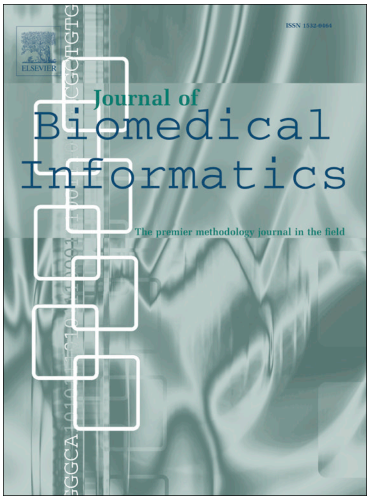


Neuroevolution as a Tool for Microarray Gene Expression Pattern Identification in Cancer Research

Bruno Iochins Grisci, Bruno César Feltes, Marcio Dorn

PII:	S1532-0464(18)30226-0
DOI:	https://doi.org/10.1016/j.jbi.2018.11.013
Reference:	YJBIN 3085
To appear in:	*Journal of Biomedical Informatics*
Received Date:	29 August 2018
Revised Date:	12 November 2018
Accepted Date:	27 November 2018

Please cite this article as: Grisci, B.I., Feltes, B.C., Dorn, M., Neuroevolution as a Tool for Microarray Gene Expression Pattern Identification in Cancer Research, *Journal of Biomedical Informatics* (2018), doi: https://doi.org/10.1016/j.jbi.2018.11.013

Neuroevolution as a Tool for Microarray Gene Expression Pattern Identification in Cancer Research

Bruno Iochins Grisci[a,b], Bruno César Feltes[a,b], Marcio Dorn[a,c]

[a] Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil
[b] These authors contributed equally to this work.
[c] Corresponding Author: mdorn@inf.ufrgs.br

Abstract

Microarrays are still one of the major techniques employed to study cancer biology. However, the identification of expression patterns from microarray datasets is still a significant challenge to overcome. In this work, a new approach using Neuroevolution, a machine learning field that combines neural networks and evolutionary computation, provides aid in this challenge by simultaneously classifying microarray data and selecting the subset of more relevant genes. The main algorithm, FS-NEAT, was adapted by the addition of new structural operators designed for this high dimensional data. In addition, a rigorous filtering and preprocessing protocol was employed to select quality microarray datasets for the proposed method, selecting 13 datasets from three different cancer types. The results show that Neuroevolution was able to successfully classify microarray samples when compared with other methods in the literature, while also finding subsets of genes that can be generalized for other algorithms and carry relevant biological information. This approach detected 177 genes, and 82 were validated as already being associated to their respective cancer types and 44 were associated to other types of cancer, becoming potential targets to be explored as cancer biomarkers. Five long non-coding RNAs were also detected, from which four don't have described functions yet. The expression patterns found are intrinsically related to extracellular matrix, exosomes and cell proliferation. The results obtained in this work could aid in unraveling the molecular mechanisms underlying the tumoral process and describe new potential targets to be explored in future works.

Keywords: Neuroevolution, Cancer, Microarray, FS-NEAT, Machine Learning, Feature Selection

Introduction

Over the past years, the fast advance of microarray technologies allowed the study of a variety of biological questions - from the basic functionality of an organism, to the understanding of complex diseases, such as cancer (Tao et al., 2017). Nowadays, microarrays are available from different platforms and provide biological information on mRNA, miRNA, lncRNA and exon arrays (Tao et al., 2017; Wang and Xi, 2013; Gorreta et al., 2012; Deniz and Erman, 2017). However, microarrays are still majorly employed to analyze mRNA, and despite the vast amount of tools available for microarray gene expression analysis, the identification of expression patterns is still a significant challenge to overcome (Walsh et al., 2015).

Classical gene expression approaches aim to obtain a list of differentially expressed genes (DEG) from expression datasets; however, such lists can be composed of hundreds or thousands of DEG and, although biologically relevant, no clear pattern can be drawn from within. In this sense, machine learning (ML) techniques could provide fast and accurate identification of expression patterns.

ML algorithms have been applied to microarray data mostly with two distinct, but complementary objectives: sample classification and gene selection. The first is a supervised learning task: given a gene expression pattern, it aims to correctly identifying its label, for instance, if it is a normal or tumoral tissue. This approach has many applications in clinical diagnostics and has successfully been tested with different algorithms in the past years (Leung and Cavalieri, 2003b). Among the available techniques, Support Vector Machines (SVM) have been considered the best option for this type of data, obtaining the best results in several comparisons (Statnikov et al., 2008; Pirooznia et al., 2008; Lee et al., 2005; Ang et al., 2016).

The other task for ML, gene selection, is a subdivision of the more general problem of feature selection (FS) (Miao and Niu, 2016), a form of dimensionality reduction. While feature extraction methods like Principal Component Analysis (PCA) perform dimensionality reduction by combining the different data dimensions, FS selects the dimensions themselves. This is fundamental for gene selection, in which the features are the expression value of genes, since it preserves their physical meaning, allowing better interpretation (Ang et al., 2016). Gene selection is essential for microarray classification because of the ”curse of dimensionality” and the ”large p, small n problem”, associated with data with a large number of dimensions but a small sample size, what can cause the model to overfit (Verleysen and François, 2005), increase memory consumption, processing time, and diminish interpretability. FS is also useful in a biological context by aiding in biomarkers identification, since it finds the subset of genes that has more discriminatory power.

One of the branches of algorithms available for ML is the field known as Neuroevolution (Sher, 2013), a combination of Artificial Neural Networks (ANN) (Haykin, 2009) and Evolutionary Computation (EC) (Eiben and Smith, 2015). EC borrows topics from evolutionary biology, such as inheritance, random variation, and selection, and adapts them to the context of computation. It does not require a significant amount of data, is easily parallelized, and can give solutions based on a fitness function (Sipper et al., 2017). Neuroevolution is a family of training methods for ANNs to obtain theirs weights, biases, and topology with EC (Ding et al., 2013). One example is the NeuroEvolution of Augmenting Topologies (NEAT) (Stanley and Miikkulainen, 2002) that uses Genetic Algorithms (GA) (Mitchell, 1998) into training and autonomously design the topology of the network.

NEAT has been successfully extended to deal with classification and selection problems (Tan et al., 2009; Sohangir et al., 2013, 2014). More specifically, Feature Selective NEAT (FS-NEAT) (Whiteson et al., 2005) showed good performance on simpler tasks (Papavasileiou and Jansen, 2016, 2017b) and preliminar results indicate it can simultaneously perform microarray classification and gene selection (Grisci et al., 2018). One significant advantage of these techniques is that they act autonomously, without the need of a predefined number of selected genes, network topology or thresholds.

Nevertheless, a robust ML pipeline is not the only requirement to extract useful and precise information from microarray data. Input quality is rarely discussed and significantly impact on the subsequent analysis (Allison et al., 2006; Leung and Cavalieri, 2003a). In a microarray study, when a nucleic acid sequence hybridizes with the one in the probe, a signal is emitted, and a value indicating the presence and abundance of the target is provided (Epstein and Butow, 2000;

Blohm and Guiseppi-Elie, 2001). Nevertheless, this raw data contains noise that came from prior manipulation during the experiment or from the platform itself (Kauffmann and Huber, 2010). Thus, using raw microarray data to test or validate computational approaches using such datasets, without concern for its precedence or quality can influence the obtained results (Kauffmann and Huber, 2010; Owzar et al., 2011).

Taking these challenges into consideration, in this work, we describe the design and application of a variant of FS-NEAT as a tool to perform classification and identify gene expression patterns in microarray data, focusing on cancer datasets. In addition, aiming to select only the most homogeneous and reliable datasets for our own application, we conducted a rigorous search protocol for datasets selection and a classical biological approach for background correction, normalization, and sample quality assessment. From the analysis of 13 datasets from three different cancer types, with a total of 1024 samples, our results autonomously and successfully classified the microarrays and selected genes at the same time. We extracted distinct expression patterns, all manually validated in the scientific literature and provided new insights on expression patterns of three types of cancer.

Materials and methods

Microarray data

To obtain multiple microarray datasets (GSEs), the raw data of leukemia, breast, and colorectal cancers were downloaded from the Gene Expression Omnibus (GEO) database using the *GEOquery* package (Davis and Meltzer, 2007) for the R platform[1] (Fig. 1 - *Data Obtainment*). The following criteria were applied to select the most homogeneous and reliable datasets (Fig. 1 - *Preprocessing*): (i) exclusion of studies that used chemotherapics, gene therapies of any kind, or that employed interfering molecules, such as miRNA, siRNA, etc; (ii) selection of studies performed only on *Homo sapiens*; (iii) microarrays that didn't use any form of Knockdown cultures, or specific selected mutations; (iv) only datasets that contained at least six normal (control) samples and six experimental (tumoral) samples; (v) studies with a clear description of the protocols used in the experiments; and (vi) studies that didn't use any kind of xenograft technique. We chose to select only data from a single company, in this case, Affymetrix, to keep the data as consistent as possible. After data obtainment, background correction and ”rma” normalization of all selected GSEs were performed by the R package *affy* (Gautier et al., 2004).

After normalization, datasets were analyzed by the R package *arrayQualityMetrics* (Kauffmann et al., 2009), to access the sample quality of the selected microarrays. Samples that displayed low quality in at least half of any parameters measured by *arrayQualityMetrics* were discarded from the final pool. Table 1 summarizes the chosen GSEs, their specifications, and the number of excluded samples.

The final expression matrices were then used as inputs for the proposed ML pipeline (Fig. 1 - *Genes Expression*). In addition to the selected GSEs, we included the original microarray dataset from Golub et al. (1999)[2] with AML and ALL leukemia subtypes, in order to provide a comparison with recent works focused on the task of microarray classification employing Neuroevolution as seen in Garro et al. (2017).

