Feature Selection

Generative AI Academy

Part 4: Neural Networks — FS-NEAT and N3O

Bruno Iochins Grisci[1]

[1] Institute of Informatics, UFRGS, Porto Alegre, Brazil

bigrisci@inf.ufrgs.br

22 de maio de 2025

Summary

Summary

NEAT

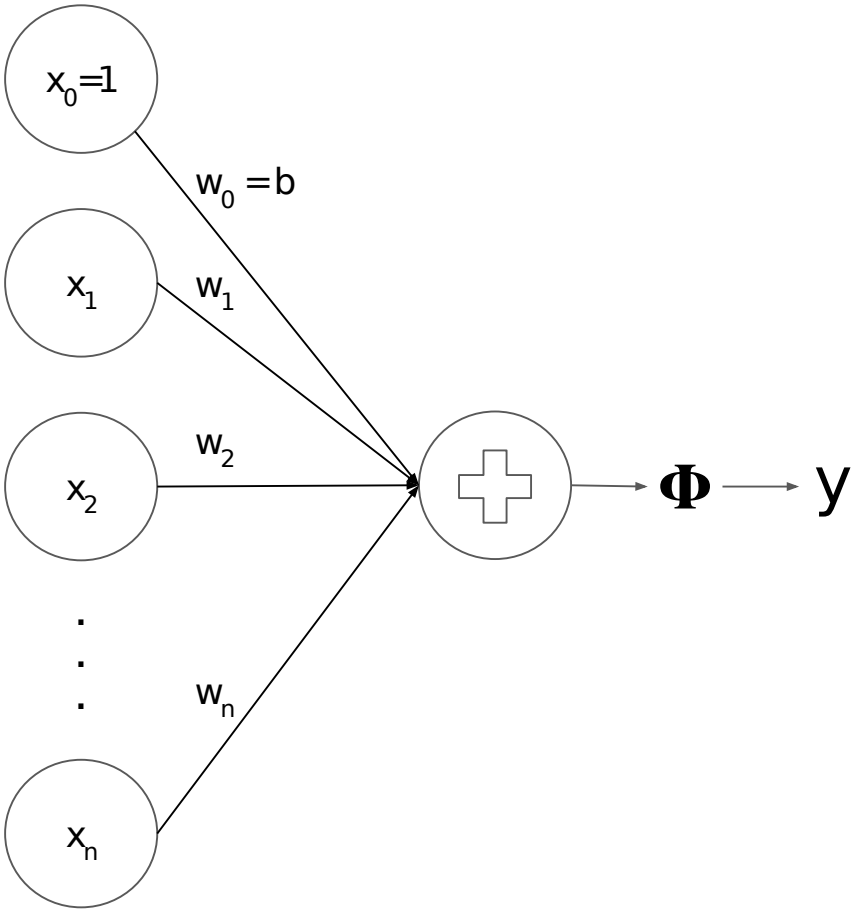
Microarray and Neuroevolution

- Two in one;
- Automaticity.

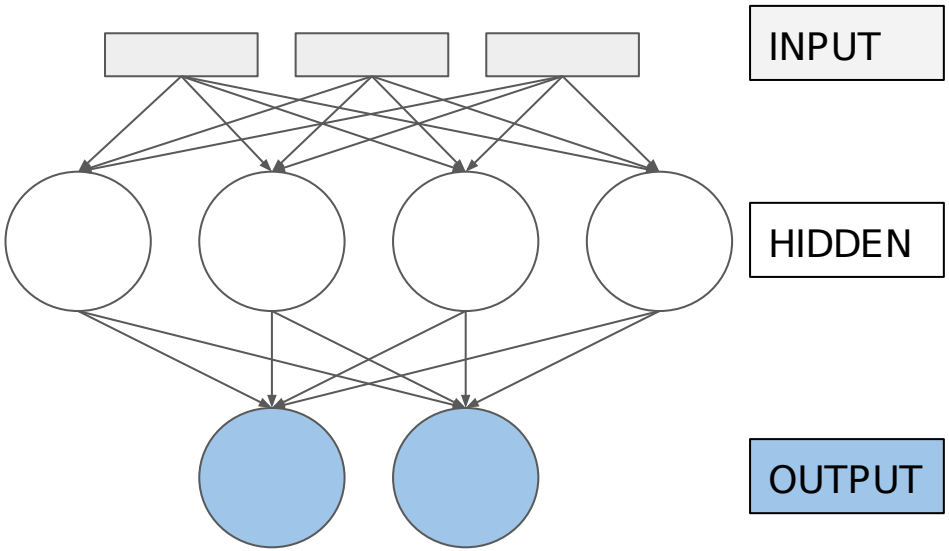
NEAT: NeuroEvolution of Augmenting Topologies

- Idea: to use GA to evolve neural networks.
- FS-NEAT: NEAT with feature selection [Whiteson et al., 2005].

