Genes & Genomics
https://doi.org/10.1007/s13258-018-0773-2

Online ISSN 2092-9293
Print ISSN 1976-9571

RESEARCH ARTICLE

Identification of tissue-specific tumor biomarker using different optimization algorithms

Shib Sankar Bhowmick[1] · Debotosh Bhattacharjee[2] · Luis Rato[3]

Received: 16 April 2018 / Accepted: 3 December 2018

Abstract

Background Identification of differentially expressed genes, i.e., genes whose transcript abundance level differs across different biological or physiological conditions, was indeed a challenging task. However, the inception of transcriptome sequencing (RNA-seq) technology revolutionized the simultaneous measurement of the transcript abundance levels for thousands of genes.

Objective In this paper, such next-generation sequencing (NGS) data is used to identify biomarker signatures for several of the most common cancer types (bladder, colon, kidney, brain, liver, lung, prostate, skin, and thyroid)

Methods Here, the problem is mapped into the comparison of optimization algorithms for selecting a set of genes that lead to the highest classification accuracy of a two-class classification task between healthy and tumor samples. As the optimization algorithms Artificial Bee Colony (ABC), Ant Colony Optimization, Differential Evolution, and Particle Swarm Optimization are chosen for this experiment. A standard statistical method called DESeq2 is used to select differentially expressed genes before being feed to the optimization algorithms. Classification of healthy and tumor samples is done by support vector machine

Results Cancer-specific validation yields remarkably good results in terms of accuracy. Highest classification accuracy is achieved by the ABC algorithm for Brain lower grade glioma data is 99.10%. This validation is well supported by a statistical test, gene ontology enrichment analysis, and KEGG pathway enrichment analysis for each cancer biomarker signature

Conclusion The current study identified robust genes as biomarker signatures and these identified biomarkers might be helpful to accurately identify tumors of unknown origin

Keywords Biomarker · Machine learning tools · Messenger RNA · Optimization algorithm · Pathway analysis

Introduction

Cancer classification based on gene expression data is of great interest in recent years. These gene expression data have been widely used to differentiate cancerous tissue samples from healthy ones as well as to identify different subtypes of cancer (Lapointe et al. 2004; Mramor et al. 2007; Liu et al. 2008). In this regard, the application of high-throughput DNA sequencing technology provides an entirely new platform for cancer classification. Gene expression data generally have high dimensionality with a relatively small number of samples (Furey et al. 2000). Therefore, it becomes a challenging task to design a robust classification algorithm that mines informative genes with definite biological significance. In this regards, a comprehensive review of feature (gene) selection methods has been illustrated by Saeys et al. (2007).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13258-018-0773-2) contains supplementary material, which is available to authorized users.

✉ Shib Sankar Bhowmick
shibsankar.bhowmick@heritageit.edu

[1] Department of Electronics and Communication Engineering, Heritage Institute of Technology, Kolkata 700107, India

[2] Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India

[3] Department of Informatics, University of Evora, 7004-516 Evora, Portugal

Published online: 08 December 2018

Available feature selection methods for high-dimensional data often fall into one of the following three categories: filter, wrapper, and embedded methods. (i) Filter methods are fast in operation and low in computational complexity. They assess the genes according to the inherent features of the expression data. Univariate filter methods include *t*-test (Olopade and Grushko 2001), correlation coefficient (Iyer et al. 1999), signal-to-noise ratio (Golub et al. 1999) etc. Whereas, correlation-based feature selection (Wang et al. 2005), and Markov blanket filter (Han and Liu 2012) are considered to be multivariable filter methods; (ii) wrapper methods (Ooi and Tan 2003), measure the usefulness of a subset of features in the space of all possible feature subsets. Local minima problem is avoided by wrapper method as it performs a search using stochastic approximations that cover a large portion of the feature space; (iii) embedded methods, make use of the internal information in a classification model to perform feature selection. Classifier such as support vector machine (SVM) with radial basis function (RBF) kernel based on recursive feature elimination (SVM–RBF–RFE) (Liu et al. 2011) and random forest (Kandaswamy et al. 2011) are an example of embedded methods.

Based on these three classes of feature selection techniques, various methods have been proposed and evaluated for correctly identifying cancer tumors (Peng et al. 2010; Liu et al. 2010; Chandra and Gupta 2011). In this regard, gene markers like PBGD (circulatory blood) (Haas et al. 2009), TGM4 (semen) (Wobst et al. 2011), HBD1 (vaginal secretion), MMP7 (Fleming and Harbison 2010), KRT4 (oral mucosa) (Richard et al. 2012), STATH (nasal secretion) and CST6 (skin) (Juusola and Ballantyne 2007) plays very crucial role to identify differentially expressed genes. The method developed by (Zhang et al. 2012) exhibits broad generalization in the genes selected using 9 two-class gene expression datasets. In addition to this, markers like (Argani et al. 2001; Wang et al. 2004) are designed for the detection of pancreatic and colon cancers.

Despite the effort, however, these implemented techniques differ according to their modeling performance. Additionally, the informative genes selected by different feature selection methods are very minute overlapped. Therefore, the evaluation of the robustness of feature selection methods deserves more attention (Chopra et al. 2010). In this work, we have searched putative gene biomarkers from a population of the healthy and tumor samples. The problem is mapped into the comparison of optimization algorithms for selecting a set of features that lead to the highest classification accuracy of a two-class classification task between healthy and tumor samples. We have solved this optimization problem by means of Artificial Bee Colony (ABC) (Karaboga and Basturk 2007), Ant Colony Optimization (ACO) (Dorigo et al. 2006), Differential Evolution (DE) (Storn and Price 1997), and Particle Swarm Optimization (PSO) (Kennedy 2011) algorithms. These algorithms iteratively select a subset of differentially expressed genes identified by a standard statistical method called, DESeq2 (Love et al. 2014). Classification of healthy and tumor samples is done using SVM (Boser et al. 1992), and genes responsible for the highest classification accuracy is returned as a candidate panel of biomarkers. In order to prevent irrelevant results to be included in the panel because of the intrinsic randomness of the method, we run our algorithm 50 times and used a majority voting scheme to select the final set of differentially expressed genes. The subset of overlapping genes responsible for highest classification accuracy is considered to be the optimal biomarker signature by different algorithms. For this analysis, next-generation-sequencing (NGS) based messenger RNA (mRNA) datasets of bladder, colon, kidney, brain, liver, lung, prostate, skin, and thyroid cancer are considered. We have investigated the biological role of our method selected genes by performing different experiments like gene ontology (GO) enrichment analysis and KEGG pathway enrichment analysis. Major highlights of this paper are:

- Comparative analysis between the healthy and a tumor group of samples, in order to identify significantly differentially expressed genes across nine cancer types.
- Performance check of the optimization algorithms like the ABC, ACO, DE, and PSO for selecting a set of genes that lead to the highest classification accuracy of a two-class classification task between healthy and tumor samples.
- Variation of modeling performance among the optimization algorithms lead to the selection of the minutely overlapped set of genes irrespective of the datasets.
- Identified genes play a key role in diverse biological processes. Biological significance tests show that most of the identified genes are involved in key oncogenesis pathways.
- The identified biomarker signatures in our experiments might be helpful to accurately identify tumors of unknown origin, as well as the proposed model itself, may be applied to other clinical queries.

