Preliminary study of entropy-based indicators to discriminate cancer-related characteristics

R.E. Monge-Gapper; Universidad Cenfotec (Costa Rica)
rmonge@ucenfotec.ac.cr

J.L. Crespo-Mariño; IEEE Electron Devices Society Chapter (Costa Rica)
juan.crespo@ieee.org

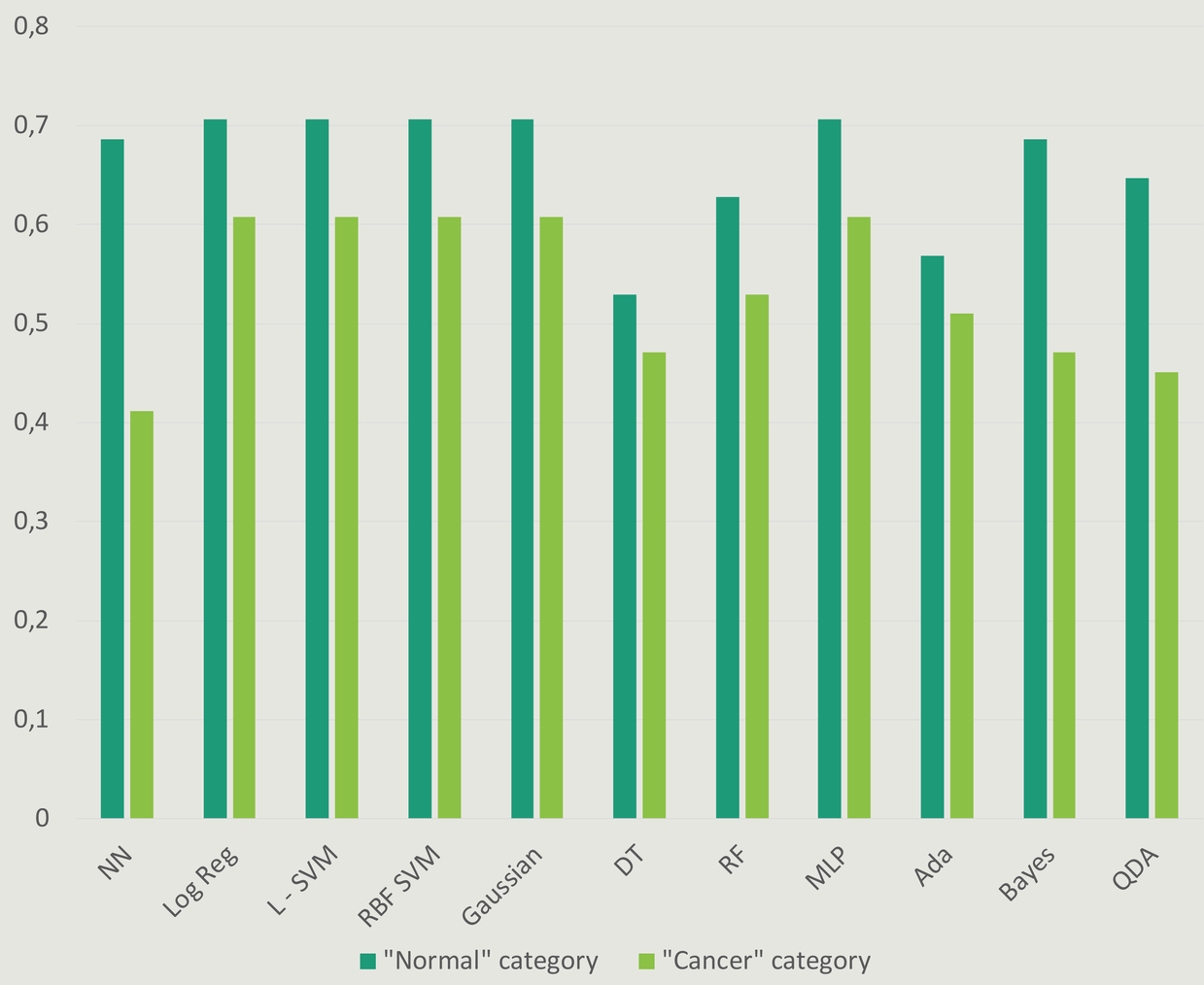
Abstract

Select entropy-based indicators have been used in this preliminary study to classify genes with acceptable results. This need for classification is driven by the interest of the scientific community in determining whether a given gene possesses or lacks cancer-related characteristics without the need for using any type of analytic model for their genetic information. A subset of genes was chosen and have been divided into two groups: those that have a relation to cancer (that is, they either cause cancer, as in oncogenes, or are tumor suppressors) and those that are not related to cancer issues (i.e., normal genes). A set of eleven classifiers were used and compared, some of which reflected an accuracy rate of over 70% of correct predictions (cancer-related or not) within a test set of genes. These results shed some light on the fact that, in effect, oncogenes and normal genes have different patterns and structures and can potentially be used as a predictor for novel genes.

Methodology

- **A gene pool of 204 genes were selected from the NCBI gene database. Of those, 76 were oncogenes (i.e. formally related to cancer); 75 were normal (i.e. required for normal functioning); and 53 were tumor suppressors (i.e. genes that, in spite of an active oncogene, can block the protein synthesis and this, prevent the proliferation). Recall that most research today is oriented toward tumor suppression, and not to tumor probabilistics.**
- **A testing set composed of 30% of the gene database (respecting the three proportions already described) was selected and set aside before the classifiers were applied on the data.**
- **Genetic data was processed using the three metrics described, a) using sliding windows of 300 DNA bases; and b) computing a single value for the full gene; and authors feed the statistically summarized results to the classifiers (minimum, maximum, average when using a sliding window) and the single value for the full gene. This gives a set of 12 discrete values for each gene in the pool of 204 genes; and one categorical result ("cancer","normal", and "suppressor").**

Results



Conclusions and Future Work