[1] `www.r-project.org`

[2] `https://github.com/ramhiser/datamicroarray`

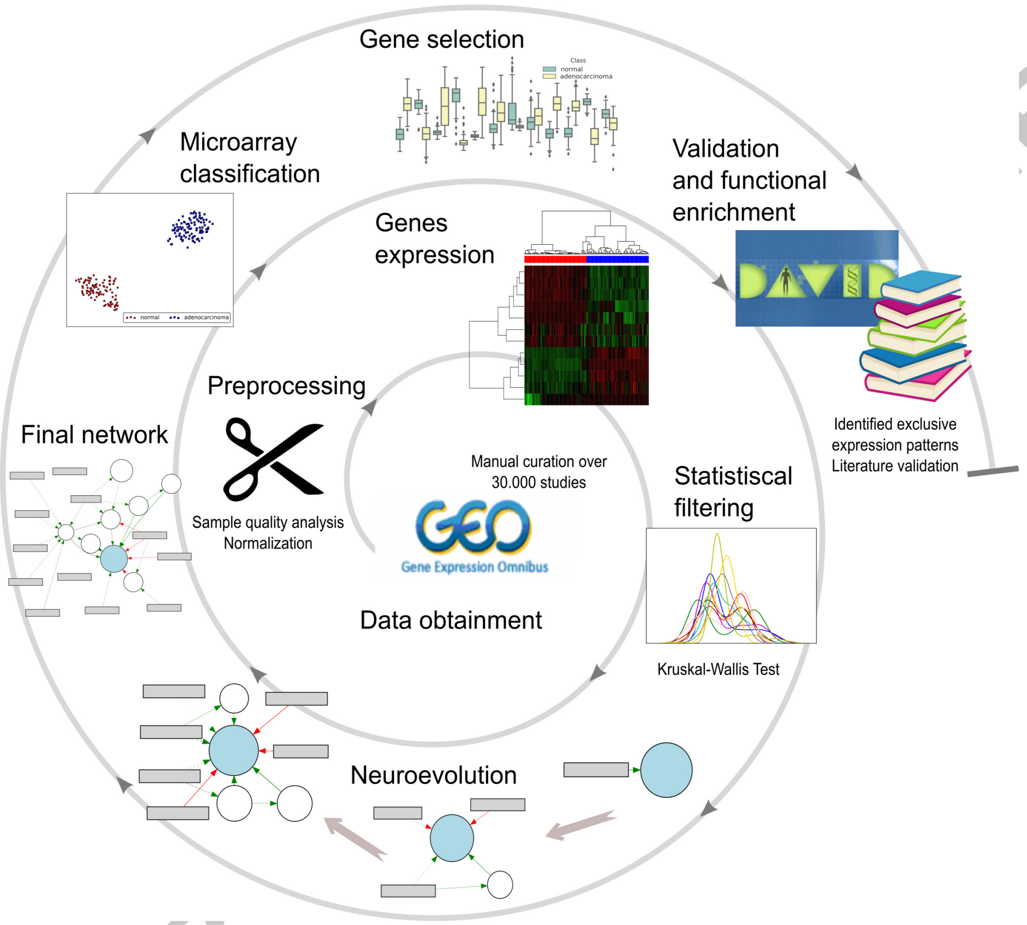


Figure 1: **Summary of the methodological steps taken in this work.** After data obtainment, the microarray datasets were normalized and the low quality samples were excluded. The genes were filtered using the Kruskal-Wallis H Test and the remaining data was employed in the Neuroevolution process, from which the best neural network was chosen. Finally, the neural network was used to perform the microarray classification, and its inputs used for gene selection. The final selected genes, which represent distinct expression patterns, were submitted to a functional enrichment analysis. Additionally, we conducted an extensive search in the scientific literature to see the types of cancer that the selected genes were associated to.

NeuroEvolution of Augmenting Topologies

Neuroevolution is applied to the resulting data after the preprocessing (Fig. 1 - *Neuroevolution*) and further filtered with Kruskal-Wallis Test (Fig. 1 - *Statistical Filtering*), employing a variation of the NEAT algorithm as described in the next Section. Generically, NEAT starts with a population of ANNs with minimal topology, i.e., all inputs fully-connected to all outputs and no hidden nodes, with random weights and biases. This minimalist start is essential to ensure that only the structural additions that in fact bring some advantage to the ANN are kept, avoiding needless complexity. From this first population, new individuals (ANNs) are generated iteratively through GA operators. The crossover operator selects two individuals from the population

Table 1: **List of GSEs and datasets employed in this work.**

Datasets	Cancer Type	Samples	Excluded Samples*	Genes	Classes
GSE42568	Breast	121	5	54675	2
GSE45827	Breast	155	4	54675	6
GSE10797	Breast	66	None	22277	3
GSE44076	Colorectal	246	52	49386	2
GSE44861	Colorectal	111	6	22277	2
GSE8671	Colorectal	64	1	54675	2
GSE21510	Colorectal	148	105	54675	2
GSE32323	Colorectal	44	11	54675	2
GSE41328	Colorectal	20	2	54675	2
GSE9476	Leukemia	64	None	22283	5
GSE14317	Leukemia	26	1	22277	2
GSE63270	Leukemia	104	3	54675	2
GSE71935	Leukemia	51	6	54675	2
Golub et al. (1999)	Leukemia	72	NA	7129	2

*: (i) samples excluded prior to the analysis, due to the presence of one or more samples that didn't met the criteria described on the Materials and Methods; (ii) samples that could generate a bias in the analysis due to treatment, tissue origin or platform mix; (iii) file corruption and errors; and (iv) samples excluded due to low quality. NA = Not Applicable.

and combines them (S1-Figure in Supplementary Data). The mutation can add a new node by splitting an existing connection (S2-Figure - *Add node* in Supplementary Data) or add a new connection between existing nodes (S2-Figure - *Add connection*), besides changing the weights and biases values. This is how the ANNs grow in complexity and diversity over the generations of the GA (Stanley and Miikkulainen, 2002).

One challenge that arises from this strategy is the correct combination of two individuals when performing crossover, since it can generate defective ANNs if the topologies do not allow a direct exchange of nodes and connections. NEAT avoids this problem by implementing historical marks, a numerical tag associated with every new structural innovation that arises during the evolution, such as a new connection between nodes. The value of the mark is assigned linearly considering when the new structure first appeared, and it is passed without change to new individuals during the crossover. By aligning all the historical marks of two networks, NEAT is capable of perfectly matching each piece of structure of the parents ANNs, creating a functional ANN with the same blocks. When aligning the individuals, these structural pieces can be disjoint (if missing from the parent) or excess (if missing from the other parent) in regard of one another. The new network receives the structures from the parent with better fitness if they are disjoint or excess, or randomly from any of them otherwise (S1-Figure).

Another major challenge is that adding new structures to a network without optimizing the weights and biases often brings disadvantageous results, creating a negative pressure towards innovation. Once again the historical marks are used to implement speciation (or niche). The compatibility between individuals is computed with Eq. 1 (Stanley and Miikkulainen, 2002) in which c_1, c_2, and c_3 are coefficients set by the user, N is the number of structures in the largest network, E is the number of excess structures, D is the number of disjoint structures, and $\bar{W}$ is the average weight differences of matching structures. If two individuals have a difference greater than a given threshold, they are placed in separated species and do not compete directly with one another, allowing for the population to diversify.

$$d = c_1 \frac{E}{N} + c_2 \frac{D}{N} + c_3 \bar{W} \tag{1}$$

FS-NEAT (Whiteson et al., 2005) is an extension of NEAT in which the minimalist start is altered, and instead of all individuals in the first population beginning with a full-connected architecture, only one random input is connected to one random output for each network. The only other needed addition is a new mutate operator, that adds inputs to a network by connecting it to any output (S3-Figure in Supplementary Data). The inputs that are not directly or indirectly connected to any output at the end of the evolutive process are discarded. Thus, FS-NEAT automatically performs FS without meta-learning or labeled data, while creating simpler and less costly networks since they only need a subset of all inputs.

Functional Enrichment

To access the most relevant bioprocesses and trace the nature of the final selected genes from all GSEs we employed the Database for Annotation, Visualization and Integrated Discovery (DAVID)v 6.8 (Huang et al., 2009b,a) (Fig. 1 - *Validation and Functional Enrichment*). The entire list of genes was used as input in DAVID, using the Benjamini FDR correction with a significance score of 0.05.

Proposed Method

We propose a method based in FS-NEAT capable of performing both tasks of microarray classification (Fig. 1 - *Microarray Classification*) and gene selection (Fig. 1 - *Gene Selection*) autonomously, without the need for specifying how many genes should be selected at the end. The first consideration about our method is that it uses One-vs-All classification for multiclass classification problems. This means that if a dataset has more than two classes, we classify each class separately against the other classes combined. While FS-NEAT can handle multiclass data, we chose the One-vs-All approach due to four assumptions: (i) most of the microarray datasets are binary; (ii) One-vs-All allows the use of only one output neuron in each ANN, simplifying their structures; (iii) it becomes easier to interpret the selection result, since for each subset of selected genes we are considering only one class; (iv) the major drawback of One-vs-All classification is the creation of size imbalance among classes, but for many microarray experiments the data is already imbalanced and, in fact, sometimes becomes better balanced with the splits created with One-vs-All.

Due to the presence of thousands of genes in each microarray dataset, before starting the evolutive process, we filtered the data using the Kruskal-Wallis H Test (KW) by comparing the expression of each gene among the two classes and removing all genes that presented no difference between them (p-value $\geqslant$ 0.01) (Fig. 1 - *Statistical Filtering*). This nonparametric approach does not assume a normal distribution and has already been used to study microarray data (Lan and Vucetic, 2011), and the use of statistical methods as a preprocessing filtering step is frequent in the literature (Luque-Baena et al., 2013). After the application of the KW, around 13% of the total amount of genes is kept for the next stages. The final preprocessing step is to normalize the gene expression using the mean normalization as described in Eq. 2, with x being a feature, and μ, x_{max}, and x_{min} being the mean, maximum value and minimum value of that feature over all the samples, respectively.

$$x_{new} = \frac{x - \mu}{x_{max} - x_{min}} \tag{2}$$

The next stage is the Neuroevolution itself (Fig. 1 - *Neuroevolution*). The output of the networks is a value between 0 and 1 that predicts to which class a sample belongs, and the inputs are the normalized values of the expression of the genes. The first population is created by connecting one random input to the output, and the initial weights and biases are randomly determined from a distribution with mean equal to zero and standard deviation equal to one. Since we are dealing with higher dimensions than usually used with FS-NEAT, we modified the algorithm to better explore the input space. In addition to the ”add node” and ”add connection” mutations from NEAT, the original crossover from NEAT and the ”add input” mutation from FS-NEAT were modified, and a new mutation was added:

Crossover operator: works similarly as the NEAT crossover operator, but if the parent with lower fitness has an input that the other parent does not, and this input is connected to a node present in the other parent, there is a 50% chance of the offspring inheriting that input (Fig. 2). This change allows the combination of the features selected by two ANNs, what is not permitted by the original crossover, since the offspring will always have the same FS as the parent with better fitness.

Swap input mutation: a new proposed mutation that randomly swaps one of the network inputs by another input not present in the ANN (Fig. 3 - *Swap input*). This allows the algorithm to explore the use of new possible features without increasing the ANNs size, and exploiting the already existing network structure.