The remainder of this article is organized as follows: in "Employed algorithms" section we describe a brief overview of the employed algorithms. The proposed method is described in "Proposed method" section. The dataset, preprocessing as well as results are shown and discussed in "Experimental results" section. Finally, in "Conclusions" section we draw our conclusions.

Employed algorithms

One fundamental idea behind the proposed method is that of finding putative gene biomarkers by means of optimization algorithms. In this regard, prediction provided by ABC,

ACO, DE, and PSO are compared. These algorithms search the best feature subsets or genes by iteratively improving their candidate solutions. All these optimization algorithms are randomized by nature and hence, the different algorithm may select different panels of genes. This randomization effect of the used algorithms is considered here to decrease the probability of returning suboptimal solutions containing false positives and/or false negatives. Although this procedure does not ensure the absence of false positives or false negatives, some considerations can be done.

Artificial Bee Colony

According to ABC (Karaboga and Basturk 2007) paradigm, there are three kinds of honey bees named: employed bees, onlookers and scouts. Conventionally, both the onlookers and the scouts are termed as unemployed bees. Here, the possible solution of the optimization problem lies in the available position of a food source where the nectar amount of a food source represents the quality of fitness. The number of possible solutions in the population represents the number of the employed bees or the onlooker bees. Numerical functions are optimized in three stages according to the ABC algorithm. At first, a random initial population of size N_{cl} (food source positions) is generated. Each solution or food source has $\mathscr{D}$ number of optimization parameters. An employed or onlooker bee probabilistically change the possible solution in her memory for finding a new food source and tests the nectar amount or the possible fitness value of the new source i.e., the new solution. This nectar information of the food source (solutions) and the position of the food sources are being shared by the employed bees with the onlooker bees on the dance area. If the nectar information is higher than that of the previous one, then the bee memorizes the new position and forgets the old one. The second stage starts with the updation of onlookers where the food sources are selected according to the probability $P_i = fit_i / \sum_{n=1}^{N_{cl}} fit_n$. Here fit_i denotes the fitness value of the i-th solution in the population. During the update process, a new candidate solution is firstly given by the following solution search equation:

$$\mathscr{Z}_{ij} = x_{ij} + \boldsymbol{\Phi}_{ij}(x_{ij} - x_{kj}) \tag{1}$$

where x_{ij} (or $\mathscr{Z}_{ij}$) denotes the jth element of x_i (or $\mathscr{Z}_i$), and j is a random index, $j \in \{1, 2, \ldots, \mathscr{D}\}$. x_k denotes another solution selected randomly from the population where $k \in \{1, 2, \ldots, \mathscr{E}_b\}$. $\boldsymbol{\Phi}_{ij}$ represents a uniform random number in $[-1, 1]$ and $\mathscr{E}_b$ corresponds to the number of employed bees. The update process is completed here by a greedy selection between x_i and Z_i. If the new food source has at least as much nectar as the old one, it replaces this latter in the memory. According to the algorithm, every solution of the employed bee is involved in the update process, while only the selected solutions have the opportunity to be updated by the onlookers. This is the third important stage which differentiates the employed bee and the onlookers. Moreover, an inactive solution of the scout bee refers to a solution that does not change over a certain number of generations.

The ABC and SVM are used here to select a feature of importance. It can effectively find potential genes that can be treated as biomarkers.

Ant Colony Optimization

ACO (Dorigo et al. 2006) algorithm is mainly applied to optimization problems, and generally consists of four main steps: initialization, construct ant solutions, local search, and global update pheromones. During the first step, all the parameters are initialized and pheromone variables are initialized to a value τ_0. Subsequently, during the construct ant solutions step, each ant begins with an empty solution $s_p = \varnothing$. Moreover, a set of m ants construct the initial solution, and during the process, an ant chooses one feasible solution component at each construction step, $c_i^j \in N(s_p) \subseteq C$. In this way, it upgrades its current partial solution. Here, $N(S_p)$ represents a set of solution component, defined mainly by an implemented solution construction process. In this regard, meaningfully in-feasible partial solutions during the construction mechanism are penalized, depending on the violation of the problem constraints. At each construction step, a probabilistic method is used to choose the solution component. One of the most commonly used by ACO is described below:

$$p(c_i^j | s_p) = \frac{\tau_{ij}^{\alpha}.[\eta(c_i^j)]^{\beta}}{\sum_{c_i^l \in N(S_p)} \tau_{il}^{\alpha}.[\eta(c_i^l)]^{\beta}}, \forall c_i^j \in N(S_p) \tag{2}$$

Here, $\alpha = 0$ corresponds to the selection probabilities that is proportional to $[\eta_{ij}]^{\beta}$. Generally, a high heuristic solution component is selected whereas, $\beta = 0$ represents the pheromone amplification at work. Here, the local search algorithm step is used in a problem specific manner to improve the complete candidate solution further that cannot be enhanced by individual ant. The pheromone update is implemented here to make the desired solution components for the next iteration. Generally, a mechanism called *pheromone deposit*, and *pheromone trial evaporation* is used for updating the pheromone information. During the pheromone deposit operation, the level of pheromone of a chosen set of solution component S_{upd} is increased. Pheromone trial evaporation decreases the level of pheromone deposited over time by the previous ants. This process is necessary to avoid a rapid convergence of the algorithm to a suboptimal region. The pheromone is updated as follows:

$$\tau_{ij} = (1-\rho)\tau_{ij} + \sum_{s \in S_{upd} | c_i^j \in s} g(s) \tag{3}$$

Here, $g(s)$ is called the evaporation function and S_{upd} represents the set of solutions used for depositing pheromone. Typically, based on the pheromone update mechanism ACO algorithm differs. Various way of determining the S_{upd} results in different pheromone update mechanism.

In the context of gene selection, an ACO algorithm has the ability to perform a flexible and robust search for a good combination of genes.