Artificial neural network

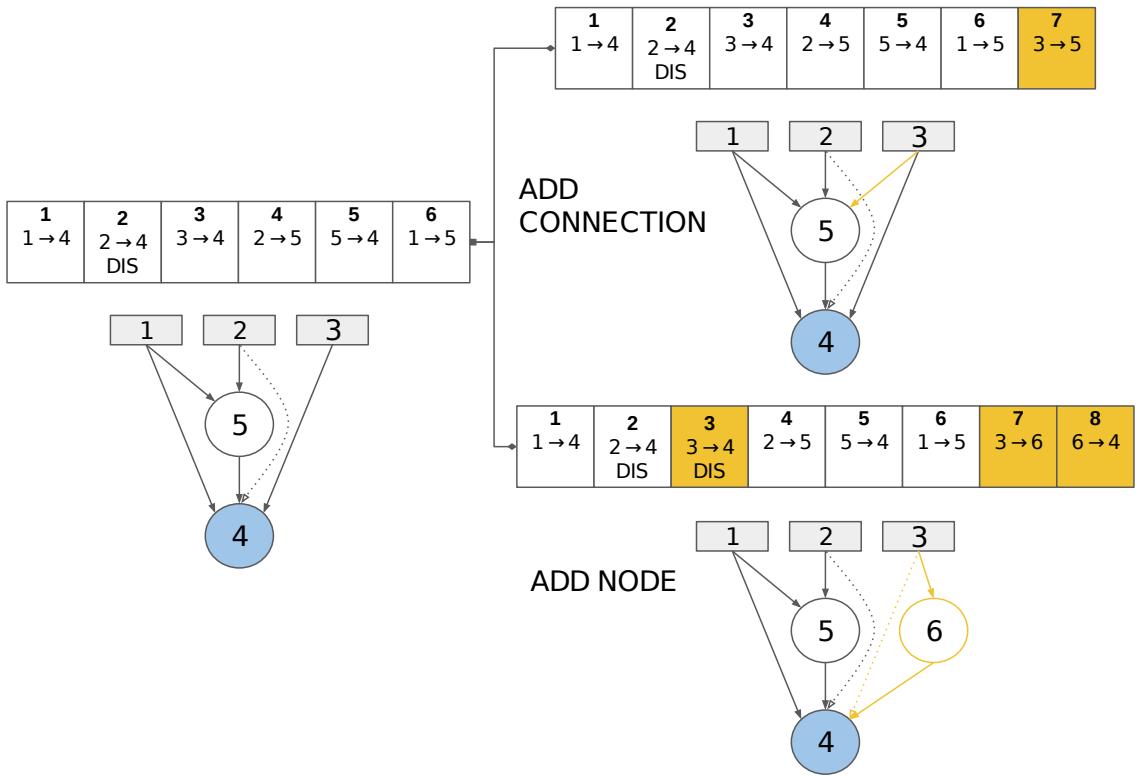


(a) Neuron



(b) MLP

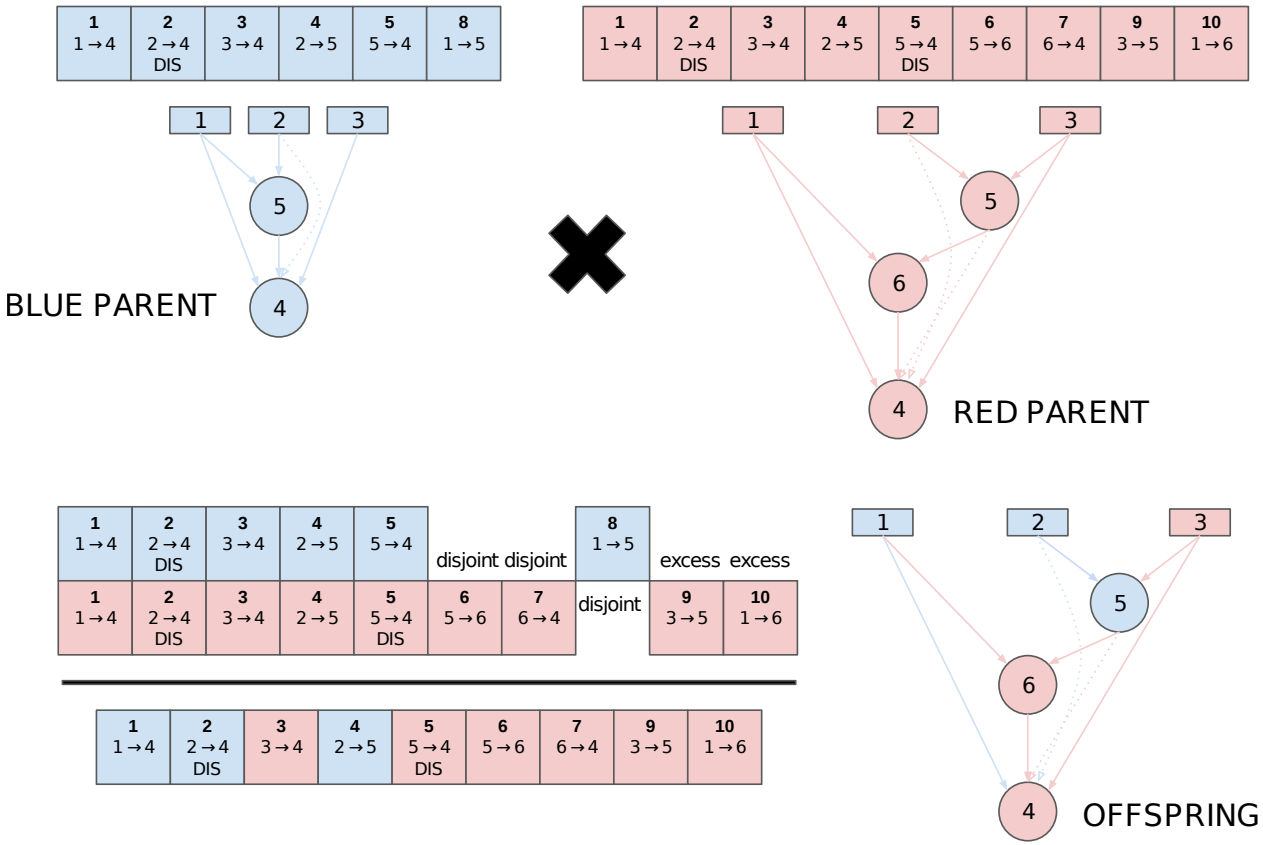
NEAT representation



Genome representation and structural mutations in NEAT

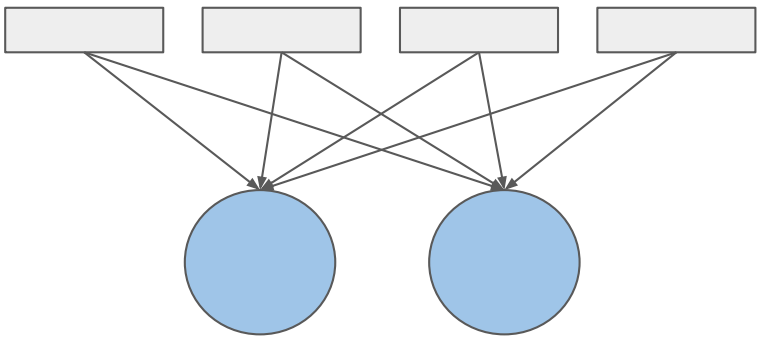
Gray: input / White: hidden / Blue: output / Yellow: new

NEAT crossover

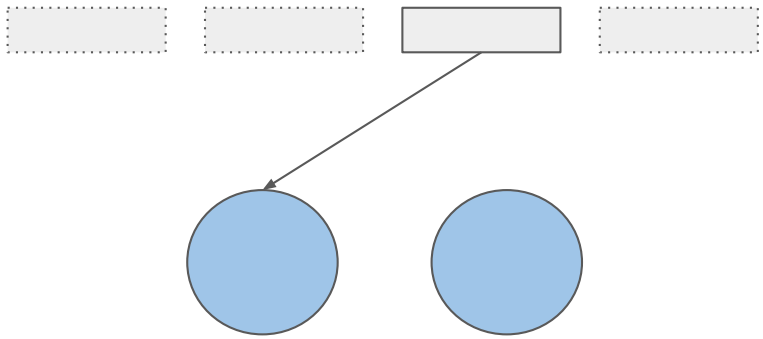


Example of NEAT crossover between two individuals, in which the red parent has better fitness than the blue parent

FS-NEAT initialization



(a) NEAT

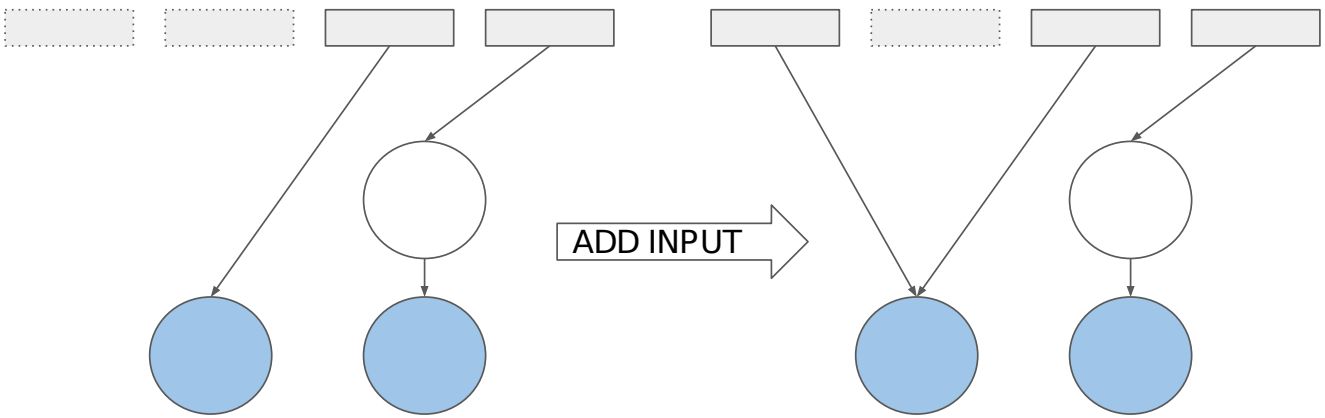


(b) FS-NEAT

Initial topology comparison between NEAT and FS-NEAT

Gray: input / White: hidden / Blue: output

FS-NEAT mutation



Example of the extra FS-NEAT structural mutation, that adds a new input in a network

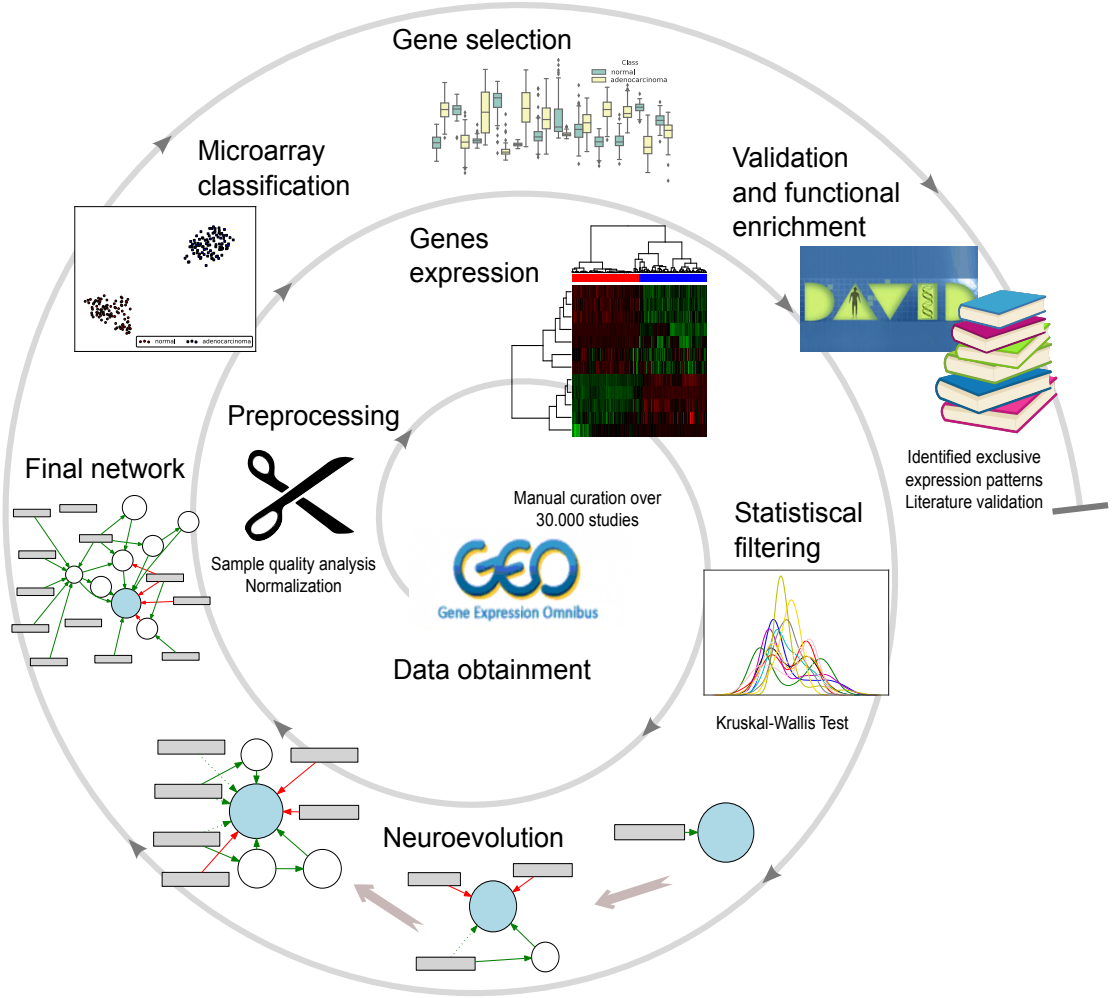
Gray: input / White: hidden / Blue: output

Summary

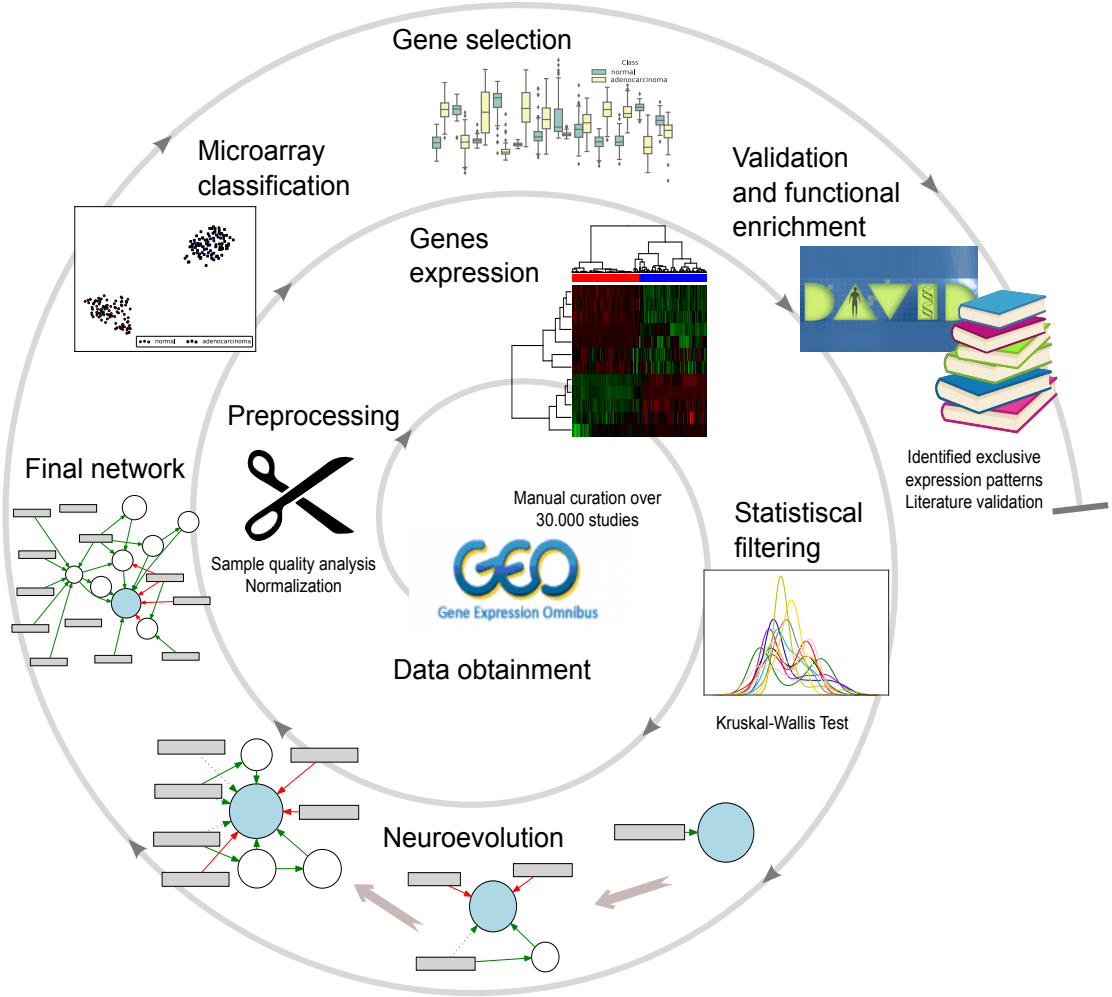
The proposed method

- Gene selection and classification;
- Preprocessing, filtering, and Neuroevolution;
- Autonomous;
- No selection threshold;
- Hybrid.

