Identification of tissue-specific tumor biomarker using different optimization algorithms

Shib Sankar Bhowmick¹ · Debotosh Bhattacharjee² · Luis Rato³

Received: 16 April 2018 / Accepted: 3 December 2018
© The Genetics Society of Korea 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).
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 \( \mathcal{D} \) 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 \( P_i = \frac{fit_i}{\sum_{n=1}^{N_{cl}} fit_n} \). 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:

\[
\mathcal{Z}_j = x_j + \Phi_j (x_j - x_k)
\]

where \( x_j \) (or \( \mathcal{Z}_j \)) denotes the \( j \)-th element of \( x_i \) (or \( \mathcal{Z}_i \)), and \( j \) is a random index, \( j \in \{ 1, 2, \ldots, \mathcal{D} \} \). \( x_k \) denotes another solution selected randomly from the population where \( k \in \{ 1, 2, \ldots, \mathcal{D}_k \} \). \( \Phi_j \) represents a uniform random number in \([-1, 1]\) and \( \mathcal{D}_k \) corresponds to the number of employed bees. The update process is completed here by a greedy selection between \( x_i \) and \( \mathcal{Z}_i \). 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 \( r_0 \). Subsequently, during the construct ant solutions step, each ant begins with an empty solution \( s_p = \emptyset \). 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, \( c_j \in N(s_p) \subseteq \mathcal{C} \). In this way, it upgrades its current partial solution. Here, \( N(s_p) \) represents a set of solution components, defined mainly by an implemented solution construction process. 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:

\[
p(c_j|s_p) = \frac{r_0^a \eta^\beta(c_j)}{\sum_{c_j \in N(s_p)} r_0^a \eta^\beta(c_j)}, \forall c_j \in N(s_p)
\]

(2)

\( \mathcal{Z}_p \)

Here, \( \alpha = 0 \) corresponds to the selection probabilities that is proportional to \( [\eta_j]^{\beta} \). 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:
\[ \tau_{ij} = (1 - \rho)\tau_{ij} + \sum_{s \in S_{upd}, i' \in s} g(s) \]  

(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 \( \mathbb{R}^D \). DE is an evolutionary optimization tool that has a wide range of applications. Here, DE is used to code features. As in any other evolutionary algorithm, population initialization in DE begins with a randomly initiated population of \( N_p \) 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.,

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

(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:

\[ \mathcal{V}_{i,L}(t+1) = \mathcal{G}_{i,L}(t) + F(\mathcal{G}_{n,L}(t) - \mathcal{G}_{m,L}(t)) \]  

(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.

\[ \mathcal{Q}_{i,j}(t+1) = [\mathcal{Q}_{i,1}(t+1), \mathcal{Q}_{i,2}(t+1), \ldots, \mathcal{Q}_{i,L}(t+1) \]  

(6)

where

\[ \mathcal{Q}_{i,j}(t+1) = \begin{cases} \mathcal{V}_{i,j}(t+1), & \text{if } rand(0, 1) \leq CR, \text{ or } j = rand(i), \\ \mathcal{G}_{i,j}(t), & \text{if } rand(0, 1) > CR, \text{ and } j \geq rand(i) \end{cases} \]  

(7)

In Eq. 7, \( rand(0, 1) \) 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 \( \mathcal{Q}_{i,j}(t+1) \) gets at least one parameter from \( \mathcal{V}_{i,j}(t+1) \).

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:

\[ \mathcal{G}_i(t+1) = \begin{cases} \mathcal{Q}_{i,j}(t+1), & \text{if } f(\mathcal{Q}_{i,j}(t+1)) > f(\mathcal{G}_i(t)) \\ \mathcal{G}_i(t), & \text{if } f(\mathcal{Q}_{i,j}(t+1)) \leq f(\mathcal{G}_i(t)) \end{cases} \]  

(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):

\[ V_i(t+1) = w \times V_i(t) + \varphi_1 \times (P_{loc}(t) - P_i(t)) + \varphi_2 \times (P_{glob}(t) - P_i(t)) \]  

(9)

\[ P_i(t+1) = P_i(t) + V_i(t+1) \]  

(10)

where \( t \) represents a time stamp of different iteration, \( w \) is the inertia weight \( \in [0.5, 1] \), \( \varphi_1 \) and \( \varphi_2 \) are the cognitive and social constants. Similarly, \( P_{loc} \) particle represents local best of the current iteration, while global best particle (till the current iteration) is \( P_{glob} \). 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 \(\mathcal{P}\) number of patients. Therefore each particle, colony or population, is made up of a distinct dataset of size \(\mathcal{P} \times L\). 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 \(\mathcal{P} \times L\) 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
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’.