Differential Evolution

DE (Storn and Price 1997) searches for a global optimum solution in a D-dimensional real parameter space $\mathcal{R}^D$. DE is an evolutionary optimization tool that has a wide range of applications. Here, DE is used to encode features. As in any other evolutionary algorithm, population initialization in DE begins with a randomly initiated population of N_{ip} vectors. Each vector often denoted as a chromosome, forms a candidate solution to the optimization problem. The i-th individual vector of the population at time-step (generation) t has L components (dimensions), chosen randomly from the set $1, 2, \ldots, D$, where D is the total number of features, i.e.,

$$\mathcal{G}_i(t) = [\mathcal{G}_{i,1}(t), \mathcal{G}_{i,2}(t), \ldots, \mathcal{G}_{i,L}(t)] \tag{4}$$

The operation of mutation is employed by DE for producing a mutant vector concerning each parent vector. Here for every vector in the current population, three other vectors are selected to perform arithmetic mutation. Actually, using this process mutually exclusive values are randomly generated within the range $[1, L]$. In other words, the L-th component of each trial vector is generated as follows:

$$\mathcal{V}_{i,L}(t+1) = \mathcal{G}_{j,L}(t) + F(\mathcal{G}_{n,L}(t) - \mathcal{G}_{m,L}(t)) \tag{5}$$

Here, F is a mutation factor. In order to enhance the potential diversity of the population, a uniform crossover has been applied after generating the mutant vector. Here, crossover rate, CR is used as a user-specified constant within the range [0, 1], which controls the fraction of parameter values to be copied from the mutant vector into the trial vector. The rest of the parameters of the trial vector is taken from the corresponding parent vector.

$$\mathcal{Q}_i(t+1) = [\mathcal{Q}_{i,1}(t+1), \mathcal{Q}_{i,2}(t+1), \ldots, \mathcal{Q}_{i,L}(t+1) \tag{6}$$

where

$$\mathcal{Q}_{i,j}(t+1) = \left\{ \begin{array}{ll} \mathcal{V}_{i,j}(t+1), & if\ rand_j(0,1) \le CR,\ or\ j = rand(i) \\ \mathcal{G}_{i,j}(t), & if\ rand_j(0,1) > CR,\ and\ j \ge rand(i) \end{array} \right\} \tag{7}$$

In Eq. 7, $rand_j(0, 1)$ is the j-th evaluation of a uniform random number generator with outcome $\in [0, 1]$. $rand(i)$ is a randomly chosen index $\in \{1, 2, \ldots, L\}$, which ensures that $\mathcal{Q}_i(t+1)$ gets at least one parameter from $\mathcal{V}_i(t+1)$.

During the selection process of the current population, the objective function value of each trial vector is compared to that of its corresponding parent vector. If the new trial vector yields an equal or lower value of the objective function, then the corresponding parent vector is replaced in the next generation. Otherwise, the parent is retained in the population. Hence, the population never deteriorates, either get better or remains the same in fitness status. The next generation is represented as follows:

$$\mathcal{G}_i(t+1) = \left\{ \begin{array}{ll} \mathcal{Q}_i(t+1), & \text{if } f(\mathcal{Q}_i(t+1)) > f(\mathcal{G}_i) \\ \mathcal{G}_i(t), & \text{if } f(\mathcal{Q}_i(t+1)) \le f(\mathcal{G}_i) \end{array} \right\} \tag{8}$$

where $f(\cdot)$ is the objective function to be maximized. Finally, elitism kept the best vector of the current population for the next iteration, based on its objective function value. The above-mentioned processes are repeated for a given number of generations until stopping criteria are met.

For this experiment, the goal of DE is to find the subset of genes that maximizes classification accuracy with the help of SVM.

Particle Swarm Optimization

PSO (Kennedy and Eberhart 1995) is a simple, robust and effective optimization technique. According to PSO, a population of N_{par} candidate solutions (called a swarm) is represented as particles P_i, where $i = \{1, 2, \ldots, N_{par}\}$. The elements of a particle are called positions and the length of the particle are denoted with L. The velocity (V_i) and position of each particle is updated during its movement in the search space as following equations (Shi and Eberhart 1998):

$$V_i^{(t+1)} = \omega \times V_i^{(t)} + \varphi_1 \times (P_{l_{best}}^{(t)} - P_i^{(t)}) + \varphi_2 \times (P_{g_{best}}^{(t)} - P_i^{(t)}) \tag{9}$$

$$P_i^{(t+1)} = P_i^{(t)} + V_i^{(t+1)} \tag{10}$$

where t represents a time stamp of different iteration, w is the inertia weight $\in [0.5, 1]$, φ_1 and φ_2 are the cognitive and social constants. Similarly, $P_{l_{best}}$ particle represents local best of the current iteration, while global best particle (till the current iteration) is $P_{g_{best}}$. The PSO algorithm terminates after a predetermined number of iterations (N_{itr}).

For this experiment, a swarm is prepared by considering the N_{par} number of particles where each particle having L gene indices selected randomly from the preprocessed dataset of gene. The encoded swarm is used to compute the fitness function with the help of a SVM classifier in 5 Fold Cross Validation mode.

Proposed method

A cohort of 4127 tumor patients are divided into nine different tissue types is considered for this experiment. Our input consists of a matrix where each row corresponds a gene and columns are the samples. In addition, we know the class label of each sample. According to this input description, our problem reduces to that of finding a subset of significantly differentially expressed genes, called gene markers. Steps of the proposed method are described below:

Differential expression analysis

In gene expression analysis, a fundamental task is the analysis of read counts per gene in RNA-sequence to measure the systematic changes across experimental conditions. For this analysis, a method called DESeq2 (Love et al. 2014) is considered here, that measure the differential analysis of count data, using shrinkage estimation for dispersions and logarithmic fold changes. The use of DESeq2 enhances the quantitative analysis of comparative RNA-sequence data by integrating methodological advances with several novel features. However, DESeq2 only uses the raw counts of data and does not actually use normalized counts. The differential analysis assumes the null hypothesis that the logarithmic fold change between the healthy and tumor samples for a gene's expression is exactly equal to zero, i.e., genes are not affected by treatment. Moreover, differential expression analysis produces a list of genes passing multiple test criteria, ranked by adjusted *P*-value.

Optimization

In this study, we have considered a set of four optimization methods to handle the model. Published performance (Karaboga et al. 2014; Dorigo et al. 2008) and the results for a set of benchmark problems (Abu-Mouti and El-Hawary 2011; Dorigo and Stützle 2003; Cai et al. 2008; Eberhart and Shi 2001) have motivated us to select ABC, ACO, DE, and PSO algorithms for the optimization problems. Although these methods do not ensure the absence of a sub-optimal solution, the researcher can solve a given problem using different optimization methods and compare the outcomes to reach a final decision. Usually, all these methods converge to the best solution. Moreover, the results cannot be treated as a global optimum.