Summary of the proposed method



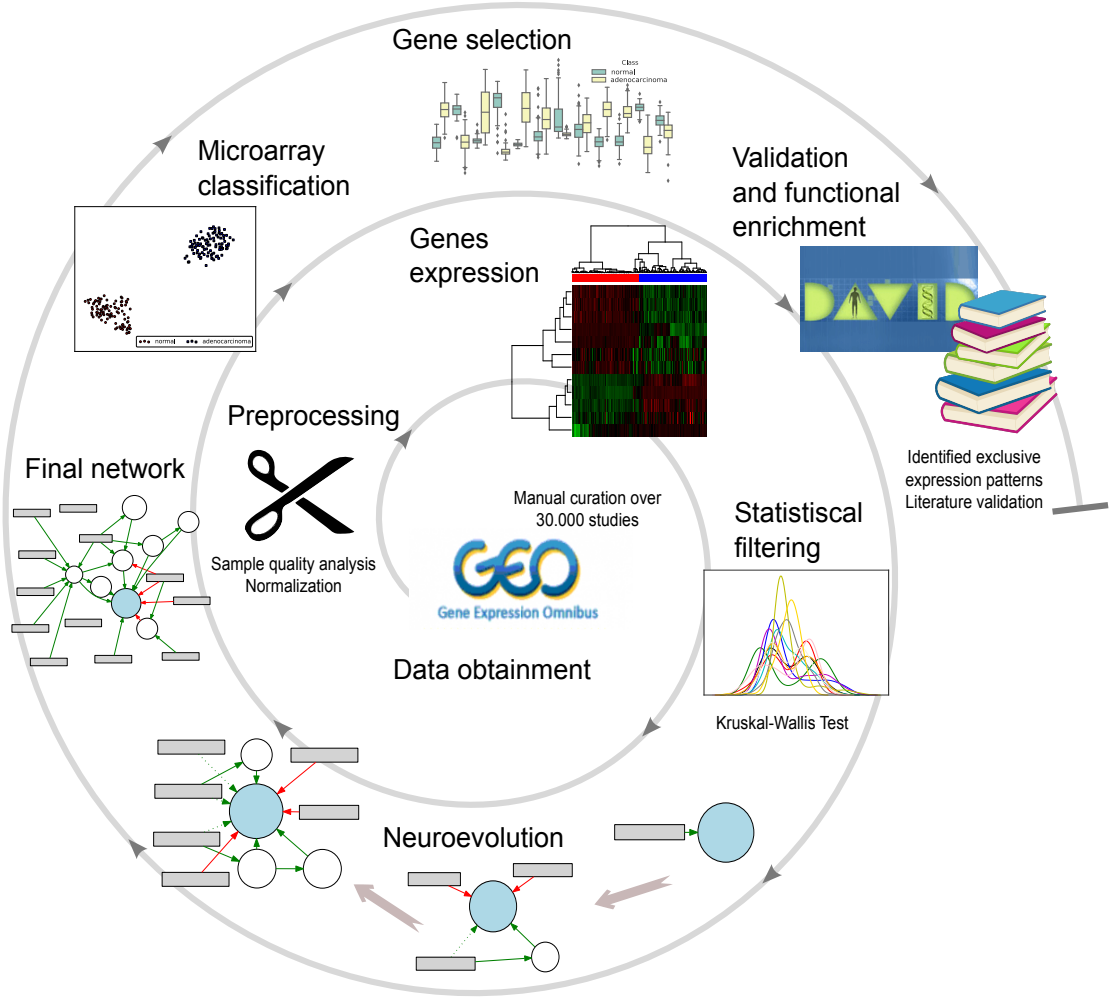
Proposed method: Filtering



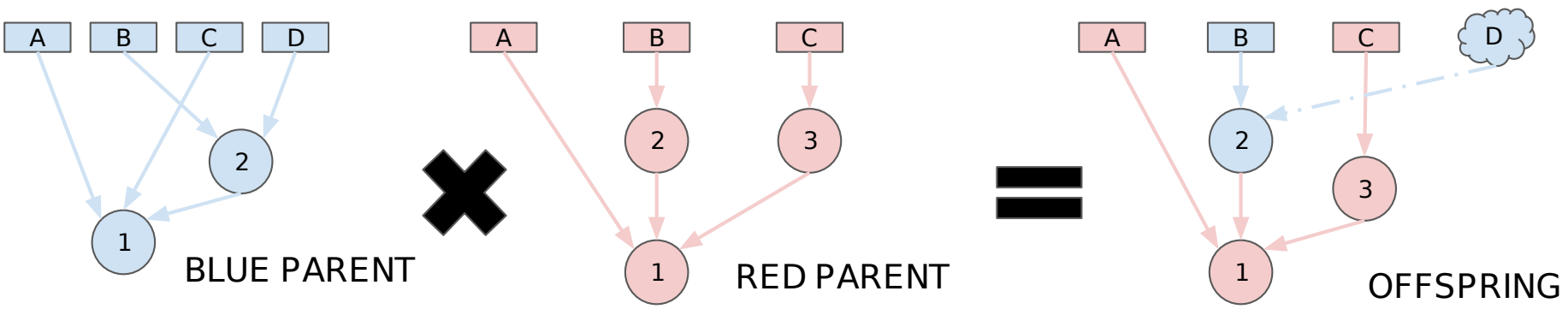
Filtering

- Data does not follow normal distribution;
- Kruskal-Wallis H Test;
- Allow different sample sizes;
- Test whether samples originate from the same distribution;
- Remove all genes whose expression distribution does not differ among classes.

Proposed method: Neuroevolution

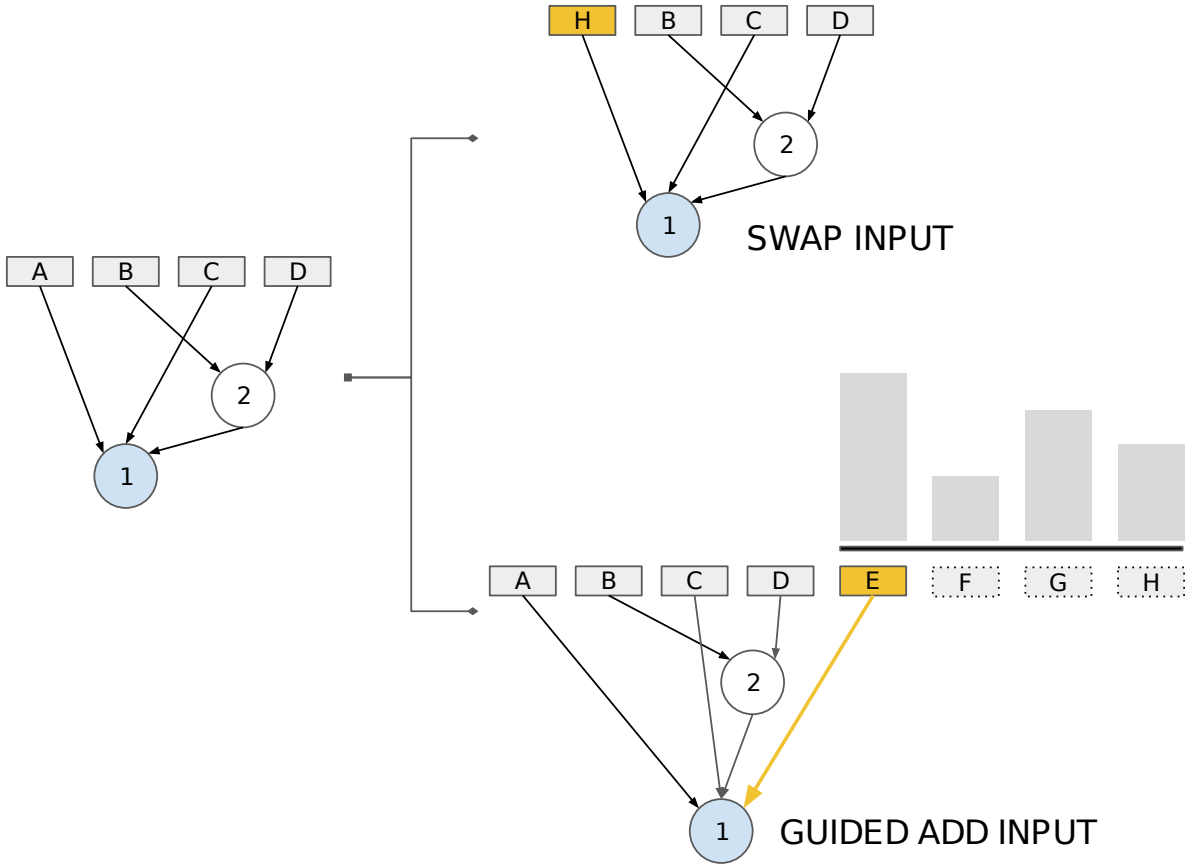


The proposed crossover operator



Additive crossover operator

The proposed mutation operators



The two new possible structural mutations for the proposed method.

