *Nucleic Acids Research*, gkq973.

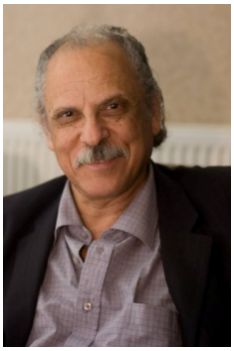
[Vafaie & De Jong, 1992]. Vafaie, H. & De Jong, K. „ Genetic algorithms as a tool for feature selection in machine learning. In *Tools with Artificial Intelligence.'', 1992, TAI'92, Proceedings., Fourth International Conference on* (pp. 200–203).

[Waterland & Michels, 2007]. Waterland, R. A. & Michels, K. B. „ Epigenetic epidemiology of the developmental origins hypothesis.'', *Annu . Rev . Nutr.*, *27*, 363–388.

[Xu & Zhang, 2005]. Xu, X. & Zhang, A. , „Selecting informative genes from microarray dataset by incorporating gene ontology.'' , In Bioinformatics and Bioengineering, 2005, BIBE 2005, Fifth IEEE Symposium on , (pp. 241–245).

[Zhang, Lu, Shi, Xu, Hon-chiu, Harris, Wong, 2006]. Zhang, X., Lu, X., Shi, Q., Xu, X., Hon-chiu, E. L., Harris, L. N., … Wong, W. H. , „Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data.'', BMC Bioinformatics, 7(1), 1.

Author's information



Abdel-Badeeh M. Salem - *Professor of Computer Science, Head of Artificial Intelligence and Knowledge Engineering Research Labs,*

Faculty of Computer and Information sciences,

Ain Shams University, Cairo, Egypt .

e-mail: abmsalem@yahoo.com

Mohamed Gawesh - *Teacher assistant of faculty of computer and information science,*

Ain shams university, Cairo, Egypt.

e-mail: mygawish@cis.asu.edu.eg

Mohammed Maher - *Bsc student at Department of Bioinformatics*

Faculty of Computer and Information Sciences,

Ain Shams University, Cairo, Egypt.

e-mail: mohammed.maher2013@hotmail.com

Yasmin Amr - *Bsc student at Department of Bioinformatics*

Faculty of Computer and Information Sciences,

Ain Shams University, Cairo, Egypt.

e-mail: yasminamrfcis@gmail.com

Amany Hussein *- Bsc student at Department of Bioinformatics*

Faculty of Computer and Information Sciences,

Ain Shams University, Cairo, Egypt.

e-mail: amanyhusseinhasan@hotmail.com

Keryakous Zarif *- Bsc student at Department of Bioinformatics*

Faculty of Computer and Information Sciences,

Ain Shams University, Cairo, Egypt.

e-mail: keryakous.zarif@hotmail.com

Basant Mohamed *- Bsc student at Department of Bioinformatics*

Faculty of Computer and Information Sciences,

Ain Shams University, Cairo, Egypt.

e-mail: basant.mohamed.ramdan@fcis.asu.edu.eg

Hebatullah Mohamed *- Bsc student at Department of Bioinformatics*

Faculty of Computer and Information Sciences,

Ain Shams University, Cairo, Egypt.

e-mail: hebafouad.cs2016@gmail.com