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, \(\mathcal{D E}S\). 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 (\(\mathcal{D E}S_g\)) genes are identified from

---

\(^1\) [http://gdac.broadinstitute.org/](http://gdac.broadinstitute.org/).

---

**Table 1** Details of TCGA data used in the experiments

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

Each tumor class is coupled with 465 samples of healthy class.

---

**Table 2**

<table>
<thead>
<tr>
<th>Disease</th>
<th>Code</th>
<th>Tumor samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder urothelial carcinoma</td>
<td>BLCA</td>
<td>408</td>
</tr>
<tr>
<td>Colon adenocarcinoma</td>
<td>COAD</td>
<td>287</td>
</tr>
<tr>
<td>Kidney renal clear cell carcinoma</td>
<td>KIRC</td>
<td>534</td>
</tr>
<tr>
<td>Brain lower grade glioma</td>
<td>LGG</td>
<td>530</td>
</tr>
<tr>
<td>Liver hepatocellular carcinoma</td>
<td>LIHC</td>
<td>373</td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>LUAD</td>
<td>517</td>
</tr>
<tr>
<td>Prostate adenocarcinoma</td>
<td>PRAD</td>
<td>498</td>
</tr>
<tr>
<td>Skin cutaneous melanoma</td>
<td>SKCM</td>
<td>471</td>
</tr>
<tr>
<td>Thyroid carcinoma</td>
<td>THCA</td>
<td>509</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>4127</td>
</tr>
</tbody>
</table>
these initial filtered \( \mathcal{E} \) 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 \( m_{\text{opt}} \) 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 \( \text{KAT2B, SGCE, SLC35B3, ACO1, ACOT1, ANKHD1, APBA1, ARHGAP31, ARM CX1, BBS7, C3orf38, and C9orf82} \) 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 (\( m_{\text{appe}} \) ) 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

<table>
<thead>
<tr>
<th>Rank</th>
<th>BLCA</th>
<th>ACO</th>
<th>DE</th>
<th>PSO</th>
<th>COAD</th>
<th>ACO</th>
<th>DE</th>
<th>PSO</th>
<th>KIRC</th>
<th>ACO</th>
<th>DE</th>
<th>PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>KAT2B</td>
<td>IFRD1</td>
<td>CPVL</td>
<td>PECI</td>
<td>NUP62</td>
<td>ADORA2A</td>
<td>ATP2A2</td>
<td>RHBDF2</td>
<td>APEH</td>
<td>MTMR10</td>
<td>C7orf42</td>
<td>ENG</td>
</tr>
<tr>
<td>2</td>
<td>SGCE</td>
<td>IQGAP2</td>
<td>PPIPK1</td>
<td>AFMID</td>
<td>RRP9</td>
<td>DTNB1P</td>
<td>CD99</td>
<td>ZNF83</td>
<td>ASPSCR1</td>
<td>NUP155</td>
<td>RABEP1</td>
<td>EPHB2</td>
</tr>
<tr>
<td>3</td>
<td>SLC35B3</td>
<td>BOC</td>
<td>SLC8A1</td>
<td>ANKS1A</td>
<td>LOC652276</td>
<td>PNPT1</td>
<td>FAM98B</td>
<td>ANKR27</td>
<td>CSTF2T</td>
<td>PRDM2</td>
<td>AIMP2</td>
<td>LOC440957</td>
</tr>
<tr>
<td>4</td>
<td>AC01</td>
<td>C10orf58</td>
<td>TANC1</td>
<td>ANXA2</td>
<td>MRM1</td>
<td>POLR2H</td>
<td>IQGAP3</td>
<td>ARHGAP26</td>
<td>FOXRED1</td>
<td>PRKCD</td>
<td>AK1</td>
<td>AAAS</td>
</tr>
<tr>
<td>5</td>
<td>ACOT1</td>
<td>C17orf53</td>
<td>AGK</td>
<td>ARHGAP10</td>
<td>PRMT2</td>
<td>SLC35A2</td>
<td>ITCH</td>
<td>ARSB</td>
<td>SEPHS2</td>
<td>SIRT6</td>
<td>ARHGAP10</td>
<td>ACN9</td>
</tr>
<tr>
<td>6</td>
<td>ANKHD1</td>
<td>C1RL</td>
<td>AMACR</td>
<td>ASF1B</td>
<td>TMEM185A</td>
<td>ALDH1A1</td>
<td>TH1L</td>
<td>C12orf51</td>
<td>AGPAT2</td>
<td>ALAD</td>
<td>ARHGAP19</td>
<td>ADO</td>
</tr>
<tr>
<td>7</td>
<td>APBA1</td>
<td>CCDC50</td>
<td>AMFR</td>
<td>BCAP29</td>
<td>ARHGAP9</td>
<td>ALDOA</td>
<td>UCHL3</td>
<td>C20orf20</td>
<td>C19orf63</td>
<td>AP3B1</td>
<td>ASS1</td>
<td>AFAP1L1</td>
</tr>
<tr>
<td>8</td>
<td>ARHGAP31</td>
<td>APPL2</td>
<td>C1orf57</td>
<td></td>
<td>ATP1B3</td>
<td>ASH1L</td>
<td>ACSF3</td>
<td>CCDC109A</td>
<td>C1orf21</td>
<td>ATPL3</td>
<td>AGPS</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>ARMCX1</td>
<td>ARFGAP2</td>
<td>C1QTNF6</td>
<td></td>
<td>C2CD2</td>
<td>ACTRIB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>LGG</td>
<td>ADC</td>
<td>CORO2A</td>
<td>PANK1</td>
<td>CTNNAL1</td>
<td>ADC</td>
<td>CORO2A</td>
<td>PANK1</td>
<td>CTNNAL1</td>
<td>ADC</td>
<td>CORO2A</td>
<td>PANK1</td>
</tr>
<tr>
<td>11</td>
<td>C3orf38</td>
<td>ARRD3</td>
<td>C3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>C9orf82</td>
<td>ASF1B</td>
<td></td>
<td></td>
<td>C9orf167</td>
<td>BAIAP2</td>
<td>L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>AP9A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>CEP120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>CFL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 3 (continued)