Gray: input / White: hidden / Blue: output / Yellow: new

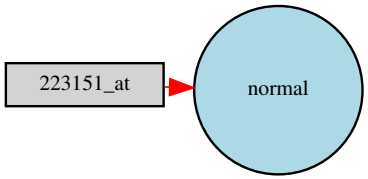
Fitness and Artificial Neuron

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

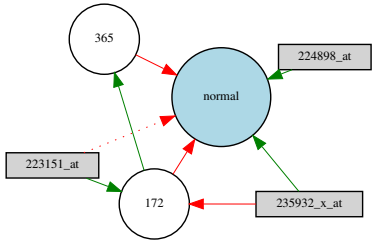
$$+ \frac{\lambda}{2n} \frac{1}{c} \sum_{k=1}^{c} w_k^2 \tag{1b}$$

$$a_h = \Phi\left(\frac{1}{m_h} \sum_{j=1}^{m_h} w_{hj} x_{hj} + b_h\right) \tag{2}$$

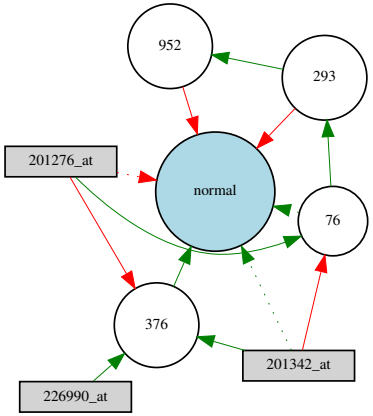
Evolution



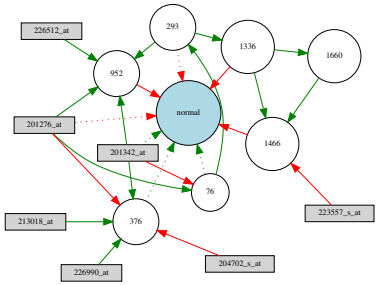
(a) 1º generation



(b) 19º generation



(c) 66º generation

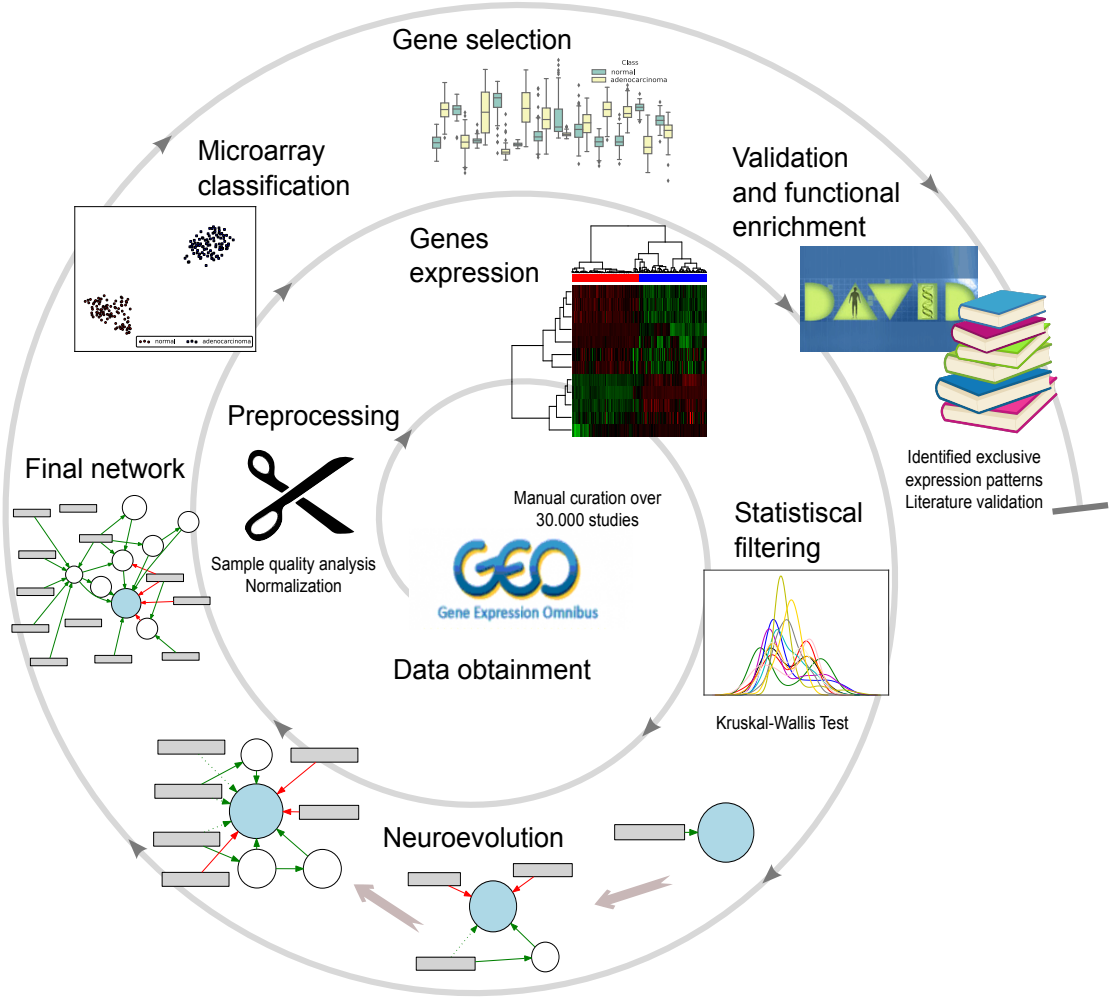


(d) 98º generation

Gray: input / White: hidden / Blue: output

Summary

Proposed method: Classification and selection



N3O x FS-NEAT

Stratified 3-fold cross-validation x 31 runs

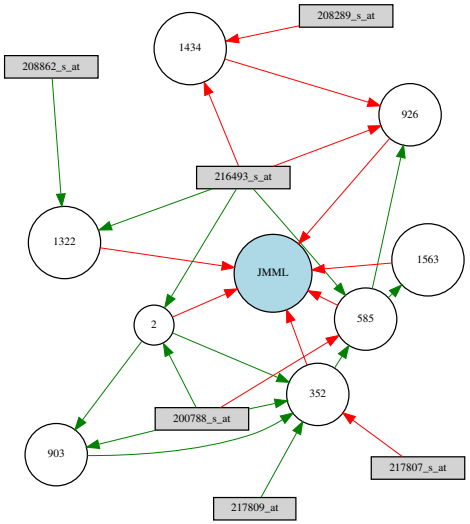
Accuracy and FS comparison of N3O with FS-NEAT.

		Accuracy		FS	
Datasets	**Class**	**N3O**	**FS-NEAT**	**N3O**	**FS-NEAT**
GSE10797	Cancer Epithelial	**0.736** ± .058	0.725 ± .043	**13.65** ± 2.36	32.33 ± 10.77
	Cancer Stroma	**0.744** ± .035	0.734 ± .044	**13.85** ± 2.76	37.12 ± 12.53
	Normal	**0.930** ± .024	0.921 ± .024	**12.92** ± 4.19	20.09 ± 9.13
GSE8671		**0.984** ± .018	0.980 ± .020	**15.16** ± 3.99	17.53 ± 7.98
GSE32323		**0.939** ± .040	0.934 ± .043	**15.74** ± 4.02	20.29 ± 8.97
GSE41328		**0.968** ± .045	0.955 ± .071	18.67 ± 6.35	**18.60** ± 9.24
GSE14317		**0.964** ± .040	0.960 ± .044	**14.80** ± 4.76	20.77 ± 9.44
GSE71935		**0.902** ± .046	0.860 ± .047	**14.60** ± 3.42	26.13 ± 11.54
golub1999molecular		0.900 ± .032	**0.901** ± .038	**12.51** ± 2.43	28.58 ± 11.97
Average		**0.896** ± .093	0.886 ± .095	**14.65** ± 1.83	24.60 ± 6.84

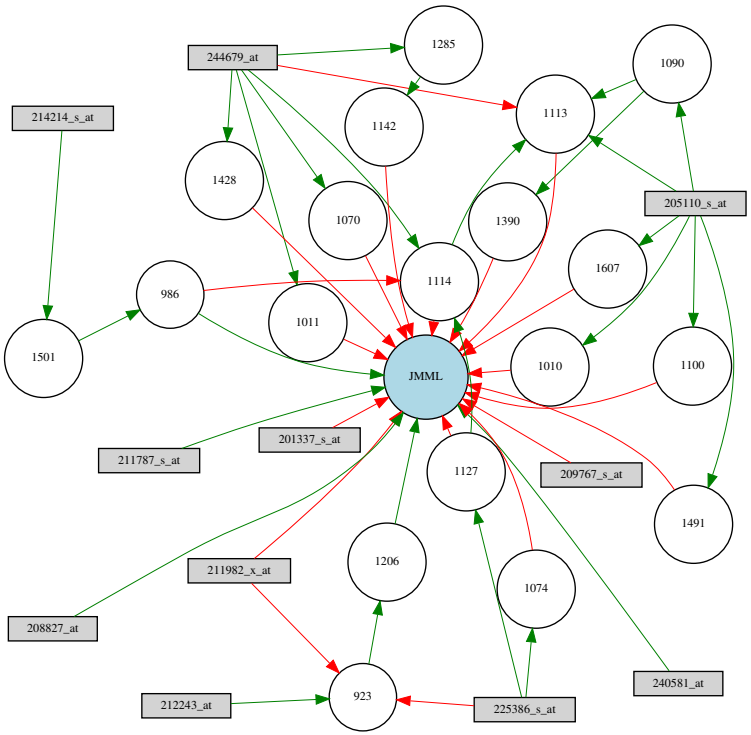
In bold are the best average accuracy and smallest average FS of each dataset. Best results with statistical significance ($p < 0.01$) are marked in blue.