Guided add input mutation: the p-values from the KW step are transformed by the formula $-\log_{10}(p)$ and scaled by the softmax function (Eq. 3, Z being a vector of probabilities z). The outputs are probabilities that are larger for smaller p-values. They are used as the probability of an input being selected by the ”add input mutation”, meaning that the genes that showed the largest difference between classes are more likely to be selected by the mutation (Fig. 3 - *Guided add input*).

$$\phi(Z) = e^{z} \div \sum_{z' \in Z} e^{z'}, \forall z \in Z \tag{3}$$

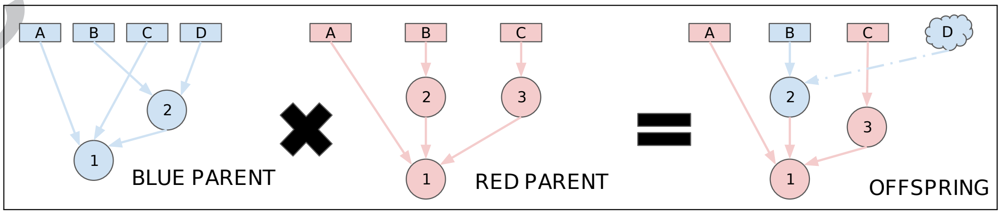


Figure 2: **The proposed crossover operator**. Given two parents, red (better fitness) and blue (worse fitness), the offspring ANN will be a combination of the two, inheriting the structures from both randomly when both have it, and from red otherwise. The major difference from FS-NEAT is that if there is an input in blue that is not connected to red, and this input in blue is connected to a node that is in red, the offspring has 50% of chance of inheriting it as well, here represented by input ”D”.

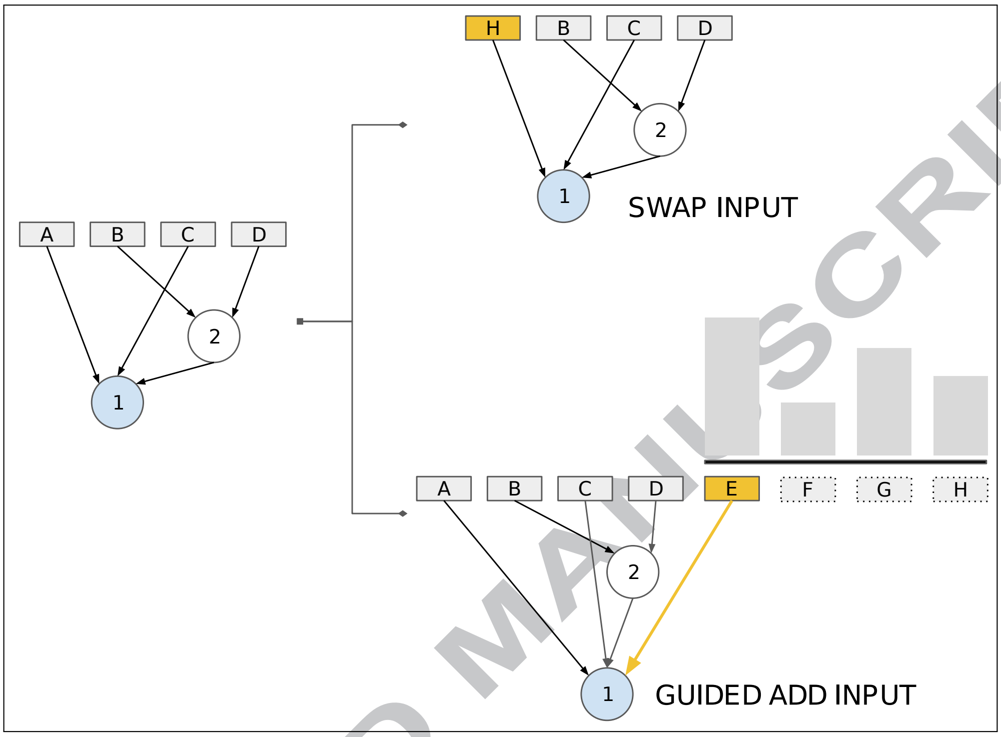


Figure 3: **The two new structural mutations in the proposed method, derived from the central neural network.** Rectangles represent inputs, blue circles indicate outputs, white circles represent hidden nodes, and arrows are the connections between nodes. The new structures are marked in yellow. The histogram above the network that had an added input represents the probabilities of each new input being selected by this operator.

The fitness function that guides the evolutive process is the cross-entropy, also known as the log loss, in its binary form. It is a popular cost function for supervised learning but does not account for data imbalance, which is common in microarray data. Thus, we altered the cross-entropy so that it will be computed individually for each class q and then averaged, as shown in Eq. 4a, in which: n^q is the number of samples of the class q, y_i is the true label of the i^{th} sample, and a_i is the ANN output for the i^{th} sample. This way, all classes have the same contribution to the fitness independently of their sizes (Grisci et al., 2018).

The second term of the fitness function, given by Eq. 4b, stands for the L2 regularization or weight decay, a technique commonly used to avoid the problem of over-fitting (Ng, 2004), when the model performs well on the training data, but fails to generalize and has poor performance in new data. The L2 regularization penalizes networks with large weight and bias values, under the assumption that simpler models are better in generalizing. Since the number of inputs of a neuron can change during the evolution, we added the term $\frac{1}{c}$, so that the regularization would not have a negative impact in the addition of new connections and nodes. The c is the number of connections and biases, n is the number of samples, w_k is the weight or bias of the connection of node k, and λ is the regularization parameter. Due to the minimalist start of FS-NEAT, our method does not require a component to minimize the number of features selected, as used in Luque-Baena et al.

(2013) for instance, making for easier fitness function design.

$$fit = \frac{1}{|Q|} \sum_{q \in Q} \left\{ -\frac{1}{n^q} \sum_{i=1}^{n^q} [y_i \ln a_i + (1 - y_i) \ln(1 - a_i)] \right\} \tag{4a}$$

$$+ \frac{\lambda}{2n} \frac{1}{c} \sum_{k=1}^{c} w_k^2 \tag{4b}$$

The structure of the neurons in our method is given by Eq. 5, in which a_h is the output, m_h is the number of inputs, b_h is the bias, w_{hj} is the weight of the j^{th} input, and x_{hj} is the j^{th} input of the neuron h, respectively. Once again the aggregation is the average of the inputs instead of the sum because the number of inputs can vary during the evolution. The f stands for the activation function of the neuron, that for the output neuron is the Gaussian function in Eq. 6, and for all the hidden nodes is the modified hyperbolic tangent in Eq. 7. These two functions combined have shown the best performance in the context of FS-NEAT in comparative studies (Papavasileiou and Jansen, 2017a).

$$a_h = f(\frac{1}{m_h} \sum_{j=1}^{m_h} w_{hj} x_{hj} + b_h) \tag{5}$$

$$f(x) = exp(-\frac{5(x - \mu)^2}{2\sigma^2}), \mu = 0, \sigma = 1 \tag{6}$$

$$f(x) = tanh(4.9 \times 0.5x) \tag{7}$$

The GA that evolves the neural networks uses the operators presented before (Fig. 3 and Fig. 2), in addition to the listed modifications. The selection for crossover uses k tournament, and elitism is adopted to preserve the best individuals from each generation. The used parameters are listed in S1-Table in the Supplementary Data and were chosen based on experimental results and literature revision (Papavasileiou and Jansen, 2017a). The selected genes are the subset of inputs directly or indirectly connected to the output node in the neural network with the best fitness at the end of the algorithm.

Results

Microarray Classification and Gene Selection

The classifiers performance was measured by the geometric mean (G-mean) as presented in Sun et al. (2007) because of the presence of datasets with class imbalance, and also by the accuracy. Statistical difference, when applicable, was measured with the Post Hoc Kruskal Wallis Dunn Test with Bonferroni adjustment from the *PMCMR* R package (Pohlert, 2014), with a significance value of $p < 0.01$. Two versions of the proposed method were tested against each other. The first one, called N3O, employs the structural changes in Neuroevolution presented in the last Section, while the other, FS-NEAT, only uses the regular algorithm as described in the literature. The proposed method is being called N3O here only as a way to facilitate the distinction in the text, but it can be considered a new variation or addition to the FS-NEAT, and not necessarily a new algorithm by itself. Both employed the same fitness, neuron structure, preprocessing and filtering steps, including the KW test, and received as input the same set of genes. The hyperparameters were the same, the only difference being that FS-NEAT had a larger probability for the "add connection" mutation, in order to compensate for the use of the two mutations in N3O.

From the results in S2-Table in the Supplementary Data, after 31 runs, N3O consistently achieved better accuracy and G-mean than regular FS-NEAT. Considering the number of selected genes, N3O performed better than FS-NEAT and was able to provide smaller solutions with at least the same predictive power. N3O also showed less variance in the number of selected genes than FS-NEAT for all studied cases.

Taking as an example a single run with the dataset GSE71935 (leukemia), it is possible to see this difference in the two final networks created by each method in S4-Figure in the Supplementary Data. Their reported accuracies for the testing set were 100.0% for N3O and 75.0% for FS-NEAT, selecting 6 and 11 genes, respectively. As can be observed, FS-NEAT required a larger neural network structure than N3O. This is also visible in S5-Figure in the Supplementary Data. Both algorithms showed roughly the same regularized error convergence, and the total number of genes visited during the evolution, as well as a similar amount of new genes being explored at each generation. FS-NEAT, however, kept a larger number of genes in the population at each generation, making for larger networks. The difference between the genes selection of N3O and FS-NEAT is even more visible in S6-Figure in the Supplementary Data. Despite visiting the same total amount of genes considering all generations, N3O required fewer features at each generation, and showed a better spread of genes among the individuals in the population.

To validate our method, from now on also referred to as N3O because of the three new structural operators, we tested it on all datasets listed in Table 1 using stratified 3-fold cross-validation, dividing the data in three folds that kept the same distribution of samples per class as the total dataset. Cross-validation is an efficient and unbiased error estimator, ideal for microarray data, and the use of three folds is in agreement with other works in the literature (Ang et al., 2016). In this particular case, a larger number of folds would not be suitable for the studied data as some of the datasets contain classes with few samples and would be underrepresented in some of the folds. For each iteration of the cross-validation, we performed the filtering using the KW test, the normalization, and the overall evolution in two folds and tested the G-mean, accuracy, and FS on the third fold. The code uses some methods from *Scikit-learn* (Pedregosa et al., 2011) and *NEAT-Python*[3] libraries.

Tables 2 and 3 show the reported G-mean and the number of genes selected by our method for all datasets over 31 runs of the cross-validation. S3-Table in the Supplementary Data shows the obtained accuracy for the same data. For datasets with more than two classes, it is indicated which class was being discriminated against the others. The "Proportion" column is the proportion of samples belonging to the class being discriminated.

Table 4 reports the most selected genes for each dataset, considering the experiments in Table 3. It shows which genes appeared as selected the most in the final networks. The gene ERBB2 (HER2), for instance, was the most selected gene for the dataset GSE45827 - HER (breast cancer), being selected by 90.6% of the neural networks, while the gene SCNN1B was the most selected gene for the dataset GSE8671 (colorectal cancer), but appeared only in 3.1% of the networks. Even those genes with a small number of repetitions are significant, however, when the probability of it happening at random is considered, what, as shown in the fifth column of Table 4, is highly unlikely.