The used optimization techniques require an initial population that is named differently. For example, in ABC, colony size is chosen as N_{cl} whereas the number of the particle for ACO as N_{par}, the initial population for DE as N_{ip} and a swarm is prepared in PSO by considering the N_{par} number of particles. Index encoding is used here to prepare the initial population. For this analysis, L number of gene indices are selected randomly from the preprocessed differentially expressed genes. Here, the value of L is chosen very small in order to make the classifier robust. For our experiment, L is considered as 20. These L genes are presented as attributes for $\mathscr{P}$ number of patients. Therefore each particle, colony or population, is made up of a distinct dataset of size $\mathscr{P} \times L$. Here, ABC, ACO, DE, and PSO finds the best performing feature subsets or genes by iteratively improving their candidate solutions. The algorithm terminates after N_{itr} iterations.

Fitness computation

The objective of a fitness function is to quantify the quality of solution of the optimization algorithms. In fact, we are interested in a function that maximizes the chances of identifying differentially expressed genes among the healthy and tumor samples. This, in turn, can be seen as a two-class classification problem. In this regard, a SVM classifier with RBF kernel is used for classification. The distinct datasets of size $\mathscr{P} \times L$ are passed to the SVM for computing the fitness function based on classification accuracy. Aimed at improving the stability of the method, we have applied tenfold cross-validation to classification. Block diagram of the proposed method is presented in Fig. 1.

Optimal signature

The used optimization algorithms are randomized by nature. Different runs of the algorithms may select different sets of a gene, hence, running this algorithm once may not be a good idea. Therefore, we have set the number of iterations to be as $N_{itr} = 50$. In order to find an optimal signature, a selection strategy based on majority voting is introduced here. According to that selection mechanism, the set of genes selected by an optimization algorithm after N_{itr} iterations are represented as $M = \{m_1, m_2, \ldots, m_{itr}\}$ and m_{apr} be the number of times a gene appear after 50 runs. Thereafter, the selected genes are sorted in decreasing order of their magnitude such that $m_{apr_i} \geq m_{apr_j}$ if $i < j$. As a result, most appeared gene tops the list. Finally, for increasing value of $i \in [1, itr]$ we make a new subset of gene and classify them using SVM classifier. The subset of a gene responsible for highest classification accuracy is considered as the optimal signature.

Biological validation

Biological validation of the optimal biomarker signature is done by means of KEGG pathway enrichment analysis and GO enrichment analysis. For KEGG pathway enrichment

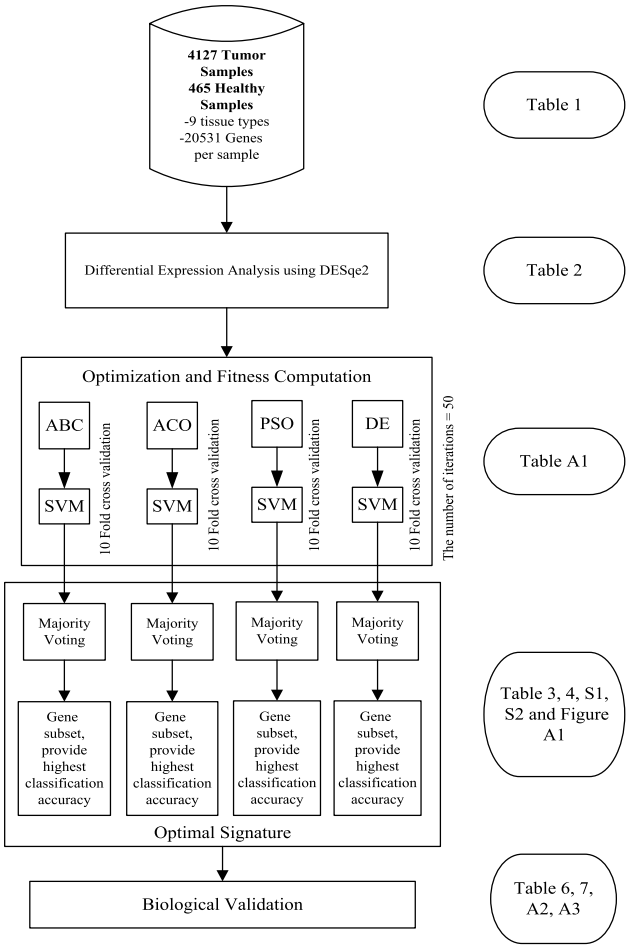

Fig. 1 Block diagram of the proposed workflow

analysis, the overlap between the known KEGG pathways and constructed protein-protein interaction (PPI) networks, is searched to find out PPI enriched KEGG pathways. GO enrichment analysis is used to perform the enrichment analysis on gene sets.

Experimental results

Datasets

In this work, NGS based mRNA expression datasets from The Cancer Genome Atlas (TCGA)[1] have been chosen (TCGA data version 2016.01.28). Expression levels of 9 of the most diffused cancer types (*Bladder, Colon, Kidney, Brain, Liver, Lung, Prostate, Skin*, and *Thyroid*) are studied in our experiment. TCGA exploits the illumina technology for sequencing and provides the expression data in the form of raw read count (RRC). A patient in TCGA is represented by a barcode-like 'TCGA-4H-AAAU-01A-11R-A41B-07'.

Table 1 Details of TCGA data used in the experiments

Disease	Code	Tumor samples
Bladder urothelial carcinoma	BLCA	408
Colon adenocarcinoma	COAD	287
Kidney renal clear cell carcinoma	KIRC	534
Brain lower grade glioma	LGG	530
Liver hepatocellular carcinoma	LIHC	373
Lung adenocarcinoma	LUAD	517
Prostate adenocarcinoma	PRAD	498
Skin cutaneous melanoma	SKCM	471
Thyroid carcinoma	THCA	509
Total		4127

Each tumor class is coupled with 465 samples of healthy class

Different fields of this barcode consist of a collection of identifiers that are specifically decoded to identify the tumor and healthy samples. For our convenience, we have transformed the expression values into log_2 scale. Table 1 summarizes the datasets and provides the information on tumor samples. In order to eliminate the biases due to a limited number of healthy samples during the experiment, we have grouped them together regardless of the cancer type obtaining a total of 465 healthy samples. Moreover, the Cancer Genome Atlas provides the expression level of 20,531 genes for the chosen cancer types, belonging to a variable number of samples. While dividing the datasets into training and test sets, it has been ensured that each of these sets had an equal proportion of healthy or cancerous samples.

Results

The main objective here is to find putative gene biomarkers from a population of a healthy and tumor samples. For this analysis, disease status for the 4592 (4127 tumors and 465 healthy) individuals with gene expression data are investigated. Significant changes in gene expression profiles between the healthy and tumor samples are considered to be the underline criteria to identify differentially expressed genes, $\mathscr{DE}$. In this regards, absolute logarithmic fold change (FC) value > 1.5 and an *adjusted p*-value of < 0.05 are considered to be the underline criteria by DESeq2 to identify differentially expressed genes. Details of the differentially expressed gene counts are summarized in Table 2. The test result shows, DESeq2 consistently identified near-about 9000 (minimum of 8744 for KIRC and maximum of 9987 for PRAD dataset) differentially expressed genes independent from the cancer type.