N3O x FS-NEAT

Final neural networks (GSE71935 Leukemia)



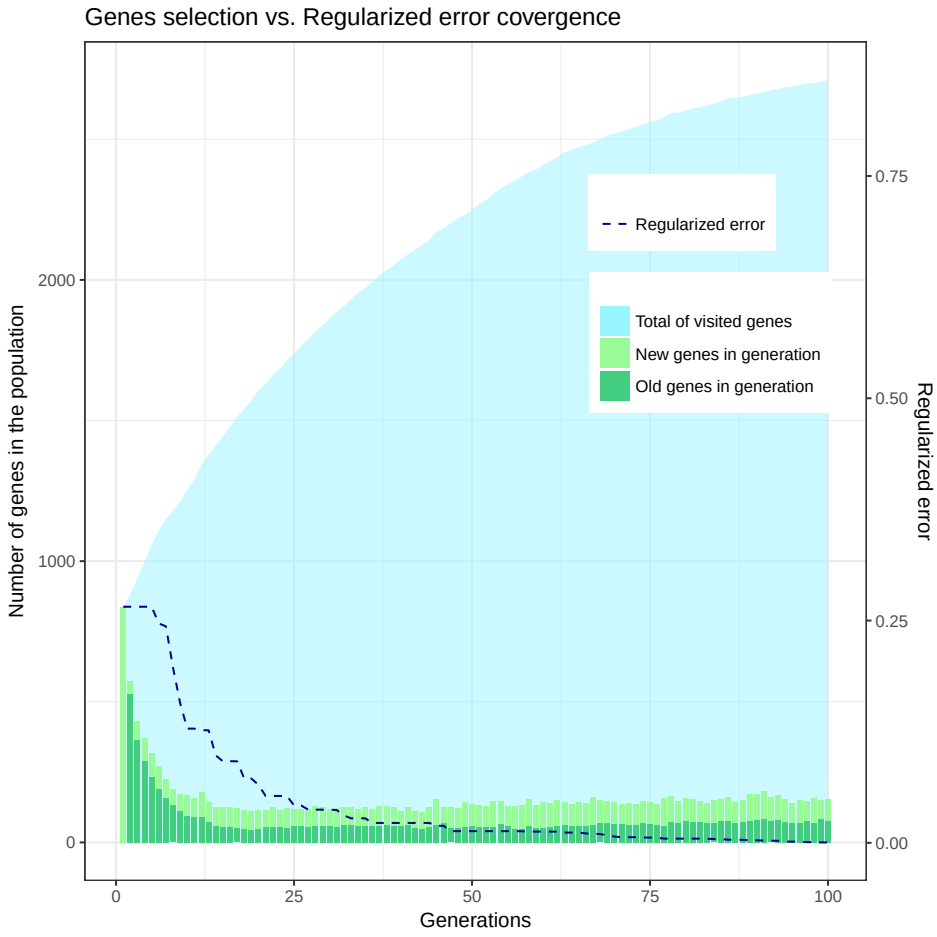
(a) N3O



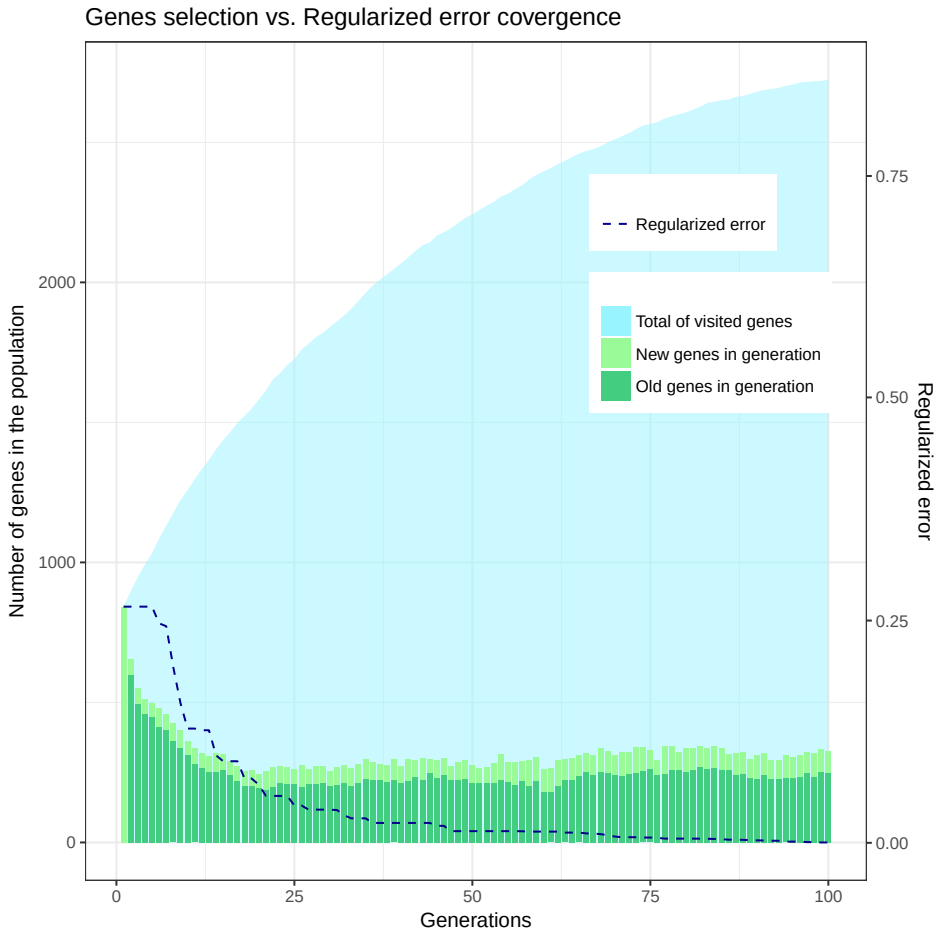
(b) FS-NEAT

N3O x FS-NEAT

Genes selection and error convergence for N3O and FS-NEAT



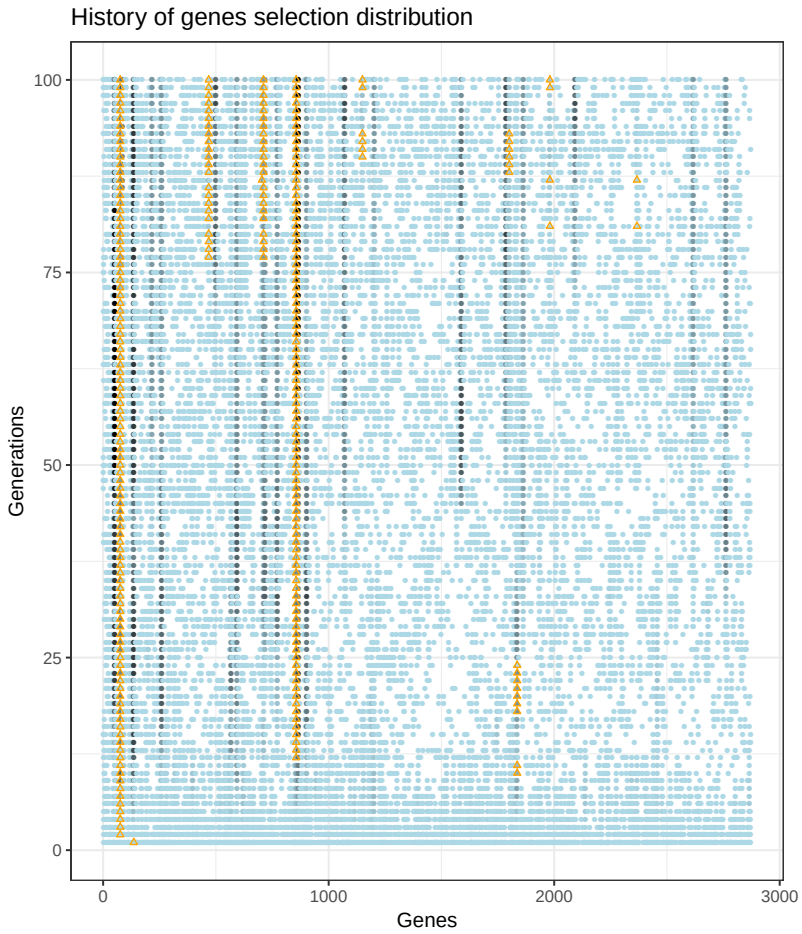
(a) N3O



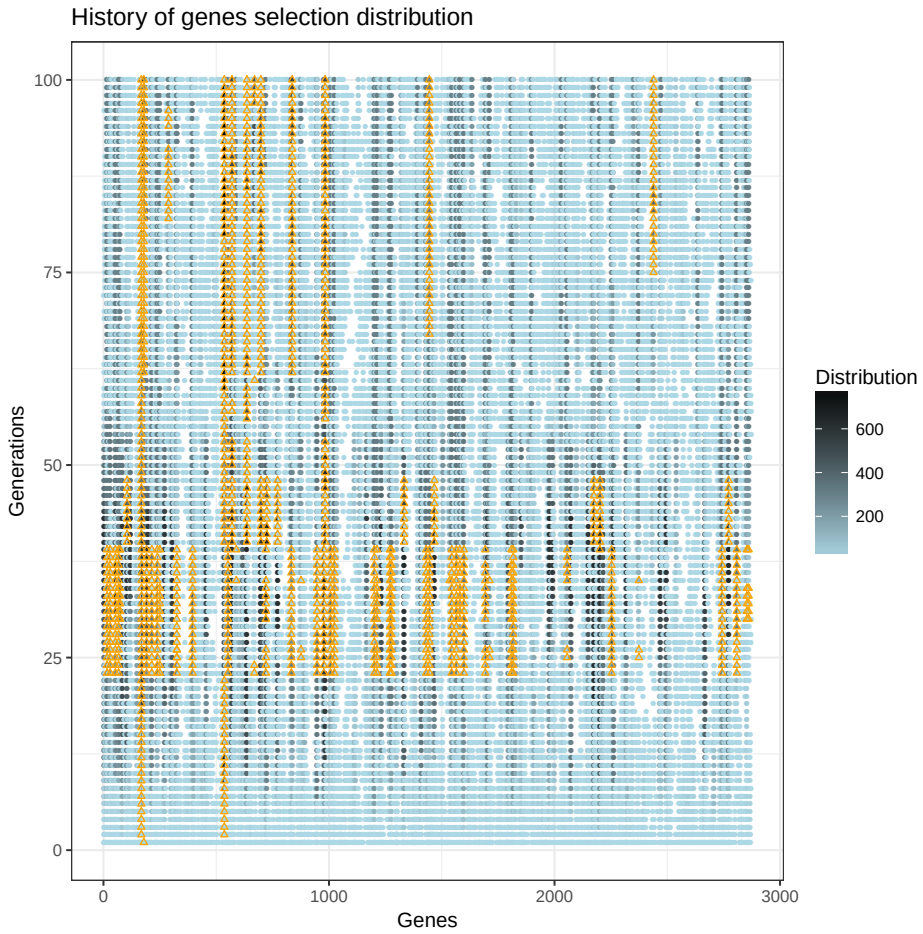
(b) FS-NEAT

N3O x FS-NEAT

Selection history of candidate genes for N3O and FS-NEAT



(a) N3O



(b) FS-NEAT

N3O x another Neuroevolution method

Method	**Dataset**	**Accuracy**	**FS**
N3O	[Golub et al., 1999]	$0.917 \pm .095$	6.27 ± 2.38
ABC&DE	[Golub et al., 1999]	$0.912 \pm .067$	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.