[3] `http://neat-python.readthedocs.io/en/latest/`

Table 2: **Stratified 3-fold cross-validation statistical report of G-mean for N3O.**

Datasets	Class	Proportion	Mean±std	Median	Min-Max
GSE42568		0.87	0.941 ± .027	0.931	0.88 - 1.00
GSE45827	Basal	0.27	0.912 ± .022	0.907	0.87 - 0.95
	HER	0.20	0.910 ± .036	0.919	0.83 - 0.96
	Cell Line	0.09	0.979 ± .025	0.996	0.93 - 1.00
	Luminal A	0.19	0.900 ± .034	0.901	0.81 - 0.96
	Luminal B	0.20	0.817 ± .050	0.828	0.70 - 0.88
	Normal	0.05	0.905 ± .080	0.926	0.75 - 1.00
GSE10797	Cancer Epithelial	0.42	0.724 ± .056	0.720	0.58 - 0.83
	Cancer Stroma	0.42	0.733 ± .039	0.736	0.67 - 0.83
	Normal	0.15	0.806 ± .071	0.806	0.63 - 0.94
GSE44076		0.50	0.982 ± .010	0.985	0.97 - 1.00
GSE44861		0.50	0.822 ± .031	0.826	0.74 - 0.87
GSE8671		0.49	0.983 ± .018	0.984	0.94 - 1.00
GSE21510		0.42	0.954 ± .033	0.959	0.88 - 1.00
GSE32323		0.48	0.938 ± .041	0.939	0.84 - 1.00
GSE41328		0.44	0.963 ± .051	1.000	0.82 - 1.00
GSE9476	AML	0.41	0.886 ± .040	0.883	0.81 - 0.97
	Bone Marrow	0.16	0.979 ± .040	1.000	0.84 - 1.00
	Bone Marrow CD34	0.13	0.900 ± .097	0.927	0.61 - 1.00
	PB	0.16	0.983 ± .028	1.000	0.89 - 1.00
	PBSC CD34	0.16	0.962 ± .053	0.991	0.77 - 1.00
GSE14317		0.72	0.949 ± .063	0.972	0.73 - 1.00
GSE63270		0.59	0.971 ± .021	0.971	0.90 - 1.00
GSE71935		0.80	0.783 ± .126	0.794	0.46 - 0.97
Golub et al. (1999)		0.65	0.886 ± .036	0.887	0.82 - 0.97

Reported values from 31 runs of the stratified 3-fold cross-validation. Proportion = Proportion of samples belonging to the class; Std = Standard deviation; Min = Minimum value reported in all runs; Max = Maximum value reported in all runs.

To further validate this selection, Table 4 brings a literature review of the most selected genes, considering the PubMed[4] repository. In total, 44% of those genes were already described in the literature as being relevant for the specific cancer type of their corresponding dataset, 20% were described as relevant for other cancer types, 20% were not described as relevant for any cancer type, and 16% were not yet described in the literature. Interestingly, the aforementioned gene ERBB2 (HER2) was the most selected gene in its dataset among all experiments, while also being described as one of the most relevant genes in breast cancer in general (Borges et al., 2018; Nattestad et al., 2018; Liu et al., 2018a; Soares et al., 2018).

Since the literature points to SVM as being the best classifiers of microarray data, our method was compared with the G-mean of a SVM with RBF kernel and hyperparameters tuned by grid search in three configurations: (i) over the original dataset (Table 5, column 4); (ii) after filtering the genes with KW (Table 5, column 5); (iii) using only the genes selected by our method (Table 5, column 6). The tests were performed using stratified 3-fold cross-validation over 31

[4] `https://www.ncbi.nlm.nih.gov/pubmed/`

Table 3: **Stratified 3-fold cross-validation statistical report of FS for N3O.**

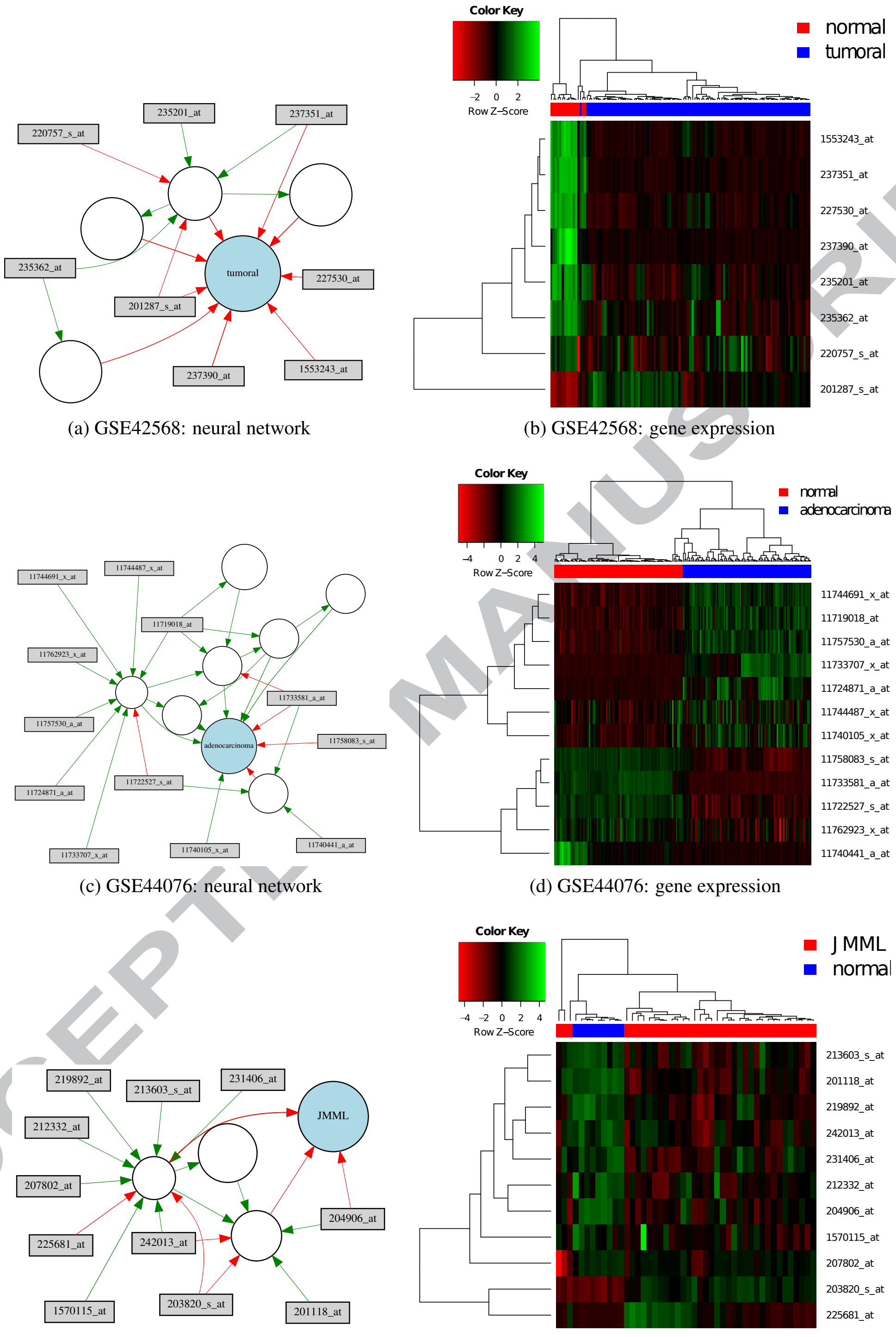
Datasets	Class	G-mean	Mean±std	Median	Min-Max
GSE42568		0.941	11.44 ± 3.12	10.67	6.33 - 19.00
GSE45827	Basal	0.912	11.76 ± 2.61	12.00	7.33 - 19.67
	HER	0.910	10.57 ± 2.63	10.33	5.33 - 17.00
	Cell Line	0.979	10.34 ± 3.97	09.67	4.33 - 21.00
	Luminal A	0.900	11.41 ± 2.02	11.33	6.00 - 16.00
	Luminal B	0.817	14.11 ± 2.38	14.00	10.0 - 18.33
	Normal	0.905	13.05 ± 4.51	12.00	7.33 - 26.00
GSE10797	Cancer Epithelial	0.724	13.65 ± 2.36	13.33	9.33 - 18.00
	Cancer Stroma	0.733	13.85 ± 2.76	13.33	7.67 - 20.00
	Normal	0.806	12.92 ± 4.19	13.00	6.67 - 20.33
GSE44076		0.982	09.65 ± 2.66	10.00	4.00 - 15.00
GSE44861		0.822	11.37 ± 2.55	10.67	6.67 - 16.33
GSE8671		0.983	15.16 ± 3.99	15.00	4.00 - 21.67
GSE21510		0.954	13.10 ± 4.47	13.00	3.00 - 22.00
GSE32323		0.938	15.74 ± 4.02	16.00	4.67 - 23.00
GSE41328		0.963	18.67 ± 6.35	18.67	3.00 - 29.33
GSE9476	AML	0.886	13.57 ± 2.80	13.00	8.00 - 19.00
	Bone Marrow	0.979	13.63 ± 3.61	13.67	5.67 - 20.33
	Bone Marrow CD34	0.900	12.52 ± 3.40	12.67	4.00 - 20.67
	PB	0.983	14.41 ± 4.77	13.67	5.00 - 26.33
	PBSC CD34	0.962	12.87 ± 3.62	13.33	7.00 - 19.00
GSE14317		0.949	14.80 ± 4.76	14.00	3.00 - 22.33
GSE63270		0.971	12.03 ± 3.11	12.33	5.33 - 18.00
GSE71935		0.783	14.60 ± 3.42	14.67	7.33 - 22.00
Golub et al. (1999)		0.886	12.51 ± 2.43	12.33	8.00 - 17.00

Reported values from 31 runs of the stratified 3-fold cross-validation. Average G-mean as reported from Table 2. Std = Standard deviation; Min = Minimum value reported in all runs; Max = Maximum value reported in all runs.

runs, and the results are in Table 5. The analogous results using accuracy instead of G-mean are shown in S4-Table in the Supplementary Data.

To further validate our method, we compared its results with the results from a recent Neuroevolution method that uses the Artificial Bee Colony (ABC) algorithm to select genes and Differential Evolution (DE) to design neural networks for microarray classification (Garro et al., 2017). The results are listed in Table 6. We compared the reported accuracy for this method on the testing set using random partitions. The reported number of selected genes was fixed at three by the authors based on experimentation.

We also evaluate the topology of ANNs found by the proposed method. From the different examples of network topologies in Fig. 4, an observation that can be made is that the proposed method found ANN architectures distinct from traditional Multilayer Perceptron models, unlikely to be designed by programmers. Many inputs are directly connected to the output, and the algorithm makes use of gates akin to Highway Networks, usually employed to improve the learning of very deep neural networks allowing information to flow between layers unrestricted (Srivastava et al., 2015).