Subsequently, during the next step of our method globally differentially expressed ($\mathscr{DE}_g$) genes are identified from

[1] http://gdac.broadinstitute.org/.

Table 2 Selection of the number of genes at different processing steps of the proposed method

Tissue	DESeq2 selected genes	Optimization method selected genes				Optimal signature genes			
		ABC	ACO	DE	PSO	ABC	ACO	DE	PSO
BLCA	9343	944	931	942	942	12	7	15	11
COAD	9376	936	930	917	934	8	15	12	8
KIRC	8744	941	919	920	931	13	15	8	11
LGG	9585	932	948	942	942	9	8	10	7
LIHC	9488	936	949	931	947	17	7	13	15
LUAD	9825	947	951	941	946	13	12	13	12
PRAD	9987	952	942	959	947	14	14	18	12
SKCM	9354	946	954	921	943	11	11	15	11
THCA	9116	929	949	912	938	12	10	8	19

these initial filtered $\mathscr{DE}$ genes. For that, ABC, ACO, DE, and PSO optimization algorithms in conjunction with an SVM classifier are used to enforce the evidence of the presence/absence of cancer. In particular, we have evaluated the individual contribution of each of our chosen optimization algorithm in the gene selection process. During the comparison, we have used K-fold cross-validation (K = 10) by fixing the SVM parameters. The choice of K must take into account the bias-variance tradeoff because large optimistic bias leads to over-fitting. Increasing K, reduces the bias but might increase variance to the point of uselessness and too small K, like twofold cross-validation, also has a large variance. 10-fold is usually considered a good compromise and sufficient to minimize any over-fitting issue. Therefore, in our analysis 10-fold cross-validation is chosen. In Online Appendix Table A1, 6 of the most frequently used classification measures (namely: Accuracy, Precision, Sensitivity, Specificity, F-measure, and Matthews correlation coefficient (MCC)) are reported for each cancer type. Results indicate that all the used optimization methods achieve a satisfactory classification accuracy. Moreover, ABC has an average performance better than the other methods in all the tested datasets.

Optimal gene signature

Different run of the optimization algorithms is likely to produce a slightly different panel of genes. In order to create a stable and reliable panel in which the probability of false positives or false negatives is minimized, we run the optimization algorithms 50 times. In fact, our experiment shows that running optimization algorithm 50 times, return heavily overlapping sets of genes. Thereafter, majority voting schema is used to select stable group. Although this procedure does not ensure the absence of false positives or false negatives, some considerations can be done. Hence, we have sorted the overlapping set genes in decreasing order of their appearance. Thus, most appeared gene by a selection process tops the list. The same set of a list is prepared for all optimization algorithms belonging to different cancer types (see Table S1 in the supplementary material for details). Next, for an increasing number of genes in that list, a new subset is prepared (taking a minimum of 2 genes and maximum of m_{itr} genes) and classified using a SVM classifier. The classification accuracy of the different subset of genes is reported in Table S1 in the supplementary. According to that result, the subset of a gene responsible for highest classification accuracy is considered as the optimal gene signature corresponding to the different optimization algorithm. For eg., in BLCA data for ABC algorithm, the subset of genes *KAT2B, SGCE, SLC35B3, ACO1, ACOT1, ANKHD1, APBA1, ARHGAP31, ARMCX1, BBS7, C3orf38, and C9orf82* are responsible for highest classification accuracy, hence considered as the optimal biomarkers (classification accuracy 91.40%). The final list of our method selected optimal biomarkers for different algorithms and datasets can be found in Table 3. Logarithmic fold change (FC), *adjusted* p-value, and appearance (m_{apr}) information of the optimal genes are reported in Table S2 of the supplementary and all these information signifies the importance of optimal genes.

We have further investigated the relationships among the identified overlapping sets of a gene by different optimization algorithms. Results reported in Fig. A1 of Online Appendix, confirm that informative genes selected by different optimization algorithms are minutely overlapped irrespective of used datasets. Variation of modeling performance among the optimization algorithms could be the major cause of these results. Hence for this experiment, best performing subset of genes from all four optimization algorithms are considered to be the optimal biomarkers in different cancer data. Selected gene counts at different processing steps are depicted in Table 2.

Statistical analysis

As the optimal signatures are predicted by different optimization algorithms, i.e., ABC, ACO, DE, and PSO. Hence, the measurement of significant contributions of these