N3O x SVM

Datasets	Class	Baseline	N3O	SVM	KW&SVM	N3O&SVM
GSE42568		0.87	0.978 ± .011	0.985 ± .007	0.985 ± .006	**0.990** ± .006
GSE45827	Basal	0.73	0.934 ± .016	**0.972** ± .003	0.971 ± .004	0.968 ± .012
	HER	0.80	0.946 ± .019	0.962 ± .010	0.950 ± .011	**0.973** ± .026
	Cell Line	0.91	0.994 ± .006	**1.000** ± .000	**1.000** ± .000	0.999 ± .003
	Luminal A	0.81	0.934 ± .019	0.968 ± .014	**0.979** ± .007	0.965 ± .017
	Luminal B	0.80	0.890 ± .026	**0.931** ± .013	0.928 ± .016	0.923 ± .024
	Normal	0.95	0.988 ± .009	**0.995** ± .003	0.993 ± .000	0.994 ± .005
GSE10797	Cancer Epithelial	0.57	0.736 ± .058	**0.857** ± .028	**0.857** ± .028	0.850 ± .053
	Cancer Stroma	0.57	0.744 ± .035	0.761 ± .036	0.761 ± .036	**0.825** ± .062
	Normal	0.85	0.930 ± .024	0.924 ± .019	0.924 ± .019	**0.965** ± .018
GSE44076		0.50	0.982 ± .009	0.983 ± .003	0.984 ± .003	**0.987** ± .008
GSE44861		0.50	0.823 ± .031	**0.829** ± .045	**0.829** ± .045	**0.829** ± .059
GSE8671		0.51	**0.984** ± .018	0.698 ± .065	0.698 ± .065	0.667 ± .000
GSE21510		0.58	0.956 ± .032	**0.986** ± .021	**0.986** ± .021	**0.986** ± .039
GSE32323		0.51	**0.939** ± .040	0.692 ± .066	0.692 ± .066	0.686 ± .050
GSE41328		0.55	**0.968** ± .045	0.695 ± .061	0.697 ± .040	0.722 ± .000
GSE9476	AML	0.59	0.901 ± .035	0.947 ± .016	0.920 ± .019	**0.954** ± .039
	Bone Marrow	0.84	0.989 ± .017	0.984 ± .000	**0.998** ± .005	0.997 ± .007
	Bone Marrow CD34	0.87	0.963 ± .023	**0.997** ± .007	0.980 ± .019	0.984 ± .018
	PB	0.84	0.994 ± .009	0.985 ± .013	**1.000** ± .000	0.999 ± .004
	PBSC CD34	0.84	0.976 ± .022	0.984 ± .010	**0.997** ± .006	0.995 ± .012
GSE14317		0.72	0.964 ± .040	0.957 ± .044	0.991 ± .025	**0.996** ± .012
GSE63270		0.59	0.969 ± .022	**0.999** ± .003	0.998 ± .004	0.991 ± .011
GSE71935		0.80	0.902 ± .046	0.896 ± .034	0.923 ± .034	**0.966** ± .030
[Golub et al., 1999]		0.65	0.900 ± .032	0.961 ± .022	**0.978** ± .012	0.943 ± .028
Average			**0.931** ± .070	0.918 ± .102	0.921 ± .103	0.926 ± .102

Generalization

- Gene selection performed with a classifier is only specific to that given algorithm [Ang et al., 2016].
- SVMs are usually insensitive to a large number of irrelevant genes, and FS often biases down their accuracy [Statnikov et al., 2008].
- When the genes selected by the N3O were applied to SVM, its performance was not hurt, and for most of the datasets, it actually had a slight improvement.

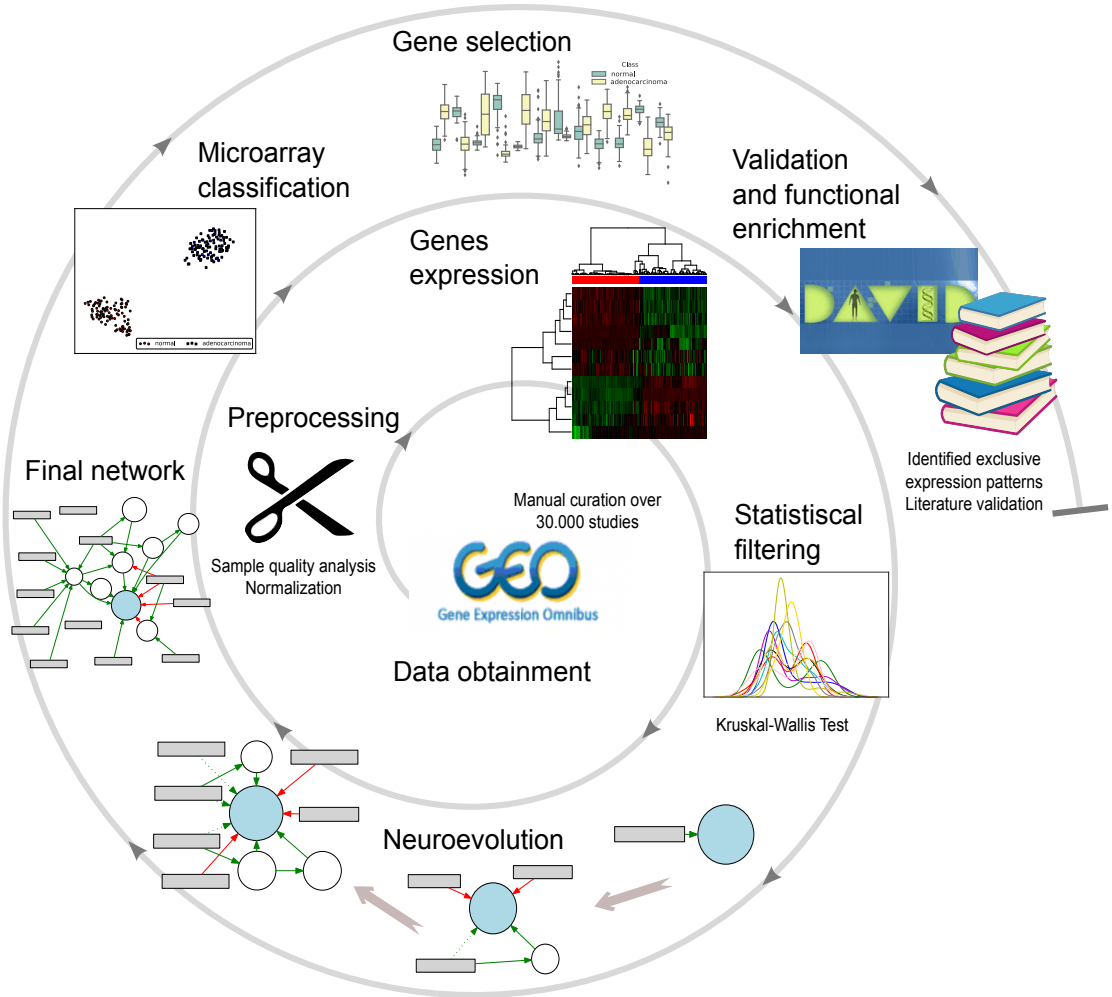
Most selected genes

Most selected genes over all runs for each dataset:

- 44%: Related to the specific type of cancer;
- 20%: Related to another type of cancer;
- 20%: Not related to any cancer;
- 16%: Not yet described.

ERBB2 (HER2) was the most selected gene in its dataset among all experiments, appearing at 90.6% of the networks. It is described as the most relevant gene in breast cancer HER2 Status [Borges et al., 2018].

Proposed method: Functional enrichment



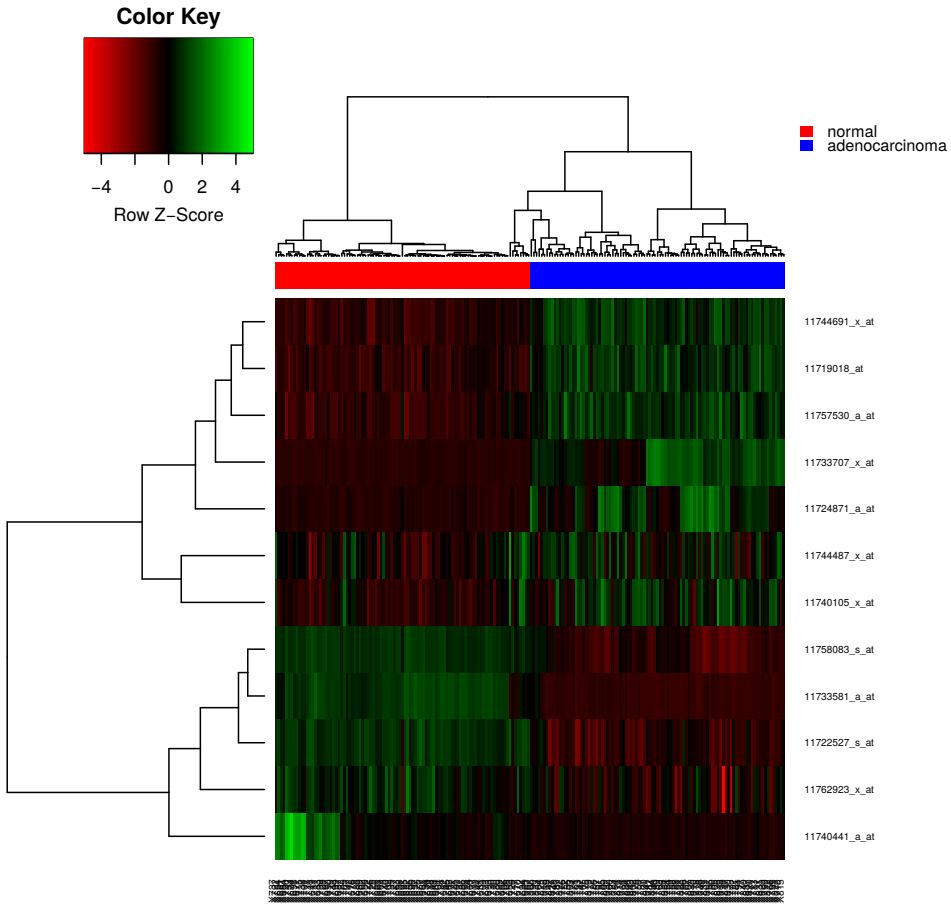
Gene selection experiment

177 genes obtained from running the method for the chosen cancer types considering all available samples.

- 82 associated to their given cancer type;
- 45 observed to be altered in other cancer types;
- 50 don't possess a clear described function, or were just not related to any tumoral condition.