(a) GSE42568: neural network

(b) GSE42568: gene expression

(c) GSE44076: neural network

(d) GSE44076: gene expression

(e) GSE71935: neural network

(f) GSE71935: gene expression

Figure 4: **Neural networks and corresponding selected genes expression for the three types of studied cancers**. The results are represented by Breast GSE42568 (a, b), Colorectal GSE44076 (c, d), and Leukemia GSE71935 (e, f). In the neural networks, rectangles represent the inputs, blue circles represent outputs and hidden nodes are represented by the white circles. Red connections identify negative weights, whereas green connections characterize positive ones. The thickness of the lines is proportional to the module of the connection weights. The heatmaps rows (genes) and columns (samples) were grouped by hierarchical clustering (dendrograms) using their correlation as a distance metric. The red and blue bar at the top shows the true label of each sample.

Table 4: **Most selected genes by N3O for each dataset.**

Datasets	Class	Gene	d	P	References
GSE42568		🟩 LYVE1	0.052	$2.05e^{-11}$	Martínez-Iglesias et al. (2016), Newman et al. (2012)
GSE45827	Basal	🟩 MLPH	0.219	$3.33e^{-16}$	Thakkar et al. (2010), Thakkar et al. (2015)
	HER	🟩 ERBB2 (HER2)	0.906	0.00	Borges et al. (2018), Nattestad et al. (2018), Liu et al. (2018a), Soares et al. (2018)
	Cell Line	AA702946	0.042	$3.68e^{-9}$	
	Luminal A	🟩 PGR	0.135	0.00	Kunc et al. (2018)
	Luminal B	Hs.444858	0.125	$4.66e^{-15}$	
	Normal	C7orf41	0.031	$1.74e^{-6}$	
GSE10797	Cancer Epithelial	🟦 BPY2	0.125	$1.55e^{-15}$	Dasari et al. (2002)
	Cancer Stroma	🟥 UBR2	0.104	$2.44e^{-15}$	
	Normal	🟥 KIT	0.354	$1.11e^{-16}$	
GSE44076		🟩 GREM2	0.062	$4.01e^{-14}$	Liu et al. (2011)
GSE44861		ENSG00000253701	0.250	0.00	
GSE8671		🟩 SCNN1B	0.031	$2.71e^{-6}$	Shangkuan et al. (2017)
GSE21510		🟦 GPSM2	0.031	$1.76e^{-6}$	Liu et al. (2015)
GSE32323		🟥 C4orf43	0.042	$1.96e^{-8}$	
GSE41328		🟩 SLC7A5	0.042	$3.87e^{-8}$	Kalmar et al. (2013)
GSE9476	AML	🟩 ALDH1A1	0.187	0.00	Longville et al. (2015), Gasparetto and Smith (2017)
	Bone Marrow	🟩 GYPA	0.125	$3.33e^{-16}$	Li et al. (2015)
	Bone Marrow CD34	🟦 TMSB15A	0.073	$1.65e^{-13}$	Darb-Esfahani et al. (2012)
	PB	🟥 GK	0.083	$6.44e^{-15}$	
	PBSC CD34	🟦 CACNB2	0.114	0.00	Tomoshige et al. (2015), Chen et al. (2016)
GSE14317		🟩 DAPK1	0.052	$6.41e^{-9}$	Tao et al. (2015), Ng et al. (2014) Celik et al. (2015)
GSE63270		🟥 AMT	0.114	$3.33e^{-15}$	
GSE71935		🟦 PCOLCE2	0.125	0.00	Thutkawkorapin et al. (2016)
Golub et al. (1999)		🟩 GST	0.281	$1.11e^{-15}$	Lavrov et al. (2017), Tang et al. (2018b)

Reported values from 31 runs of the stratified 3-fold cross-validation. Gene = gene corresponding to the probe; d = how many times the most selected gene was selected (proportion). P = probability of the most selected gene being randomly selected with proportion d or larger (if 0.00 the system lacked enough float precision to represent the number). The computation of P assumes a binomial distribution and considers the probability of a gene being randomly selected by a single neural network as the average number of final inputs of a network over the total number of genes in the dataset. If the gene is (i) green: reported in the literature as relevant for the corresponding cancer type of the dataset; (ii) blue: reported as relevant for another cancer type; (iii) red: not reported as relevant for any cancer type; (iv) white: gene not described.

Expression Patterns and Gene Selection

By applying our method, a set of genes representing an expression pattern was extracted from each GSE when creating ANNs considering all available samples (three examples can be seen at Fig. 4). Table 7 lists: (i) the number of genes that were selected for each GSE, per class. In this sense, the algorithm selects the set of genes that differ in a given condition from the other. For the GSEs with more than two classes, we only discuss the gene expression patterns that are exclusive of the tumoral classes; (ii) the number of genes that were already associated to the GSE's cancer type; (iii) the quantity of long non-coding (LnC) RNAs; (iv) the amount of genes that were not found to be related to any type of cancer, or that don't have a clear described function, such as predicted genes; and (v) the number of genes that were not observed to be related to the GSE's

Table 5: **G-mean comparison of N3O and SVM.**

Datasets	Class	N3O	SVM	KW&SVM	N3O&SVM
GSE42568		0.941 ± .027	0.939 ± .030	0.941 ± .025	**0.961** ± .023
GSE45827	Basal	0.912 ± .022	**0.957** ± .004	0.956 ± .008	**0.957** ± .020
	HER	0.910 ± .036	0.921 ± .022	0.891 ± .031	**0.939** ± .067
	Cell Line	0.979 ± .025	**1.000** ± .000	**1.000** ± .000	0.994 ± .017
	Luminal A	0.900 ± .034	0.942 ± .042	**0.961** ± .017	0.933 ± .043
	Luminal B	0.817 ± .050	0.841 ± .035	0.839 ± .052	**0.850** ± .047
	Normal	0.905 ± .080	**0.949** ± .035	0.935 ± .024	0.936 ± .059
GSE10797	Cancer Epithelial	0.724 ± .056	0.836 ± .032	0.821 ± .044	**0.841** ± .056
	Cancer Stroma	0.733 ± .039	0.755 ± .043	0.779 ± .031	**0.817** ± .068
	Normal	0.806 ± .071	0.698 ± .099	**0.895** ± .044	0.891 ± .049
GSE44076		0.982 ± .010	0.983 ± .003	0.984 ± .003	**0.987** ± .008
GSE44861		0.822 ± .031	0.825 ± .055	0.771 ± .065	**0.987** ± .008
GSE8671		**0.983** ± .018	0.614 ± .096	0.568 ± .000	0.568 ± .000
GSE21510		0.954 ± .033	0.988 ± .019	**0.990** ± .017	0.984 ± .046
GSE32323		**0.938** ± .041	0.602 ± .097	0.555 ± .007	0.591 ± .077
GSE41328		**0.963** ± .051	0.550 ± .109	0.557 ± .091	0.612 ± .000
GSE9476	AML	0.886 ± .040	0.939 ± .017	0.904 ± .026	**0.947** ± .044
	Bone Marrow	0.979 ± .040	0.949 ± .000	**0.993** ± .017	0.990 ± .025
	Bone Marrow CD34	0.900 ± .097	**0.987** ± .031	0.911 ± .093	0.938 ± .070
	PB	0.983 ± .028	0.951 ± .045	**1.000** ± .000	0.997 ± .013
	PBSC CD34	0.962 ± .053	0.953 ± .032	**0.993** ± .017	0.989 ± .036
GSE14317		0.949 ± .063	0.917 ± .086	0.983 ± .049	**0.994** ± .019
GSE63270		0.971 ± .021	**0.999** ± .004	0.998 ± .004	0.991 ± .012
GSE71935		0.783 ± .126	0.733 ± .172	0.794 ± .119	**0.931** ± .070
Golub et al. (1999)		0.886 ± .036	0.951 ± .027	**0.970** ± .017	0.930 ± .036
Average		**0.903** ± .079	0.871 ± .135	0.880 ± .140	0.902 ± .128

The G-mean is the result of 31 runs of the stratified 3-fold cross-validation. All SVM versions used the RBF kernel and had their hyperparameters tuned by grid search. N3O = average G-mean of the proposed method; SVM = average G-mean of SVM; KW&SVM = average G-mean of SVM after filtering the data with Kruskal-Wallis H Test; N3O&SVM = average G-mean of SVM using only the genes selected by the proposed method. In bold is the best average G-mean of each dataset. Best results with statistical significance ($p < 0.01$) are marked in blue.

Table 6: **Comparison of N3O with another Neuroevolution method.**

Method	Dataset	Accuracy	FS
N3O	Golub et al. (1999)	0.917 ± .095	6.27 ± 2.38
ABC&DE	Golub et al. (1999)	0.912 ± .067	3

N3O = average accuracy and number of selected features of our method for the testing set (20%) with random partition over 30 repetitions; ABC&DE = accuracy reported by the method from (Garro et al., 2017) for the testing set (20%) with random partition over 30 repetitions; FS = number of selected features.

cancer type, but found in others. The complete list of selected genes and their associated cancer type can be found in S5-Table in the Supplementary Data. In summary, among the 177 genes, 82 genes were already associated with their given cancer type (LnC RNA apply here), 5 were LnC RNAs, 44 are not yet related to the GSE's cancer type, but were observed to be altered in other cancer types, and a total of 50 genes didn't return any hits from the scientific literature search, either because they don't possess a clear described function, or were just not related to any tumoral condition (LnC RNA apply here) (see S5-Table for the full list). Interestingly, each expression pattern was unique, and only the REC8 Meiotic Recombination Protein (REC8) was common between a set of Leukemia and one of Colorectal cancer (CRC).

To understand the nature of the obtained expression patterns, we employed the DAVID tool to search for the significant Gene Ontologies (GO) and cellular localization of the 177 selected genes, which will be discussed in the next Section.

Table 7: **Number of associated genes obtained from each GSE.**

GSEs-Cancer	Genes	Hits	lncRNA	NHF	Other
GSE42568 - Breast Cancer	8	4	1	2	1
GSE45827 - Breast Basal	9	5	NA	2	2
GSE45827 - Breast LuminalA	6	5	1	1	NA
GSE45827 - Breast LuminalB	7	2	NA	5	NA
GSE45827 - Breast HER	6	3	NA	3	NA
GSE10797 - Breast Epithelium	8	5	1	1	2
GSE10797 - Breast Stromal	12	6	NA	3	3
GSE44076 - CRC Adenocarcinoma	12	7	NA	2	3
GSE44861 - CRC	8	5	NA	2	1
GSE8671 - CRC Adenoma	23	10	1	7	6
GSE21510 - CRC	9	1	NA	3	5
GSE32323 - CRC	6	2	NA	1	3
GSE41328 - CRC Adenocarcinoma	24	10	NA	7	7
GSE9476 - AML	18	8	NA	6	4
GSE14317 - ATL	4	1	NA	2	1
GSE63270 - AML	6	3	1	2	1
GSE71935 - JMML	11	5	NA	1	5

Hits = Genes that were already observed to be expressed in the GSE's cancer type; lncRNA = Long non-coding RNA; NHF = No Hits Found. Number of genes that were either not found to be related to any type of cancer in the scientific literature, or that don't have a clear described function so far; Other = Number of genes not observed in the GSE's cancer type, but already found to be expressed in other types of cancer; NA = Not Applicable; CRC = Colorectal Cancer; AML = Acute Myeloid Leukemia; ATL = Adult T-Cell Leukemia/Lymphoma; JMML = Juvenile myelomonocytic Leukemia; HER = Breast Cancer - HER Status.