Table 3 Our method selected optimal gene signatures

Rank	BLCA				COAD				KIRC			
	ABC	ACO	DE	PSO	ABC	ACO	DE	PSO	ABC	ACO	DE	PSO
1	KAT2B	IFRD1	CPVL	PECI	NUP62	ADORA2A	ATP2A2	RHBDF2	APEH	MTMR10	C7orf42	ENG
2	SGCE	IQGAP2	PPIP5K1	AFMID	RRP9	DTNBP1	CD99	ZNF83	ASPSCR1	NUP155	RABEP1	EPHB2
3	SLC35B3	BOC	SLC8A1	ANKS1A	LOC652276	PNPT1	FAM98B	ANKRD27	CSTF2T	PRDM2	AIMP2	LOC440957
4	ACO1	C10orf58	TANC1	ANXA2	MRM1	POLR2H	IQGAP3	ARHGAP26	FOXRED1	PRKCD	AK1	AAAS
5	ACOT1	C17orf53	AGK	ARHGAP10	PRMT2	SLC35A2	ITCH	ARSB	SEPHS2	SIRT6	ARHGAP10	ACN9
6	ANKHD1	C1RL	AMACR	ASF1B	TMEM185A	ALDH18A1	TH1L	C12orf51	AGPAT2	ALAD	ARHGAP19	ADO
7	APBA1	CCDC50	AMFR	BCAP29	ARHGAP9	ALDOA	UCHL3	C20orf20	C19orf63	AP3B1	ASS1	AFAP1L1
8	ARHGAP31		APPL2	C15orf57	ATP1B3	ASH1L	ACSF3	CCDC109A	C1orf21	ATL3	ATP11A	AGPS
9	ARMCX1		ARFGAP2	C1QTNF6		C2CD2	ACTR1B		CCDC85C	BCAT1		ALS2CL
10	BBS7		ARHGEF2	C20orf3		C9orf167	BAIAP2L1		CCT7	BFAR		ARFRP1
11	C3orf38		ARRDC3	C3		C9orf3	BBC3		COX4NB	C1orf122		ASH2L
12	C9orf82		ASF1B			COQ10A	BDH1		DOCK1	CCDC109A		
13			ATP9A			DHX35			EHD3	CHD2		
14			CEP120			DLG3				CMBL		
15			CFL1			FBXL12				CRADD		
Rank	LGG				LIHC				LUAD			
	ABC	ACO	DE	PSO	ABC	ACO	DE	PSO	ABC	ACO	DE	PSO
1	CTNNAL1	ADC	CORO2A	PANK1	MPST	CDC25B	LCAT	DIS3	GYPC	PTEN	ABLIM3	AK3
2	PIGB	AKAP8	PLK2	AES	ABCF3	ABCD3	NIPSNAP1	DPH2	PTRF	ACAD9	CEBPA	ATP5G3
3	TSPAN7	AKTIP	SHROOM3	ANKRD39	ACADSB	ADSS	ABHD3	KDM5A	ACACA	ALDH3A2	CYTH1	CD93
4	ABAT	ALG5	SRRM2	ANO8	ALS2CR4	ALG3	AES	ANKRD13C	ACADS	ALG6	ABCG1	ADAMTS9
5	AFF1	AP1S3	AGAP2	ARF4	ARSD	ATAD3A	ALDH2	AP1G1	ALDH6A1	C17orf48	AKAP9	ADRBK2
6	AHNAK	BCAP29	ALPK1	ARRDC1	C12orf72	ATPAF1	ANKRD13C	ARID2	ATP6V1D	C8orf76	ARAP1	AFAP1L2
7	APOL3	BCL2L11	ALPL	BAP1	C6orf145	C10orf88	ARID3B	ASH2L	BAIAP2	C9orf130	C12orf43	ALAD
8	ARHGEF11	C17orf100	APOBEC3G		C8orf55		ARRDC2	ATP5O	BCL2L11	CFL2	C14orf129	ARRDC4
9	ATP13A2		AQP3		CRIPT		ARSK	C1orf85	C13orf27	DDX23	C14orf139	ASB13
10			ARHGAP33		DPP7		ASAP3	C4orf42	C19orf12	ELOVL6	C9orf45	ATP5F1
11					EPS8		BLMH	CCDC55	C1S	EXOSC10	CAB39L	AVL9
12					ERLIN1		BLOC1S3	CERCAM	CAB39L	GOLGA1	CAD	C1orf112
13					FKBP2		C17orf65	CLEC11A	CAT		CALM2	
14					GART			COMMD4				
15					GMPPA			CSRNP2				
16					GPAM							
17					GRK4							

Table 3 (continued)

Rank	PRAD				SKCM				THCA			
	ABC	ACO	DE	PSO	ABC	ACO	DE	PSO	ABC	ACO	DE	PSO
1	PVRL2	MICALL1	CD302	NSFL1C	DHRS11	LYAR	OBFC2B	PNPLA2	MFHAS1	PRMT3	C16orf88	LBR
2	SMPD1	RAD54L2	HMGCR	PYCR1	SLC7A2	AATF	ANKRA2	A4GALT	RAB27A	ADIPOR2	DLG4	BCL9L
3	AMT	AIMP2	KLHL18	ANAPC2	AK2	ADRM1	EXOC6	AGPS	ARHGEF12	AKTIP	IDH2	CDCA4
4	ANXA2	ALKBH3	SAPS2	ANKS1A	ARL8A	APOE	GRK5	APOA1BP	ARPC3	ANKRD37	KLHL22	METAP1
5	C19orf42	AMZ2	ACOT8	ANO6	ATP8B1	AVEN	HHAT	ATG2A	C6orf125	AP1B1	LIMK1	ADAP2
6	CDCA5	AP4M1	ANXA5	AP1G1	C10orf26	AZI2	PEMT	C1orf123	C6orf134	APP	POLK	ANKH
7	CEMP1	BPNT1	BCAS3	B3GNT7	C10orf58	C1orf57	AGPS	CACHD1	CALCRL	ARHGAP29	RIMS3	ARMCX2
8	CNOT8	C1orf93	BRD8	BCAM	C1S	CCNB1	AHNAK	CAMSAP1L1	CDR2L	BCL7A	ARL11	BCAS3
9	CSNK1G2	C5orf42	C12orf49	C10orf2	C9orf103	CCNY	ARFIP2	CD63	CHCHD3	BEX4		BEX4
10	CXorf40B	CLIC1	C19orf63	C11orf17	CABLES2	CHST11	C10orf32	CGNL1	COPS8	C11orf49		BPTF
11	DOM3Z	CRYZ	C1QC	C14orf166	COL6A2	CRLS1	CBARA1	CHPF2	COTL1			C10orf26
12	ETV5	CTSA	DCAF15	C17orf37			CCDC80		CTSB			C6orf145
13	FAAH	DHX40	ENTPD6				CEP135					CALCRL
14	FAHD2A	DNAJC19	FLJ10038				CEP170					CASZ1
15			GTPBP5				CHERP					CATSPER2
16			KIAA0195									CCDC102A
17			LRIG1									CCDC21
18			LRTOMT									CD59
19												CDC37L1

Table 4 Friedman test ranks of the used optimization algorithms

Tissue	ABC	ACO	DE	PCO
BLCA	1	3	2	4
COAD	1	3	2	4
KIRC	1	3	2	4
LGG	1	2	3	4
LIHC	1	3	2	4
LUAD	1	3	2	4
PRAD	1	3	2	4
SKCM	1	3	2	4
THCA	1	3	2	4
Sum of ranks	9	26	19	36

Table 5 Parameters used in the experiments

Method	Symbol	Value	Description
ABC	N_{cl}	50	Colony size
	N_{itr}	50	Number of iterations
	a_{ABC}	1	Acceleration coefficient
			Upper bound
ACO	N_{par}	50	Number of particles
	N_{itr}	50	Number of iterations
	$\mathcal{Q}$	0.5	Intensification factor
	zeta	1	Deviation–distance ratio
DE	N_{ip}	50	Initial populations
	N_{Gr}	50	Number of generations
	$N_{g_{cp}}$	0.8	Cross over probability
PSO	N_{par}	50	Number of particles
	N_{itr}	50	Number of iterations
	φ_1	2	Cognitive constant
	φ_2	2	Social constant
	IW_{max}	0.9	Max inertia weight
	IW_{min}	0.4	Min inertia weight
SVM	$\mathscr{C}$	0.01	SVM $\mathscr{C}$ constant
	N_{exe}	50	Number of executions

algorithms is important to our research. For that analysis, a non-parametric statistical significance test called the Friedman test (Friedman 1937) at 5% significance level is chosen. According to that test outcome in Table 4, the sum of ranks given by Friedman test corresponding to ABC, ACO, DE, and PSO are 9, 26, 19, and 36 respectively. Since ABC achieve the lowest rank among the used algorithms hence, it has an advantage with statistical test ranking. The ranks reveal average Chi-square value and a corresponding p-value of 25.93 and 0.00000986, respectively. Therefore, it indicates the acceptance of an alternative hypothesis, i.e., among the used optimization algorithms, ABC comes out to be the superior.