<table>
<thead>
<tr>
<th>Rank</th>
<th>PRAD</th>
<th>SKCM</th>
<th>THCA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABC</td>
<td>ACO</td>
<td>DE</td>
</tr>
<tr>
<td>1</td>
<td>PVRL2</td>
<td>MICALL1</td>
<td>CD302</td>
</tr>
<tr>
<td>2</td>
<td>SMPD1</td>
<td>RAD54L2</td>
<td>HMGC1R</td>
</tr>
<tr>
<td>3</td>
<td>AMT</td>
<td>AIMP2</td>
<td>KLHL18</td>
</tr>
<tr>
<td>4</td>
<td>ANXA2</td>
<td>ALKBH3</td>
<td>SAPS2</td>
</tr>
<tr>
<td>5</td>
<td>C19orf42</td>
<td>AMZ2</td>
<td>ACOT8</td>
</tr>
<tr>
<td>6</td>
<td>CDC5</td>
<td>AP4M1</td>
<td>ANXA5</td>
</tr>
<tr>
<td>7</td>
<td>CEMP1</td>
<td>BPNT1</td>
<td>BCAS3</td>
</tr>
<tr>
<td>8</td>
<td>CNOT8</td>
<td>C1orf93</td>
<td>BRD8</td>
</tr>
<tr>
<td>9</td>
<td>CSNK1G2</td>
<td>C5orf42</td>
<td>C12orf49</td>
</tr>
<tr>
<td>10</td>
<td>CXorf60B</td>
<td>CLIC1</td>
<td>C19orf63</td>
</tr>
<tr>
<td>11</td>
<td>DOM3Z</td>
<td>CRYZ</td>
<td>CIQ1</td>
</tr>
<tr>
<td>12</td>
<td>ETV5</td>
<td>CTSA</td>
<td>DCAF15</td>
</tr>
<tr>
<td>13</td>
<td>FAAH</td>
<td>DHX40</td>
<td>ENTPD6</td>
</tr>
<tr>
<td>14</td>
<td>FAHD2A</td>
<td>DNAJC19</td>
<td>FLJ10038</td>
</tr>
</tbody>
</table>
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 €. We set γ = 0.5 and € = 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:0044444 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 euukaryotes, 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 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)
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 $\mathcal{G}$ 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 ($\mathcal{G}$) 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