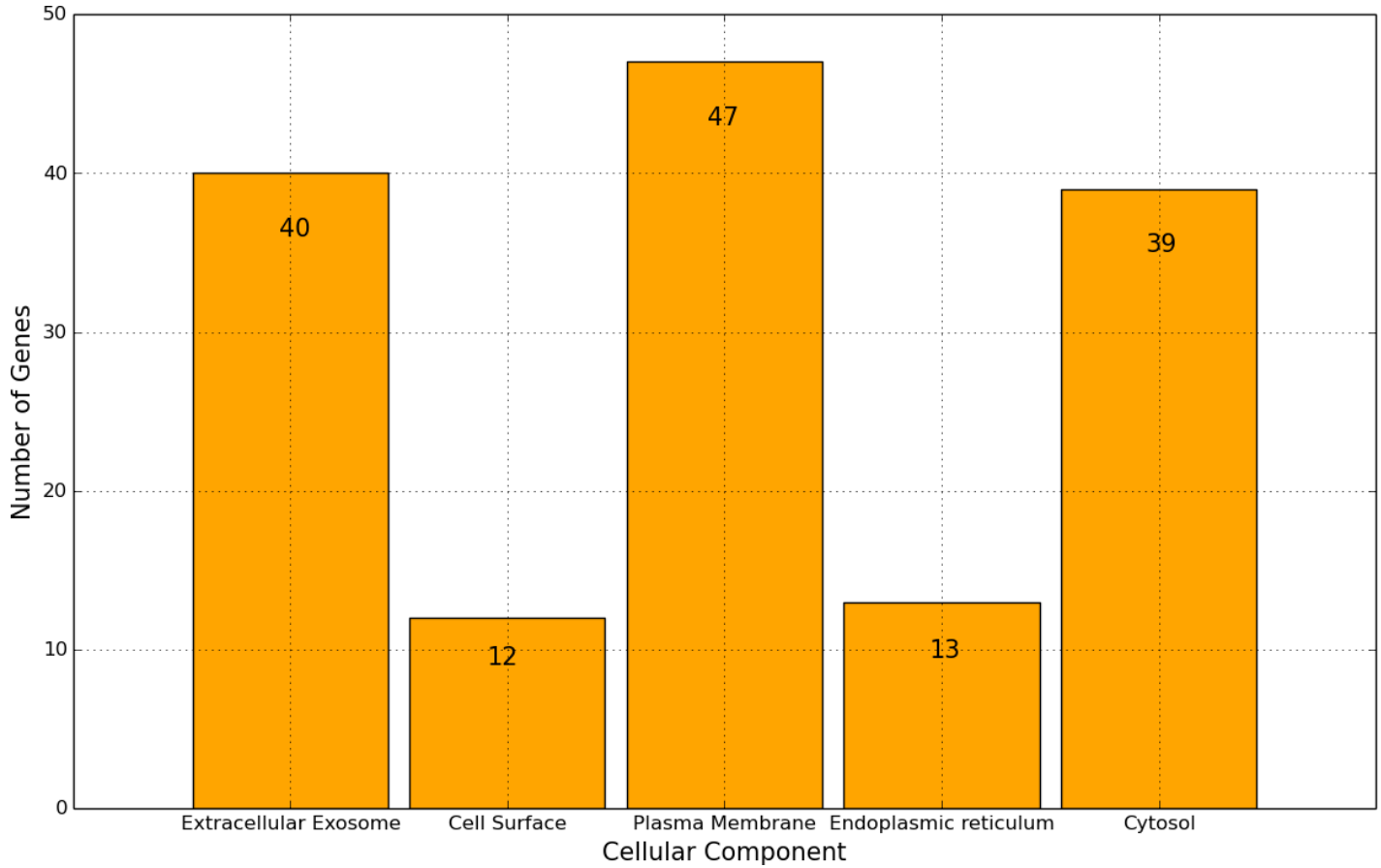
Gene selection experiment

Detailment of gene selection for GSE44076 - CRC Adenocarcinoma.



Cellular components

Major bioprocesses were acessed using DAVID.



The number of genes related to the major cellular components.

Major Gene Ontology

Major GO derived from all selected genes.

Bioprocesses	**Corrected 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}

Biological role

- Components that act in the plasma membrane and extracellular exossomes: fundamental in cancer biology;
- Extracellular Matrix possesses proteins related to cell adhesion and cytoskeleton organization that are fundamental for tumor invasion and colony formation;
- Exossomes are critical for cell-cell signaling, influencing pathological conditions;
- Plasma membrane is part of a dynamic system of external and internal signals, intimately associated to cancer molecular mechanisms.

Summary

Conclusion

A pipeline for microarray classification and gene selection by employing Neuroevolution as a method capable of efficiently performing both tasks was presented.

- Improve upon regular FS-NEAT;
- Autonomous;
- Results matching the ones other methods;
- Generalizable;
- Selected genes validated as envolved in the cancer process.

References


Ang, J. C., Mirzal, A., Haron, H., et al. (2016).
Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection.
IEEE/ACM transactions on computational biology and bioinformatics, 13(5):971–989.


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.

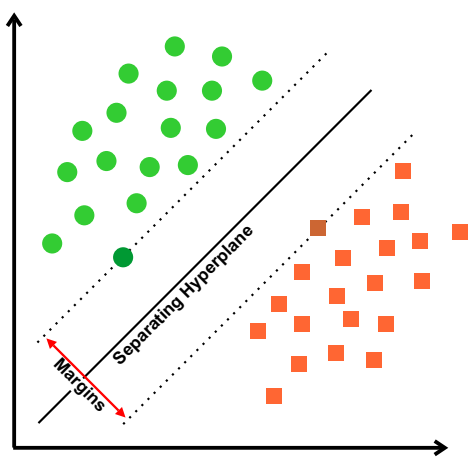

Garro, B. A., Rodríguez, K., and Vazquez, R. A. (2017).
Designing artificial neural networks using differential evolution for classifying dna microarrays.
In *Evolutionary Computation (CEC), 2017 IEEE Congress on*, pages 2767–2774. IEEE.


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(5439):531–537.

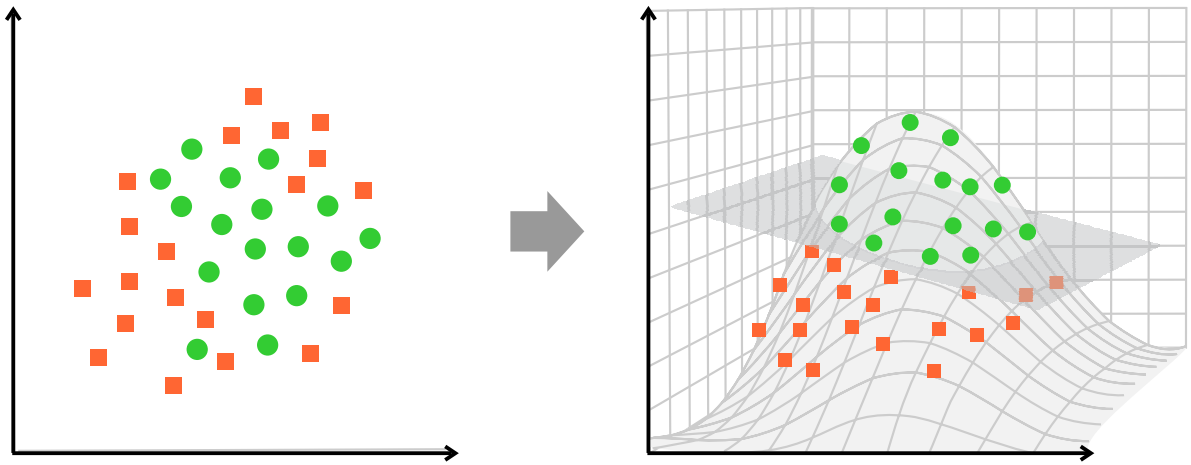

Statnikov, A., Wang, L., and Aliferis, C. F. (2008).
A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.
BMC bioinformatics, 9(1):319.


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

SVM



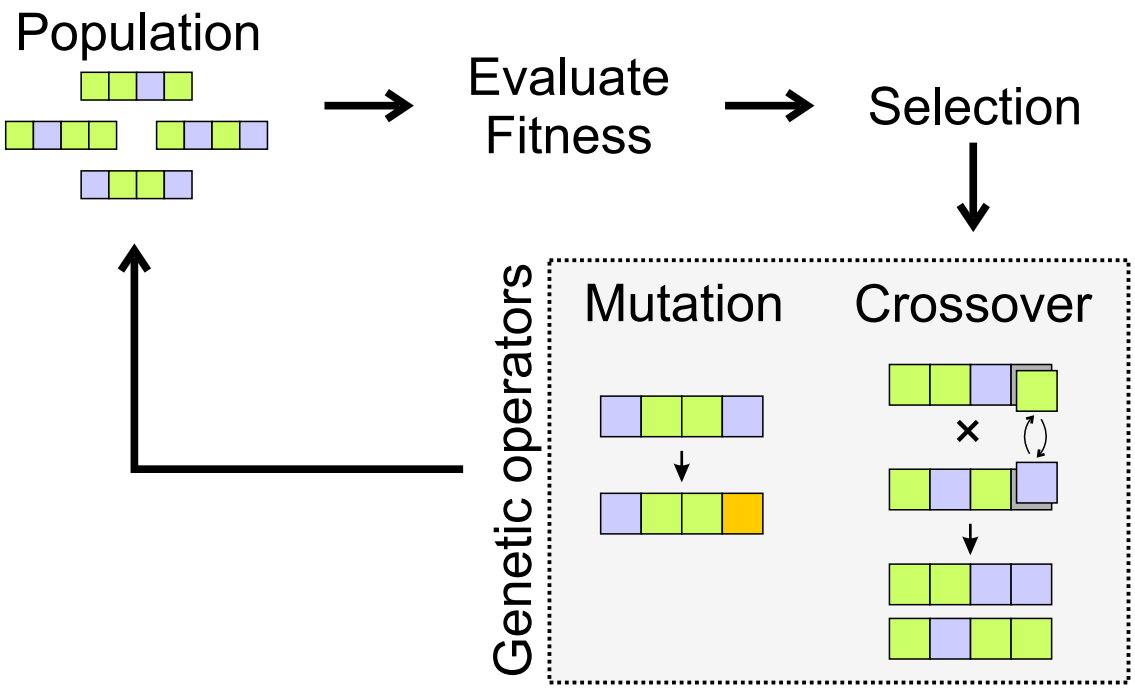
(a) Support Vector Machines



(b) Kernel transformation to higher dimension

(a) Example of an SVM classifying data (represented by dots and squares) in 2D. In this case, the separating hyperplane is the line that best splits the data into two classes. (b) In this case, the data is not linearly separable, so a kernel transformation is applied, mapping it to a higher dimension, where a separating hyperplane exists.