Details of the parameters used in this experiment are summarized in Table 5. Best practices from the literature are considered for choosing most parameters while problem specific experimental evaluation is also considered in some cases. The RBF kernel used by SVM is controlled by means of two parameters: γ, and the trade-off between training error and margin $\mathscr{C}$. We set $\gamma = 0.5$ and $\mathscr{C} = 2.0$.

Biological significance

The biological significance of the optimal signature genes are examined in terms of GO enrichment analysis to identify the different biological process that is associated with those genes and KEGG pathway enrichment analysis to find out the associated pathways of the informative genes.

Gene ontology analysis

GO enrichment analysis of the optimal biomarkers in different cancer data are performed via GO consortium (Ashburner et al. 2000). Here in Table 6, significant functionally enriched GO Biological processes related to the genes of our panel are reported, while that for Celular component and Molecular function are mentioned in Table A2 and A3 in Online Appendix, respectively. Results shows *GO:0005737 Cytoplasm* and *GO:0044444 Cytoplasmic part* are the most stressed cellular components. This activity is more likely due to the part of all the contents of a cell excluding the plasma membrane and nucleus. Other GO terms including *GO:0071840 Cellular component organization or biogenesis* involve in a process that results in the biosynthesis of constituent macromolecules, assembly, arrangement of constituent parts, or disassembly of a cellular component, *GO:0051130 Positive regulation of cellular component organization* involved in the formation, arrangement of constituent parts, or disassembly of cell structures, including the plasma membrane and any external encapsulating structures such as the cell wall and cell envelope, *GO:0044424 Intracellular part* is a part of the living contents of a cell. In eukaryotes, it includes the nucleus and the cytoplasm. As the associated GO terms are related to cell cycle regulation hence, our selected panel of genes is considered to be significant for cancer diagnosis. The present study on enrichment analysis may provide a basis for the improved understanding of the GO enrichment analysis corresponding to bladder, colon, kidney, brain, liver, lung, prostate, skin and thyroid cancer.

KEGG pathway enrichment analysis

In order to perform the KEGG pathway analysis of the $\mathscr{DE}_g$ genes of our panel, Enrichr (Kuleshov et al. 2016) tool is chosen. Enrichr computes the overlap between known KEGG pathways and Protein-Protein-Interaction (PPI)

Table 6 Most significant GO terms associated with the optimal signature genes for biological process are obtained through enrichment analysis

GO biological process	BLCA	COAD	KIRC	LGG	LIHC	LUAD	PRAD	SKCM	THCA
GO:0071840 Cellular component organization or biogenesis	✓	✓	✓			✓	✓		
GO:0043482 Cellular pigment accumulation		✓		✓	✓		✓	✓	
GO:0051234 Establishment of localization	✓	✓	✓	✓					✓
GO:0006996 Organelle organization	✓	✓	✓		✓				✓
GO:0006082 Organic acid metabolic process	✓				✓	✓	✓	✓	
GO:0043436 Oxoacid metabolic process		✓			✓	✓	✓	✓	
GO:0043476 Pigment accumulation		✓		✓	✓		✓	✓	
GO:0043547 Positive regulation of GTPase activity	✓	✓	✓	✓					✓
GO:0051130 Positive regulation of cellular component organization			✓			✓	✓	✓	✓
GO:0044093 Positive regulation of molecular function	✓	✓	✓	✓					✓
GO:0043087 Regulation of GTPase activity	✓	✓	✓	✓					✓
GO:0044282 Small molecule catabolic process		✓		✓	✓	✓	✓		
GO:0044281 Small molecule metabolic process		✓			✓	✓	✓	✓	
GO:0006810 Transport	✓	✓	✓	✓					✓

Table 7 Most common KEGG pathways associated with the optimal signature genes of our cancer panels

KEGG Pathway	BLCA	COAD	KIRC	LGG	LIHC	LUAD	PRAD	SKCM	THCA
hsa01210: 2-Oxocarboxylic acid metabolism	✓		✓						✓
hsa01100: Metabolic pathways			✓	✓	✓	✓		✓	
hsa04360: Axon guidance						✓			✓
hsa01230: Biosynthesis of amino acids		✓	✓						
hsa01040: Biosynthesis of unsaturated fatty acids	✓	✓		✓		✓			
hsa00650: Butanoate metabolism									✓
hsa00020: Citrate cycle (TCA cycle)	✓				✓	✓			
hsa00071: Fatty acid degradation	✓				✓	✓			
hsa04068: FoxO signaling pathway				✓		✓			
hsa00564: Glycerophospholipid metabolism					✓			✓	
hsa00340: Histidine metabolism					✓	✓			
hsa04142: Lysosome							✓		✓
hsa00670: One carbon pool by folate					✓		✓		
hsa00770: Pantothenate and CoA biosynthesis			✓	✓					
hsa00120: Primary bile acid biosynthesis	✓						✓		
hsa00640: Propanoate metabolism				✓		✓			
hsa00920: Sulfur metabolism					✓		✓		
hsa00280: Valine, leucine and isoleucine degradation					✓	✓			
hsa00410: beta-Alanine metabolism				✓		✓			

networks for the input set of genes. According to this analysis, lower *p*-values represent the higher probability of the pathway to be enriched with the set of genes. Most significant common pathways for each investigated cancer type are reported in Table 7.

Among the important pathways, *hsa01100: Metabolic pathways* involved in enzyme-mediated biochemical reactions that lead to the breakdown of natural product small molecules within a cell or tissue, *hsa01210: 2-Oxocarboxylic acid metabolism* composed of 2-oxocarboxylic acids, are the most elementary set of metabolites that includes pyruvate (2-oxopropanoate), 2-oxobutanoate, oxaloacetate and 2-oxoglutarate. Other commonly appearing pathways are, *hsa00071: Fatty acid degradation*, *hsa01040: Biosynthesis of unsaturated fatty acids*, *hsa04068: FoxO signaling pathway*, and *hsa00020: Citrate cycle (TCA cycle)* regulates cellular proliferation. In particular *hsa04068: FoxO signaling pathway* actively involved in the regulation of the expression of genes in cellular physiological events

including apoptosis, cell-cycle control, glucose metabolism, oxidative stress resistance, and longevity.

Conclusions

In this article, we have presented a pipeline to identify robust biomarker signatures for several of the most common cancer types. For this analysis, next-generation mRNA sequencing data from TCGA have been used to highlight significantly differentially expressed genes between healthy and tumor samples. Experimental results suggest that informative genes selected by ABC, ACO, DE, and PSO algorithms are mostly independent in nature. Variation of modeling performance among the optimization algorithms leads to the selection of the minutely overlapped set of genes irrespective of the datasets. However, majority voting of the overlapping sets of genes in different runs of the optimization algorithm is considered for choosing optimal biomarkers. The subset of overlapping genes responsible for highest classification accuracy is considered to be the optimal signature by different algorithms. Moreover, classification accuracy is considered as the underline objective for optimization and results indicate that all the used optimization algorithms achieve a satisfactory classification accuracy. In particular, ABC gains a slightly higher accuracy in all the tested datasets.