- **The initial experiments conducted here point to the fact that standard classifiers and some non-classic classifiers (such as a multilayer neural network) have enough information to give a highly usable characterization of whether a gene has cancer-related traits or not. In spite of not being optimally high (70%), the analysts estimate is greatly enhanced.**
- **Future work will be directed toward the use of more robust classification techniques, as higher order networks, convolutional ones and associated deep-learning classifiers to better discern within cancer-causing and tumor-suppressing genes.**

References and Acknowledgements

[1] Monge, Ricardo E., and Juan L. Crespo. "Analysis of data complexity in human dna for gene-containing zone prediction." Entropy 17, no. 4 (2015): 1673-1689.

[2] Namazi, Hamidreza, and Mona Kiminezhadmalaie. "Diagnosis of lung cancer by fractal analysis of damaged DNA." Computational and mathematical methods in medicine 2015 (2015).

[3] Barman, S., M. Roy, S. Biswas, and S. Saha. "Prediction of cancer cell using digital signal processing." Annals of the Faculty of Engineering Hunedoara 9, no. 3 (2011): 91.

[4] Namazi, Hamidreza, Amin Akrami, Jamal Hussaini, Osmar N. Silva, Albert Wong, and Vladimir V. Kulish. "The fractal based analysis of human face and DNA variations during aging." Bioscience trends (2016).

[5] Namazi, Hamidreza, Vladimir V. Kulish, Fatemeh Delaviz, and Ali Delaviz. "Diagnosis of skin cancer by correlation and complexity analyses of damaged DNA." Oncotarget 6, no. 40 (2015): 42623.

[6] Frost, J. James, Kenneth J. Pienta, and Donald S. Coffey. "Symmetry and symmetry breaking in cancer: a foundational approach to the cancer problem." Oncotarget 9, no. 14 (2018): 11429.

[7] Skliar, Osvaldo, Ricardo E. Monge, Guillermo Oviedo, and Víctor Medina. "Indices of regularity and indices of randomness for m-ary strings." Revista de Matemática: Teoría y Aplicaciones 16, no. 1 (2009): 43-59.

[8] Kolmogorov, Andrei N. "On tables of random numbers." Theoretical Computer Science 207, no. 2 (1998): 387-395.

[9] Shannon, Claude Elwood. "A mathematical theory of communication." ACM SIGMOBILE mobile computing and communications review 5, no. 1 (2001): 3-55.

Authors thank the financial support by IEEE Electron Devices Society Costa Rica Section Chapter. A "internal report version" more detailed version of this poster can be downloaded from https://tinyurl.com/Entropy2021