evolp: A repository for mutations and gene expression data

Dario A. Leon[a1], Joan Nieves[2], Roberto Herrero[3], Frank Quintela[4,1], and Augusto Gonzalez[5,1]

[1]Institute of Cybernetics, Mathematics and Physics, 10400, Havana, Cuba
[2]Faculty of Physics, University of Havana, 10400, Havana, Cuba
[3]Instituto Nacional de Investigaciones en Metrología, 10200, Havana, Cuba
[4]University of Modena & Reggio Emilia, 41125, Modena, Italy
[5]University of Electronic Science and Technology, 610051, Chengdu, People Republic of China

We present evolp*, a repository with a set of tools for analyzing mutations and gene expression data. We discuss some examples of applications, already published elsewhere.*

The emerging fast sequencing techniques in biology have made possible the acquisition of the whole-genome data of many organisms. Moreover, sequencing provides direct observations of genetic changes that can be associated with specific phenotypes and diseases. It is also possible to measure gene expression (GE) levels in individual cells and tissues. The availability of massive databases has motivated the development of several software tools for understanding the different kind of data, and the biological processes and phenomena behind them. In this regards, we are developing an open-source repository, shortly called *evolp*, for processing and analyzing mutations and gene expression data measured in different systems.

In this brief review we summarize a few results coming from the data and python modules contained in the *evolp* repository. More detailed descriptions can be found in the cited articles and the references therein.

The Levy character of mutations

Mutations can be classified according to the number of base-pairs (*bp*) changed in the DNA molecule. Point mutations involve 1 bp whereas many bps are modified in chromosomal rearrangements. The rate of both kind of events in bacteria have been estimated in Ref. [1] from data of a long-term evolution experiment (LTEE) with E. Coli populations.

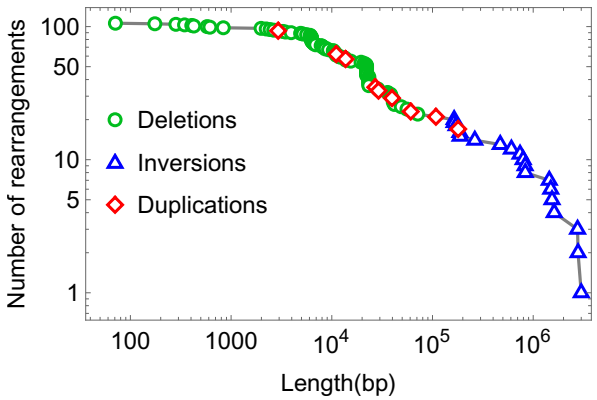

Figure 1: Log-log plot of the integrated size distribution of large rearrangements in the LTEE resulting from one of the processing tools of *evolp*.

Focusing on the size, there are many types of chromosomal rearrangements, e.g. insertions, deletions, duplications, translocations and inversions of DNA fragments, which correspond to different scales and mutation mechanisms. However, the distribution function for the lengths of the modified segments, independently of their type, is found to be a Levy or Pareto-like distribution [1]. In Fig. 1 we plotted the integrated size distribution of the mutations found in the LTEE, according to the observed types. Interestingly, the same kind of length distribution is found for copy number variants data coming from more than 100000 subjects of European ancestry [1], suggesting an universal Levy character of the mutations.

On the other hand, an algorithm for modeling the evolutionary dynamic of the fitness and the clonal competition of the LTEE populations has been developed [2], using the distribution of fitness effects of the mutations as the key ingredient.

Analyzing the gene expression space

The space of all possible configurations of the gene expressions, like the mutation space, is a multi-dimensional space with a large number of components. A region in this space defines a given biological state of the system [3], for instance, of the ancestral or the evolved populations in the LTEE. Also the progression of diseases like dementia and cancer can be thought of as a transition in GE space from an initial or normal state, to a final or disease state. The state of such complex genetic diseases is not simply described, because there is not a direct connection between their causes and unusual levels of GE in one or few genes. However, by using a technique for dimensional reduction called principal component analysis (PCA) to the GE data, the initial and final states are well separated in one or two variables [3].

The PCA processing of GE data from the LTEE, the Allen Institute study of aging and dementia and The Cancer Genome Atlas (TCGA) has revealed different kinds of transitions from the initial state to the final state, a "continuous" transition in the cases of LTEE and dementia, and a "discontinuous" transition in the case of cancer [3]. Continuous and discontinuous

transitions corresponds to slight or radical steps in the progression.

