Preliminary Study of Entropy-Based Indicators to Discriminate Cancer-Related Characteristics †

Juan Luis Crespo Mariño [1] and Ricardo Monge-Gapper [2]

[1] IEEE Electron Devices Society Chapter (Costa Rica), Cartago, Costa Rica

[2] Cenfotec University, San José, Costa Rica

† Presented at the Entropy 2021: The Scientific Tool of the 21st Century, 5–7 May 2021; Available online: https://sciforum.net/conference/Entropy2021/.

Select entropy-based indicators (such as Kolmogorov Complexity, Shannon Information Entropy and the Index of Regularity) 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. A subset of genes was chosen, based on previous studies and on random selection. These genes have been represented by their DNA sub-sequence 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). Initially, eleven classifiers were used and compared, some of which reflected an accuracy rate of over 70%. This accuracy rate represents the percentage 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 and features. This exploratory study also analyzes non-classic classifiers and evaluates the prospects of clustering and advanced machine-learning algorithms to determine significant patterns within DNA sequences.