In conclusion, we can say that the current study identified robust genes as biomarker signatures and also analyzed their biological significance. For this analysis, pathway enrichment analysis has been used to study the overlapping $\mathscr{DE}_g$ genes with known genes of the KEGG pathways. Additionally, GO enrichment analysis also added some valuable insight. In particular, most of the identified genes are found to be involved in key oncogenesis pathways. Therefore, the identified biomarker signatures ($\mathscr{DE}_g$) in our experiments might be helpful to accurately identify tumors of unknown origin, as well as the proposed model itself, may be applied to other clinical queries.

Compliance with ethical standards

Conflicts of interest Shib Sankar Bhowmick, Debotosh Bhattacharjee and Luis Rato declare that they have no conflict of interest

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards

Informed consent Informed consent was obtained from all individual participants included in the study

References

Abu-Mouti FS, El-Hawary M (2011) Optimal distributed generation allocation and sizing in distribution systems via artificial bee colony algorithm. IEEE Trans Power Deliv 26(4):2090–2101

Argani P, Rosty C, Reiter RE, Wilentz RE, Murugesan SR, Leach SD, Ryu B, Skinner HG, Goggins M, Jaffee EM (2001) Discovery of new markers of cancer through serial analysis of gene expression: prostate stem cell antigen is overexpressed in pancreatic adenocarcinoma. Cancer Res 61(11):4320–4324

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on computational learning theory, Pittsburgh, pp 144–152

Cai H, Chung C, Wong K (2008) Application of differential evolution algorithm for transient stability constrained optimal power flow. IEEE Trans Power Syst 23(2):719–728

Chandra B, Gupta M (2011) An efficient statistical feature selection approach for classification of gene expression data. J Biomed Inform 44(4):529–535

Chopra P, Lee J, Kang J, Lee S (2010) Improving cancer classification accuracy using gene pairs. PLoS ONE 5(12):e14305

Dorigo M, Stützle T (2003) The ant colony optimization metaheuristic: algorithms, applications, and advances. In: Glover F, Kochenberger GA (eds) Handbook of metaheuristics. Springer, Boston, pp 250–285

Dorigo M, Birattari M, Stützle T (2006) Ant colony optimization. IEEE Comput Intell Mag 1(4):28–39

Dorigo M, Birattari M, Blum C, Clerc M, Stützle T, Winfield A (eds) (2008) Ant colony optimization and swarm intelligence: 6th International conference, ANTS 2008, Brussels, Belgium, September 22–24, 2008, Proceedings. Theoretical computer science and general issues, vol 5217. Springer, Berlin, Heidelberg

Eberhart Shi Y (2001) Particle swarm optimization: developments, applications and resources. Proc Evol Comput 1:81–86

Fleming RI, Harbison S (2010) The development of a mRNA multiplex RT-PCR assay for the definitive identification of body fluids. Forensic Sci Int: Genet 4(4):244–256

Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32:675–701

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Haas C, Klesser B, Maake C, Bär W, Kratzer A (2009) mRNA profiling for body fluid identification by reverse transcription endpoint PCR and realtime PCR. Forensic Sci Int: Genet 3(2):80–88

Han M, Liu X (2012) Forward feature selection based on approximate Markov blanket. In: International symposium on neural networks, Berlin, pp 64–72

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J, Boguski MS (1999) The transcriptional program in the response of human fibroblasts to serum. Science 283(5398):83–87

Juusola J, Ballantyne J (2007) mRNA profiling for body fluid identification by multiplex quantitative RT-PCR. J Forensic Sci 52(6):1252–1262

Kandaswamy KK, Chou KC, Martinetz T, Möller S, Suganthan P, Sridharan S, Pugalenthi G (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 270(1):56–62

Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J Global Optim 39(3):459–471

Karaboga D, Gorkemli B, Ozturk C, Karaboga N (2014) A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artif Intell Rev 42(1):21–57

Kennedy J (2011) Particle swarm optimization. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning. Springer, New York, pp 760–766

Kennedy J, Eberhart R (1995) Particle swarm optimization. Proc IEEE Int Conf Neural Netw 4:1942–1948

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44(W1):W90–W97

Lapointe J, Li C, Higgins JP, Van De Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci 101(3):811–816

Liu H, Liu L, Zhang H (2010) Ensemble gene selection by grouping for microarray data classification. J Biomed Inform 43(1):81–87

Liu J, Ranka S, Kahveci T (2008) Classification and feature selection algorithms for multi-class CGH data. Bioinformatics 24(13):i86–i95

Liu Q, Sung AH, Chen Z, Liu J, Chen L, Qiao M, Wang Z, Huang X, Deng Y (2011) Gene selection and classification for cancer microarray data based on machine learning and similarity measures. BMC Genom 12(5):S1

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550

Mramor M, Leban G, Demšar J, Zupan B (2007) Visualization-based cancer microarray data classification analysis. Bioinformatics 23(16):2147–2154

Olopade OI, Grushko T (2001) Gene-expression profiles in hereditary breast cancer. N Engl J Med 344(26):2028–2029

Ooi C, Tan P (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 19(1):37–44

Peng Y, Wu Z, Jiang J (2010) A novel feature selection approach for biomedical data classification. J Biomed Inform 43(1):15–23

Richard MLL, Harper KA, Craig RL, Onorato AJ, Robertson JM, Donfack J (2012) Evaluation of mRNA marker specificity for the identification of five human body fluids by capillary electrophoresis. Forensic Sci Int: Genet 6(4):452–460

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Shi Y, Eberhart R (1998) A modified particle swarm optimizer. In: Proceedings of IEEE international conference on evolutionary computation, Anchorage, pp 69–73

Storn R, Price K (1997) Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359

Wang Y, Jatkoe T, Zhang Y, Mutch MG, Talantov D, Jiang J, McLeod HL, Atkins D (2004) Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 22(9):1564–1571

Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW (2005) Gene selection from microarray data for cancer classification-a machine learning approach. Comput Biol Chem 29(1):37–46

Wobst J, Banemann R, Bastisch I (2011) RNA can do better-an improved strategy for RNA-based characterization of different body fluids and skin. Forensic Sci Int Genet Suppl Ser 3(1):e421–e422

Zhang H, Wang H, Dai Z, Ms Chen, Yuan Z (2012) Improving accuracy for cancer classification with a new algorithm for genes selection. BMC Bioinform 13(1):298