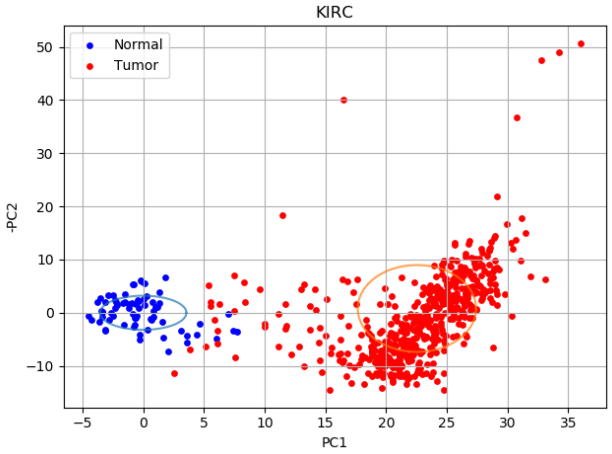

Figure 2: Principal component analysis of the gene expression data of KIRC.

In Fig. 2 we plot the first two PCA components of the GE data corresponding to Clear Cell Kidney Cancer (KIRC). The plot is similar to the one in Ref. [3] and has been generated with one of the tools of the *evolp* repository. In this basic example, only 1000 out of 60000 genes present in the databases are processed in order to reduce the computational cost which allowed to run the script in few minutes in a regular desktop computer. Nevertheless the separation of the normal and the cancer zones differs just slightly from the full processing, demonstrating the robustness of the findings.

The results of the PCA processing applied to the TCGA data of sets of tissues have been analyzed in different ways [4, 5, 6]. The concentration of samples in GE space is linked to the fitness landscape of a particular system, where the normal and tumor zones correspond to local maxima and the intermediate zone defines a low-fitness barrier. The discontinuous nature of the transition from normal to tumor zones has been related with the Levy character of the mutations [4] mentioned in the previous section. A one dimensional model of tumorigenesis, based on the first principal component of PCA, has been proposed for describing the cancer lifetime risk [4]. The model uses as ingredients only geometric properties in the GE space, such as the minimum distance between the two zones. In support of this picture, a correlation has been found between the computed model in 8 different tissues, where all the data is available, including data on cancer risk taken from other databases [4]. On the other hand, the number of available states in the normal and tumor zones of 12 different tissues has been estimated by using an entropy-like magnitude [5]. The approach uses data on the variance in PCA and some ideas from Information Theory in order to define an "effective dimension". This dimension is used to compare the hypervolumes of the basins of attraction of the normal and cancer attractors in GE space. The number of accessible states is found to be much higher for tumors than for normal samples [5]. A second magnitude characterizing the topology of the GE space is the overlapping between the normal and tumor clouds, which is found to be correlated differently with the entropy [5] for the cancer and normal states. The former have roughly the same number of accessible states, whereas the second exhibits a higher variability.

The computed weights of the gene expression profile in the PCA basis have been used to determine the most under or over-expressed genes in the cancer state, founding 6 common genes from 15 studied tumors [6]. Moreover, the application of PCA to RNA-seq data has unveiled a novel signature in GE associated to prostate cancer [7].

Conclusions

The usefulness of the processing tools of *evolp* has been exemplified in several independent works that coinvolve different databases, and it is expected to keep growing in the future.

Notes

a. Email: dario@icimaf.cu

References

[1] D. A. Leon, A. Gonzalez. Mutations as Levy flights. *Sci. Rep.* **11**:9889, 2021

[2] D. A. Leon, A. Gonzalez. Modeling evolution in a Long Time Evolution Experiment with E. Coli. `https://arxiv.org/abs/1804.02660`, 2020

[3] Augusto Gonzalez, Joan Nieves, Dario A. Leon, Maria Luisa Bringas, Pedro Valdes-Sosa. Gene expression rearrangements denoting changes in the biological state. *Sci. Rep.* **11**:8470, 2021

[4] Roberto Herrero, Dario A. Leon, Augusto Gonzalez. A one-dimensional parameter-free model for carcinogenesis in gene expression space. *Sci. Rep.* **12**:4748, 2022

[5] Augusto Gonzalez, Frank Quintela, Dario A. Leon, Maria Luisa Bringas-Vega, Pedro A. Valdes-Sosa. Estimating the number of available states for normal and tumor tissues in gene expression space. *Biophysical Reports* **2**(2):100053, 2022

[6] Augusto Gonzalez, Dario A. Leon, Yasser Perera, Rolando Perez. On the gene expression landscape of cancer. `https://arxiv.org/abs/2003.07828v3`, 2020

[7] Yasser Perera, Augusto Gonzalez, Rolando Perez. Principal component analysis of RNA-seq data unveils a novel prostate cancer-associated gene expression signature. *Arch. Can. Res.* **9**(S4):002, 2021