Table 8: **Major GO derived from all selected genes.**

Bioprocesses	corr p-Value
Extracellular Matrix Organization	1.9×10^{-1}
Response to Hypoxia	7.7×10^{-1}
Signal Transduction	8.6×10^{-1}
Positive Regulation of Cell proliferation	8.0×10^{-1}

Discussion

Classification and selection generalization

Some of the most popular classifiers in ML, neural networks and Deep Learning (DL), are extensively used in Bioinformatics (Park and Kellis, 2015; Angermueller et al., 2016; Mamoshina et al., 2016; Min et al., 2017; Ching et al., 2018), but fall short in microarray classification (Pirooznia et al., 2008; Lee et al., 2005). The most common way to create an ANN is by defining a fixed topology (layers, nodes, and connections) and train it with an algorithm such as backpropagation (LeCun et al., 1998) to set the values of the biases and weights. In Bioinformatics, however, where many of the concepts underlying biological process are only partially known (Grisci and Dorn, 2017), this can be an issue, since the design of a topology involves some prior knowledge about the problem and this structure can impact on the final predictive power of the network. Moreover, part of the motive for ANNs being behind other ML methods in microarray classification, such as SVM and Random Forests (Pirooznia et al., 2008; Lee et al., 2005) has been attributed to the use of gradient-based optimization methods (Gupta et al., 2015). Neuroevolution avoids some of the pitfalls encountered by the need of having a fixed topology and by backpropagation, regarding microarray data, and some of its components have already been used when dealing with this task (Garro et al., 2017; Gupta et al., 2015; Luque-Baena et al., 2013).

From the G-mean and accuracy results in Table 2 and S3-Table, the described Neuroevolution method was able to successfully classify the microarray data, beating the baseline in all datasets. It also showed competitive results against SVM, which is considered the best class of algorithms for this problem, as seen in Table 5 and S4-Table. Our approach performance was also comparable to other recent Neuroevolution method, as presented in Table 6, displaying favorable results. The method was able to perform the classification task simultaneously with gene selection and autonomously, without the need for any previous threshold or user decision, which is important when considering that the number of optimal genes is different for each dataset (Statnikov et al., 2008). From the average number of genes selected in Table 3, our approach was able to perform a reduction over 99.9% in the number of dimensions in all datasets.

It is known that gene selection performed with a classifier is only specific to that given algorithm, meaning that there is no guarantee that the selected features will have a good performance with other methods (Ang et al., 2016). Furthermore, SVMs are usually insensitive to a large number of irrelevant genes, and feature selection often biases down their accuracy (Statnikov et al., 2008). Even so, when the genes selected by the proposed Neuroevolution method were applied to SVM, its performance was not hurt, and for most of the datasets, it actually had a slight improvement, as shown in the last column of Table 5. This result suggests that the selected genes are not methodological artifacts, and could be generalized and further explored even by different algorithms.

Biological role of selected genes

The most important factor to be observed is that the groups of selected genes for each class in each GSE correspond to the expression pattern that makes that particular tumoral state unique. Thus, in a biological view, it is expected that each set of genes do not overlap each other, as observed by our results (S5-Table). The only exception was the gene REC8, that was selected in one case of Leukemia (GSE63270) and Colorectal cancer (GSE44861). However, in both cases, it was associated with another type of cancer, becoming a potential target instead of a bias. A total of 50 genes were either not related to any type of cancer or didn't possess a clear functional description, leaving 127 genes that were already described to be expressed in some type of cancer. It is important to highlight that selecting genes with no described function yet is normal to any expression analysis. The human genome is still replete of known DNA segments with no described functions (e.g. putative genes or open reading frames or pseudogenes) that can impact on cancer biology (Tutar et al., 2016; Emadi-Baygi et al., 2017; Poliseno et al., 2015; Shi et al., 2018; Wedge et al., 2018), and studies like these become fundamental to provide the first glimpse of the functional role for such genes. Moreover, among those 127 genes, 82 (64.5%) were related to their specific cancer types, and 44 were observed to be altered in some way in other types, becoming excellent potential targets to be explored in future works. Thus, our analysis satisfactorily selected expression patterns that are in agreement to their tumoral background, endorsing the proper manipulation and application of our approach. All genes are described in S5-Table, with their associated cancer types and corresponding references.

Concerning the cellular component, the majority of the genes were related to extracellular exosomes, cell surface, plasma membrane, endoplasmatic reticulum and the cytosol (S7-Figure in the Supplementary Data). In this sense, one curious aspect is that, in their majority, the selected genes are components that act in the plasma membrane (31.1%) and extracellular exossomes (26.4%) (S7-Figure). In addition, the GO analysis showed that the main bioprocesses were extracellular matrix organization, response to hypoxia, signal transduction, and positive regulation of cell proliferation (Table 8). In fact, the extracellular matrix (ECM) and exossomes are fundamental in cancer biology. The ECM environment possesses proteins related to cell adhesion and cytoskeleton organization that are fundamental for tumor invasion and colony formation, assembling the tumoral microenvironment for metastasis (Saitoh, 2018; Gkretsi and Stylianopoulos, 2018). Moreover, exossomes are critical for cell-cell signaling, as well for their role as carriers for a diverse array of biomolecules, strongly influencing not only normal cellular functions but also pathological conditions, such as cancer (Couto et al., 2018; Maia et al., 2018). In addition, the plasma membrane is part of a dynamic system of external and internal signals through membrane-bound transcription factor and cellular compartments that modulate gene expression, proteins folding and cellular transduction pathways, impacting in virtually all cellular functions, but intimately associated to cancer molecular mechanisms (Liu et al., 2018b; Filippini et al., 2018; Stuelten et al., 2018). In fact, as can be seen in S5-Table, the vast majority of the selected genes are membrane and transmembrane components, as well for kinases, cytoskeleton elements, and transcription factors that impact on the expression of cell cycle genes. The most common biological processes associated with the selected genes are in agreement to the biochemical function of the gene sets and cancer biology.

Another interesting fact is that our approach selected five LnC RNA (Table 7. In contrast to mRNAs, LnC RNA do not encode to proteins but are critical transcriptional regulators that modulate gene expression through multiple molecular mechanisms (Hu et al., 2018; Chan and Tay, 2018). Among the five LnC RNA selected by our approach, PVT1 was selected in breast cancer (GSE10797). Remarkably, PVT1 was already associated with triple-negative breast cancer, in which PVT1 increases KLF5 protein stability and regulate β-catenin, which is related to poor prognosis in breast cancer (Tang et al., 2018a). This further validates the accuracy of our approach in selecting relevant targets.

Finally, it is also important to highlight the importance of dataset treatment prior to any ML analysis. As we mentioned before, raw data contains noise that can severely affect the quality of the result, and blindly applying a classification technique over noisy data can impact on the biological relevance of the selected genes. In addition, it is common for microarray datasets to exhibit some samples with overall bad quality that should be removed to improve the obtained results.

Conclusion

In this work, we developed a pipeline for microarray classification and gene selection by employing Neuroevolution as a ML method capable of efficaciously perform both tasks autonomously. This evolutive method builds upon the FS-NEAT algorithm, adding new operators for better exploration of the search space, and designs unique neural networks for solving the desired tasks. Tested with microarray datasets of three different types of cancer, with varying number of samples, features, and classes, our strategy successfully overcame the baseline and showed good performance against other algorithms. Especially in the case of SVMs, the use of the features selected by our method did not disturb the classification and, for some cases, even improved it, a result not expected in the literature and that may show the strength of the performed selection. Our results also pointed to 177 genes involved in specific gene expression patterns that are closely associated to extracellular matrix, plasma membrane and exosomes, proposing new targets to be explored to uncover the molecular mechanisms underlying colorectal cancer, leukemia, and breast cancer. The successful validation of our targets in the literature also reinforces the efficacy of our approach to correctly classify expression pattern in different types of cancer. We also highlight that a proper preprocessing step of microarray datasets prior to the ML pipeline can assure better biological results.

Acknowledgments

This work was supported by grants from FAPERGS [16/2551-0000520-6], MCT/CNPq [311022/2015-4], CAPES-STIC AMSUD [88887.135130/2017-01] - Brazil, CAPES, Alexander von Humboldt-Stiftung (AvH) [BRA 1190826 HFSTCAPES-P] - Germany, CAPES, and by Microsoft Corporation under a Microsoft Azure for Research Award. We also would like to thank the Asociación de Universidades Grupo Montevideo (AUGM), and Eduardo Bassani Chandelier, Manuel Villalobos Cid, and Mario Inostroza-Ponta for the technical and theoretical support.

Data Availability Statement

The necessary source codes and datasets used for the experiments can be accessed in GitHub: `https://github.com/sbcblab/NEAT-Microarray.git` The password for unzipping the folder with the source code is ”sbcbjbi2018”. The curated data generated for this work can be accessed in the CuMiDa database: `http://sbcb.inf.ufrgs.br/cumida`.

References

Allison, D., Cui, X., Page, G., Sabripour, M., 2006. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7, 55–65.

Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A., 2016. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM transactions on computational biology and bioinformatics 13, 971–989.

Angermueller, C., Pärnamaa, T., Parts, L., et al., 2016. Deep learning for computational biology. Mol Syst Biol 12, 878.

Blohm, D., Guiseppi-Elie, A., 2001. New developments in microarray technology. Current Opinion in Biotechnology 12, 41–47.

Borges, V.F., Ferrario, C., Aucoin, N., Falkson, C., Khan, Q., Krop, I., Welch, S., Conlin, A., Chaves, J., Bedard, P.L., et al., 2018. Tucatinib combined with ado-trastuzumab emtansine in advanced erbb2/her2-positive metastatic breast cancer: A phase 1b clinical trial. JAMA oncology .

Celik, S., Akcora, D., Ozkan, T., Varol, N., Aydos, S., Sunguroglu, A., 2015. Methylation analysis of the dapk1 gene in imatinib-resistant chronic myeloid leukemia patients. Oncology letters 9, 399–404.

Chan, J., Tay, Y., 2018. Noncoding rna:rna regulatory networks in cancer. Int J Mol Sci 19, pii: E1310.