Genetic algorithms



Schematic of a simple GA pipeline. A population of random individuals is generated, each of them representing a candidate solution. These individuals are evaluated by some domain-specific metric and, based on that, selected. The selected individuals can be subjugated to crossover or mutation operators, that create new individuals.

Kruskal-Wallis H Test

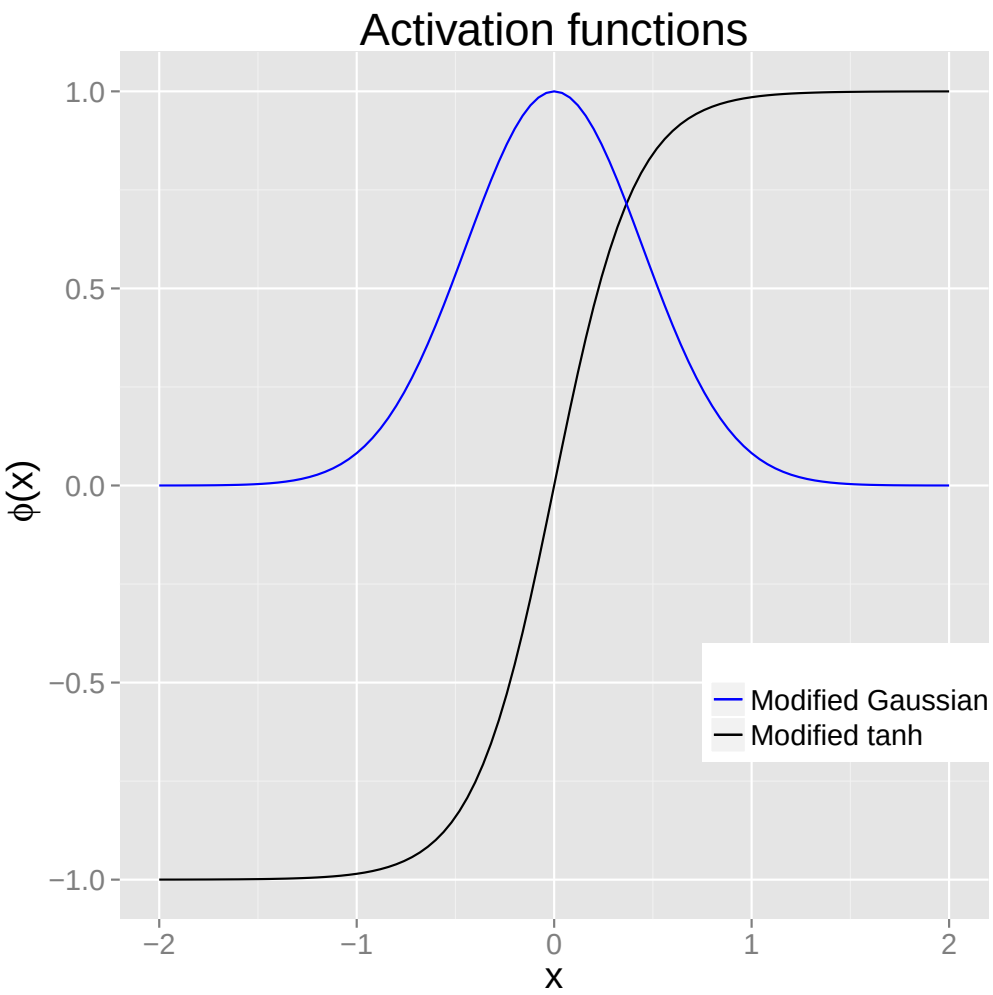
$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \tag{3}$$

$$\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i} \tag{4}$$

$$\bar{r} = \frac{1}{2}(N+1) \tag{5}$$

In which N is the total number of samples in all groups, g is the number of groups, n_i is the number of samples in the group i, r_{ij} is the rank of sample j from group i considering the rank among all samples, $\bar{r}_i$ is the average rank of samples in group i, and $\bar{r}$ is the average of all r_{ij}. The p-value is approximated using chi-squared by $\mathbf{Pr}(\chi^2_{g-1} \geq H)$.

Activation functions



List of used hyperparameters

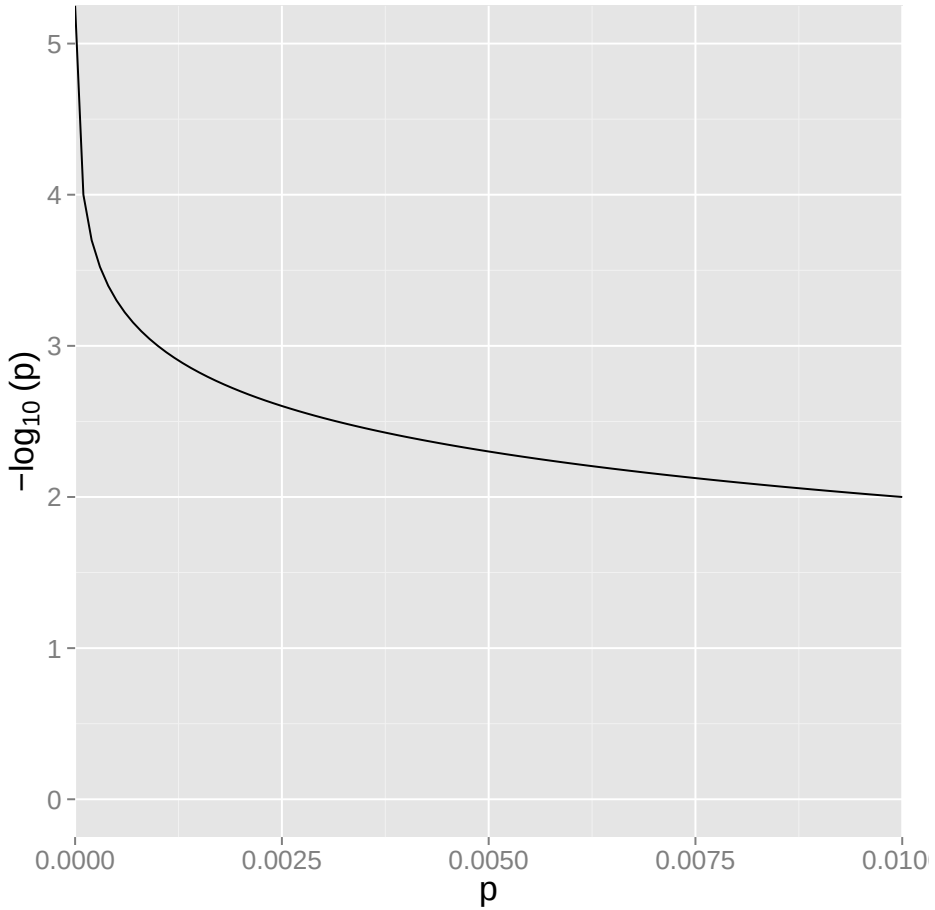
Hyperparameter	**Value**
Population size	1000
Number of generations	100
Aggregation function	mean
Activation function	tanh, Gaussian
L2 regularization λ	0.5
Probability of mutation adding input	0.05
Probability of mutation swapping input	0.05
Probability of mutation adding connection	0.05
Probability of mutation adding node	0.03
Probability of mutation changing weight	0.04
Elitism proportion	0.1
k tournament selection	2
Coefficient 1	1.0
Coefficient 2	1.0
Coefficient 3	0.4
Compatibility threshold	3.0

Mean normalization and Softmax

$$x_{new} = \frac{x - \mu}{x_{max} - x_{min}} \tag{6}$$

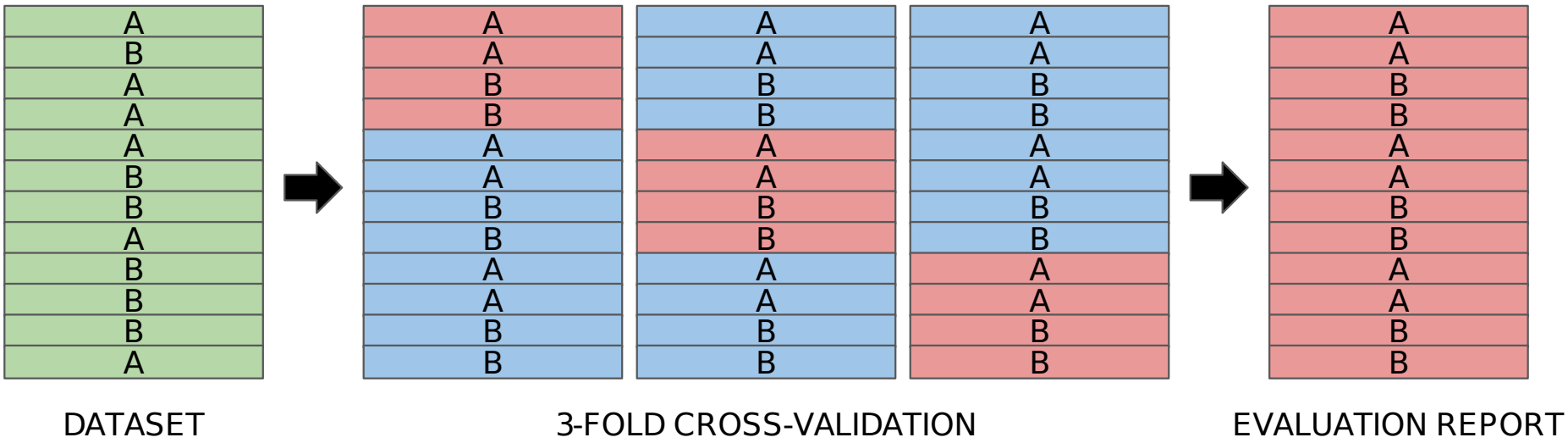
$$softmax(Z) = e^{z} \div \sum_{z' \in Z} e^{z'}, \forall z \in Z \tag{7}$$

Log transformation



The p-values $-\log_{10}$ transformation

Cross-validation



Stratified 3-fold cross-validation

Baseline

$$baseline(D) = \max(\frac{|D_A|}{|D|}, \frac{|D_B|}{|D|}) \tag{8}$$

$$p = \frac{m}{G} \tag{9a}$$

$$\binom{A}{s} = \frac{A!}{s!(A-s)!} \tag{9b}$$

$$P_g[X = s] = \binom{A}{s} p^s (1-p)^{A-s} \tag{9c}$$

$$P_g[X \geq s] = 1 - (P_g[X = 0] + P_g[X = 1] + ... + P_g[X = s-1]) \tag{9d}$$