Chen, M., Rothman, N., Ye, Y., Gu, J., Scheet, P.A., Huang, M., Chang, D.W., Dinney, C.P., Silverman, D.T., Figueroa, J.D., et al., 2016. Pathway analysis of bladder cancer genome-wide association study identifies novel pathways involved in bladder cancer development. Genes & cancer 7, 229.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.M., Zietz, M., Hoffman, M.M., et al., 2018. Opportunities and obstacles for deep learning in biology and medicine. bioRxiv , 142760.

Couto, N., Caja, S., Maia, J., Strano Moraes, M., Costa-Silva, B., 2018. Exosomes as emerging players in cancer biology. Biochimie pii: S0300-9084, 30067–1.

Darb-Esfahani, S., Kronenwett, R., Von Minckwitz, G., Denkert, C., Gehrmann, M., Rody, A., Budczies, J., Brase, J., Mehta, M., Bojar, H., et al., 2012. Thymosin beta 15a (tmsb15a) is a predictor of chemotherapy response in triple-negative breast cancer. British journal of cancer 107, 1892.

Dasari, V.K., Deng, D., Perinchery, G., Yeh, C.c., Dahiya, R., 2002. Dna methylation regulates the expression of y chromosome specific genes in prostate cancer. The Journal of urology 167, 335–338.

Davis, S., Meltzer, P., 2007. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. Bioinformatics 14, 1846–1847.

Deniz, E., Erman, B., 2017. Long noncoding rna (lincrna), a new paradigm in gene expression control. Funct Integr Genomics 17, 135–143.

Ding, S., Li, H., Su, C., Yu, J., Jin, F., 2013. Evolutionary artificial neural networks: a review. Artificial Intelligence Review , 1–10.

Eiben, A.E., Smith, J.E., 2015. Introduction to Evolutionary Computing. 2nd ed., Springer.

Emadi-Baygi, M., Sedighi, R., Nourbakhsh, N., P, N., 2017. Pseudogenes in gastric cancer pathogenesis: a review article. Brief Funct Genomics 16, 348–360.

Epstein, C., Butow, R., 2000. Microarray technology - enhanced versatility, persistent challenge. Current Opinion in Biotechnology 11, 36–41.

Filippini, A., Sica, G., D'Alessio, A., 2018. The caveolar membrane system in endothelium: From cell signaling to vascular pathology. J Cell Biochem , doi: 10.1002/jcb.26793.

Garro, B.A., Rodríguez, K., Vazquez, R.A., 2017. Designing artificial neural networks using differential evolution for classifying dna microarrays, in: Evolutionary Computation (CEC), 2017 IEEE Congress on, IEEE. pp. 2767–2774.

Gasparetto, M., Smith, C.A., 2017. Aldhs in normal and malignant hematopoietic cells: Potential new avenues for treatment of aml and other blood cancers. Chemico-biological interactions 276, 46–51.

Gautier, L., Cope, L., Bolstad, B., Irizarry, R., 2004. affy - analysis of affymetrix genechip data at the probe level. Bioinformatics 20, 307–315.

Gkretsi, V., Stylianopoulos, T., 2018. Cell adhesion and matrix stiffness: Coordinating cancer cell invasion and metastasis. Front Oncol , doi: 10.3389/fonc.2018.00145.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science 286, 531–537.

Gorreta, F., Carbone, W., Barzaghi, D., 2012. Genomic profiling: cdna arrays and oligoarrays. Methods Mol Biol 823, 89–105.

Grisci, B., Dorn, M., 2017. Neat-flex: Predicting the conformational flexibility of amino acids using neuroevolution of augmenting topologies. Journal of Bioinformatics and Computational Biology , 1750009.

Grisci, B.I., Feltes, B.C., Dorn, M., 2018. Microarray classification and gene selection with fs-neat, in: 2018 IEEE Congress on Evolutionary Computation (CEC), IEEE. pp. 1–8.

Gupta, A., Bhalla, S., Dwivedi, S., Verma, N., Kala, R., 2015. On the use of local search in the evolution of neural networks for the diagnosis of breast cancer. Technologies 3, 162–181.

Haykin, S.S., 2009. Third ed., Pearson Education, Upper Saddle River, NJ.

Hu, G., Niu, F., Humburg, B., Liao, K., Bendi, S., Callen, S., et al., 2018. Molecular mechanisms of long noncoding rnas and their role in disease pathogenesis. Oncotarget 9, 18648–18663.

Huang, D., Sherman, B., Lempicki, R., 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37, 1–13.

Huang, D., Sherman, B., Lempicki, R., 2009b. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nature Protoc 4, 44–57.

Kalmar, A., Wichmann, B., Galamb, O., Spisák, S., Tóth, K., et al., 2013. Gene expression analysis of normal and colorectal cancer tissue samples from fresh frozen and matched formalin-fixed, paraffin-embedded (ffpe) specimens after manual and automated rna isolation. Methods 59, S16–S19.

Kauffmann, A., Gentleman, R., Huber, W., 2009. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. Bioinformatics 25, 415–416.

Kauffmann, A., Huber, W., 2010. Microarray data quality control improves the detection of differentially expressed genes. Genomics 95, 138–142.

Kunc, M., Biernat, W., Senkus-Konefka, E., 2018. Estrogen receptor-negative progesterone receptor-positive breast cancer–"nobody's land "or just an artifact? Cancer treatment reviews 67, 78–87.

Lan, L., Vucetic, S., 2011. Improving accuracy of microarray classification by a simple multi-task feature selection filter. International journal of data mining and bioinformatics 5, 189–208.

Lavrov, A.V., Ustaeva, O.A., Adilgereeva, E.P., Smirnikhina, S.A., Chelysheva, E.Y., Shukhov, O.A., Shatokhin, Y.V., Mordanov, S.V., Turkina, A.G., Kutsev, S.I., 2017. Copy number variation analysis in cytochromes and glutathione s-transferases may predict efficacy of tyrosine kinase inhibitors in chronic myeloid leukemia. PloS one 12, e0182901.

LeCun, Y., Bottou, L., Orr, G.B., Müller, K.R., 1998. Efficient backprop, in: Neural networks: Tricks of the trade. Springer, pp. 9–50.

Lee, J.W., Lee, J.B., Park, M., Song, S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48, 869–885.

Leung, Y., Cavalieri, D., 2003a. Fundamentals of cdna microarray data analysis. Trends Genet 19, 649–659.

Leung, Y.F., Cavalieri, D., 2003b. Fundamentals of cdna microarray data analysis. TRENDS in Genetics 19, 649–659.

Li, J., Zhai, X., Wang, H., Qian, X., Miao, H., Zhu, X., 2015. Bioinformatics analysis of gene expression profiles in childhood b-precursor acute lymphoblastic leukemia. Hematology 20, 377–383.

Liu, P., Liu, H.B., Lu, Y., Wen, W., Jia, D., Wang, Y., You, M., 2011. Genome-wide association and fine mapping of genetic loci predisposing to colon carcinogenesis in mice. Molecular cancer research , molcanres–0540.

Liu, P., Tang, H., Wu, J., et al., 2018a. Linc01638 promotes tumorigenesis in her2+ breast cancer. Current Cancer Drug Targets 18, 1–1.

Liu, X., Wang, J., Sun, G., 2015. Identification of key genes and pathways in renal cell carcinoma through expression profiling data. Kidney and Blood Pressure Research 40, 288–297.

Liu, Y., Li, P., Fan, L., Wu, M., 2018b. The nuclear transportation routes of membrane-bound transcription factors. Cell Commun Signal 16, 12.

Longville, B.A., Anderson, D., Welch, M.D., Kees, U.R., Greene, W.K., 2015. Aberrant expression of aldehyde dehydrogenase 1a (aldh 1a) subfamily genes in acute lymphoblastic leukaemia is a common feature of t-lineage tumours. British journal of haematology 168, 246–257.

Luque-Baena, R., Urda, D., Subirats, J., Franco, L., Jerez, J., 2013. Analysis of cancer microarray data using constructive neural networks and genetic algorithms, in: Proceedings of the IWBBIO, international work-conference on bioinformatics and biomedical engineering, pp. 55–63.

Maia, J., Caja, S., Strano Moraes, M., Couto, N., Costa-Silva, B., 2018. Exosome-based cell-cell communication in the tumor microenvironment. Front Cell Dev Biol 6, 18.

Mamoshina, P., Vieira, A., Putin, E., Zhavoronkov, A., 2016. Applications of deep learning in biomedicine. Molecular pharmaceutics 13, 1445–1454.

Martínez-Iglesias, O., Olmeda, D., Alonso-Merino, E., Gómez-Rey, S., González-López, A.M., Luengo, E., Soengas, M.S., Palacios, J., Regadera, J., Aranda, A., 2016. The nuclear corepressor 1 and the thyroid hormone receptor β suppress breast tumor lymphangiogenesis. Oncotarget 7, 78971.

Miao, J., Niu, L., 2016. A survey on feature selection. Procedia Computer Science 91, 919–926.

Min, S., Lee, B., Yoon, S., 2017. Deep learning in bioinformatics. Briefings in bioinformatics 18, 851–869.

Mitchell, M., 1998. An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA, USA.

Nattestad, M., Goodwin, S., Ng, K., Baslan, T., Sedlazeck, F., Rescheneder, P., Garvin, T., Fang, H., Gurtowski, J., Hutton, E., et al., 2018. Complex rearrangements and oncogene amplifications revealed by long-read dna and rna sequencing of a breast cancer cell line. Genome research , gr–231100.

Newman, B., Lose, F., Kedda, M.A., Francois, M., Ferguson, K., Janda, M., Yates, P., Spurdle, A.B., Hayes, S.C., 2012. Possible genetic predisposition to lymphedema after breast cancer. Lymphatic research and biology 10, 2–13.

Ng, A.Y., 2004. Feature selection, l 1 vs. l 2 regularization, and rotational invariance, in: Proceedings of the twenty-first international conference on Machine learning, ACM. p. 78.

Ng, H.Y., Wan, T.S., So, C.C., Chim, C.S., 2014. Epigenetic inactivation of dapk1, p14arf, mir-34a and-34b/c in acute promyelocytic leukaemia. Journal of clinical pathology 67, 626–631.

Owzar, K., Barry, W., Jung, S., 2011. Statistical considerations for analysis of microarray experiments. Clin Transl Sci 4, 466–477.

Papavasileiou, E., Jansen, B., 2016. A comparison between fs-neat and fd-neat and an investigation of different initial topologies for a classification task with irrelevant features, in: Computational Intelligence (SSCI), 2016 IEEE Symposium Series on, IEEE. pp. 1–8.

Papavasileiou, E., Jansen, B., 2017a. The importance of the activation function in neuroevolution with fs-neat and fd-neat, in: Computational Intelligence (SSCI), 2017 IEEE Symposium Series on, IEEE. pp. 1–7.

Papavasileiou, E., Jansen, B., 2017b. An investigation of topological choices in fs-neat and fd-neat on xor-based problems of increased complexity, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, ACM. pp. 1431–1434.

Park, Y., Kellis, M., 2015. Deep learning for regulatory genomics. Nat Biotechnol 33, 825–826.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pirooznia, M., Yang, J.Y., Yang, M.Q., Deng, Y., 2008. A comparative study of different machine learning methods on microarray gene expression data. BMC genomics 9, S13.

Pohlert, T., 2014. The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR). URL: `https://CRAN.R-project.org/package=PMCMR`. r package.

Poliseno, L., Marranci, A., Pandolfi, P., 2015. Pseudogenes in human cancer. Front Med (Lausanne) 2, 68.

Saitoh, M., 2018. Involvement of partial emt in cancer progression. J Biochem , doi: 10.1093/jb/mvy047.

Shangkuan, W.C., Lin, H.C., Chang, Y.T., Jian, C.E., Fan, H.C., Chen, K.H., Liu, Y.F., Hsu, H.M., Chou, H.L., Yao, C.T., et al., 2017. Risk analysis of colorectal cancer incidence by gene expression analysis. PeerJ 5, e3003.

Sher, G.I., 2013. Introduction to Neuroevolutionary Methods. Springer, New York, NY. pp. 105–141.

Shi, Z., Wei, Z., Li, J., Yuan, S., Pan, B., et al., 2018. Identification and verification of candidate genes regulating neural stem cells behavior under hypoxia. Cell Physiol Biochem 47, 212–222.

Sipper, M., Olson, R.S., Moore, J.H., 2017. Evolutionary computation: the next major transition of artificial intelligence?

Soares, G.P., Pereira, A.A.L., Boas, M.S.V., Van Vaisberg, V., Magalhães, M.C.F., Linck, R.D.M., Mano, M.S., 2018. Value of systemic staging in asymptomatic early breast cancer. Revista Brasileira de Ginecologia e Obstetrícia/RBGO Gynecology and Obstetrics .

Sohangir, S., Rahimi, S., Gupta, B., 2013. Optimized feature selection using of augmenting topologies (neat), in: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint, IEEE. pp. 80–85.

Sohangir, S., Rahimi, S., Gupta, B., 2014. Neuroevolutionary feature selection using neat. Journal of Software Engineering and Applications 7, 562.

Srivastava, R.K., Greff, K., Schmidhuber, J., 2015. Highway networks. arXiv preprint arXiv:1505.00387 .

Stanley, K.O., Miikkulainen, R., 2002. Evolving neural networks through augmenting topologies. Evolutionary Computation 10, 99–127.

Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319.

Stuelten, C., Parent, C., Montell, D., 2018. Cell motility in cancer invasion and metastasis: insights from simple model organisms. Nat Rev Cancer 18, 296–312.

Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40, 3358–3378.

Tan, M., Hartley, M., Bister, M., Deklerck, R., 2009. Automated feature selection in neuroevolution. Evolutionary Intelligence 1, 271–292.

Tang, J., Li, Y., Sang, Y., Yu, B., Lv, D., et al., 2018a. Lncrna pvt1 regulates triple-negative breast cancer through klf5/beta-catenin signaling. Oncogene , doi:10.1038/s41388–018–0310–4.

Tang, Y., Naito, S., Abe-Kanoh, N., Ogawa, S., Yamaguchi, S., Zhu, B., Murata, Y., Nakamura, Y., 2018b. Benzyl isothiocyanate attenuates the hydrogen peroxide-induced interleukin-13 expression through glutathione s-transferase p induction in t lymphocytic leukemia cells. Journal of biochemical and molecular toxicology , e22054.

Tao, Y.F., Xu, L.X., Lu, J., Hu, S.Y., Fang, F., Cao, L., Xiao, P.F., Du, X.J., Sun, L.C., Li, Z.H., et al., 2015. Early b-cell factor 3 (ebf3) is a novel tumor suppressor gene with promoter hypermethylation in pediatric acute myeloid leukemia. Journal of Experimental & Clinical Cancer Research 34, 4.

Tao, Z., Shi, A., Li, R., Wang, Y., Wang, X., et al., 2017. Microarray bioinformatics in cancer- a review. J BUON 22, 838–843.

Thakkar, A., Raj, H., Ravishankar, Muthuvelan, B., Balakrishnan, A., Padigaru, M., 2015. High expression of three-gene signature improves prediction of relapse-free survival in estrogen receptor-positive and node-positive breast tumors. Biomarker insights 10, BMI–S30559.

Thakkar, A.D., Raj, H., Chakrabarti, D., Ravishankar, Saravanan, N., Muthuvelan, B., Balakrishnan, A., Padigaru, M., 2010. Identification of gene expression signature in estrogen receptor positive breast carcinoma. Biomarkers in cancer 2, BIC–S3793.

Thutkawkorapin, J., Picelli, S., Kontham, V., Liu, T., Nilsson, D., Lindblom, A., 2016. Exome sequencing in one family with gastric-and rectal cancer. BMC genetics 17, 41.

Tomoshige, K., Matsumoto, K., Tsuchiya, T., Oikawa, M., Miyazaki, T., Yamasaki, N., Mishima, H., Kinoshita, A., Kubo, T., Fukushima, K., et al., 2015. Germline mutations causing familial lung cancer. Journal of human genetics

60, 597.

Tutar, Y., Özgür, A., Tutar, E., Tutar, L., Pulliero, A., et al., 2016. Regulation of oncogenic genes by micrornas and pseudogenes in human lung cancer. Biomed Pharmacother 83, 1182–1190.

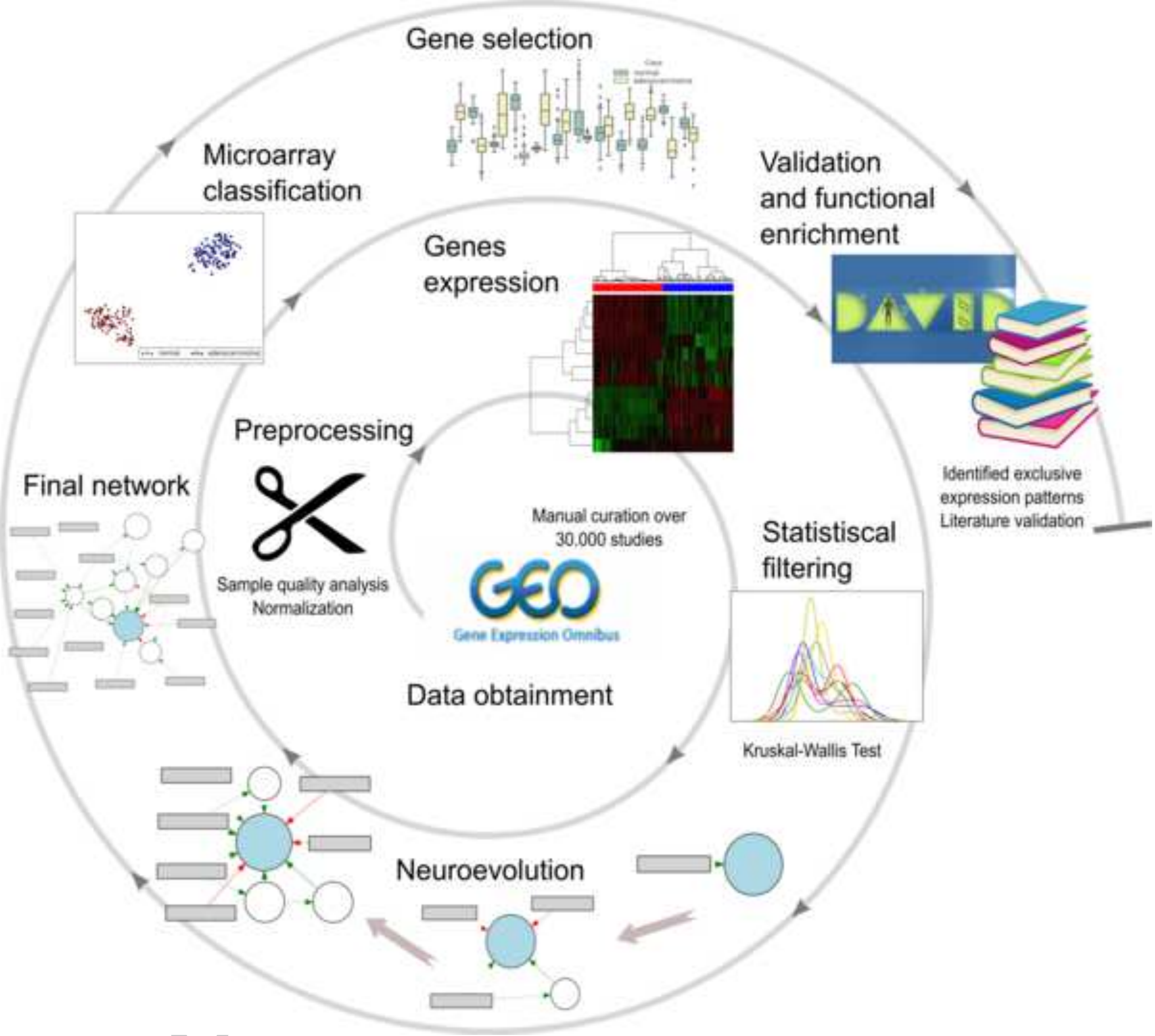
Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction., in: IWANN, Springer. pp. 758–770.

Walsh, C., Hu, P., Batt, J., Santos, C., 2015. Microarray meta-analysis and cross-platform normalization: Integrative genomics for robust biomarker discovery. Microarrays (Basel) 4, 389–406.

Wang, B., Xi, Y., 2013. Challenges for microrna microarray data analysis. Microarrays (Basel) 2, 10.3390/microarrays2020034.

Wedge, D., Gundem, G., Mitchell, T., Woodcock, D., Martincorena, I., et al., 2018. Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nat Genet 50, 682–692.

Whiteson, S., Stone, P., Stanley, K.O., Miikkulainen, R., Kohl, N., 2005. Automatic feature selection in neuroevolution, in: Proceedings of the 7th annual conference on Genetic and evolutionary computation, ACM. pp. 1225–1232.


Gene selection
Microarray classification
Validation and functional enrichment
Genes expression
DAVID
Identified exclusive expression patterns
Literature validation
Preprocessing
Final network
Sample quality analysis
Normalization
Manual curation over 30,000 studies
GEO
Gene Expression Omnibus
Statistiscal filtering
Data obtainment
Kruskal-Wallis Test
Neuroevolution

Highlights

- We propose a new Neuroevolution-based algorithm for analyzing microarray data;
- Design of new structural operators for FS-NEAT;
- High classification results were obtained when comparing to other approaches;
- A list of potential biomarkers for different types of cancer is discussed;
- Our method selected potential relevant genes to understand cancer biology;