Comment

Ensemble of correlation, parenclitic and synolitic graphs as a tool to detect universal changes in complex biological systems. Comment on "Dynamic and thermodynamic models of adaptation" by A.N. Gorban et al.

Tatiana Nazarenko[1], Oleg Blyuss[1,2,5,7], Harry Whitwell[3,4,5,6], and Alexey Zaikin[1,5,6]

[1] *Department of Mathematics and Institute for Women's Health, University College London, London, UK*

[2] *School of Physics, Astronomy and Mathematics, University of Hertfordshire, Harfield, UK*

[3] *National Phenome Centre and Imperial Clinical Phenotyping Centre, Department of Metabolism, Digestion and Reproduction, IRDB Building, Imperial College London, Hammersmith Campus, London, W12 0NN, UK*

[4] *Section of Bioanalytical Chemistry, Division of Systems Medicine, Department of Metabolism, Digestion, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK*

[5] *Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia*

[6] *Centre for Analysis of Complex Systems, Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia*

[7] *Department of Pediatrics and Pediatric Infectious Diseases, Institute of Child's Health, Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia*

Keywords: Adaptation, universality, thermodynamics, complexity, correlation graphs, parenclitic graphs

Complexity is a natural feature of biological systems. For example, a human can be characterised by a vast amount of information, packed in a super-network that is, in principle, a network of networks [1]. For each human cell, the total amount of DNA, if measured in bases, has an informational length of the 3.2 Gb. However, in total, we have approximately 10^{12} cells, and in each of them, even neglecting mosaic mutations, the genotype information is multiplied by the epigenome, transcriptome, proteome and metabolome profiles, which are organized in a huge and very dynamic connectome. Each tissue can be represented by a graph or network and connected to other networks, resulting in a super-network containing a huge number of features describing a complex biological system. When the biological system ages, or approaches a diseased state, the bodies adaptation to maintaining homeostasis is reflected in the changes of this super-network.

Due to the fascinating progress in multi-omic experimental techniques, we are getting access to increasingly huge quantities of biological data. But it is clear, that the amount of information and, more importantly, the complexity of its organisation, exceeds our analytical capability. The situation is even more complicated, because, for example, as noted by Gorban et al. in the analysis of cancer omics data, underlying changes may occur in different places of the network. A possible solution to the challenging problem of identifying biomarkers and disease-precursors is in graph theory, using well-described topological descriptors to identify changes in network structure that represent disease. In this sense, the "shape" of the network is used, rather than specific molecular biomarkers. However, a real breakthrough can only occur if we are able to define the behaviour of the networks to a relevant stimulus (e.g. disease), in a generic way, for example, based on the fundamental laws of thermodynamics.

Representing high-dimensional biological data in the form of a graph and linking features by biological and thermodynamic laws seems to be a very promising approach to deal with overwhelming complexity of biological

Email address: alexey.zaikin@ucl.ac.uk (Tatiana Nazarenko[1], Oleg Blyuss[1,2,5,7], Harry Whitwell[3,4,5,6], and Alexey Zaikin[1,5,6])

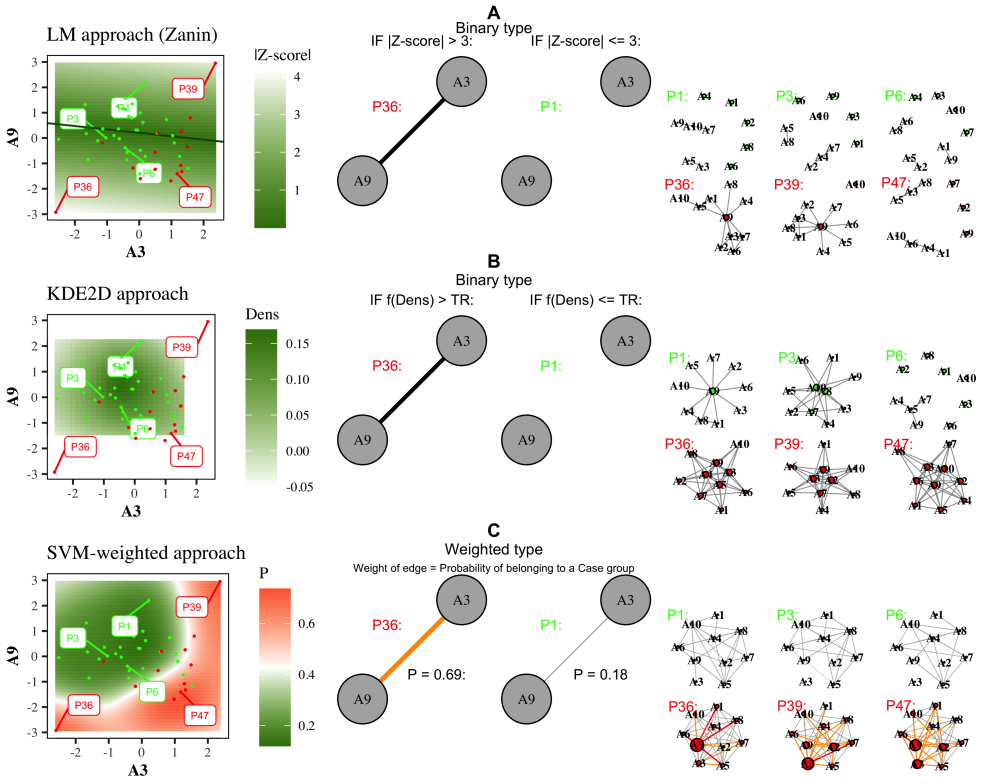

Figure 1: Parenclitic and synolytic networks as particular types of a correlation graph: **(A)** (*Original parenclitic network [3]*). This is a network where only the control group was considered as the basis for determining the normal state on the plane of two signs (based on control group the linear regression was built, the deviation of the control points from it was calculated and the distribution of such deviations was constructed). For each new sample, the edge weight was first determined as the absolute value of z-score, and then binarized (if z-score is greater than 3, then the edge is present in the sample network, otherwise there is no edge). **(B)** (*KDE-parenclitic network [4]*). This is a network in which again only the control group was considered as the basis for determining the normal state on the plane of two signs (for the control group, the KDE2D density was built, then a function was calculated that converts the density values into an analogue distance (so that the points located in the area of the highest density have the minimum weight); the distance outside the grid was continued (for more details, see [4]). For each new sample, the edge weight was first determined as the value of the entered distance, and then converted to binary form (if the distances are greater than a threshold, then the edge is present in the sample network, otherwise there is no edge). **(C)** *SVM-weighted synolytical network*. This is a network in which both groups participate in defining normal and abnormal states. On the plane of two features, with the help of the radial SVM, the best boundary separating the classes is drawn. Automatically, each point in such a model gets a value for the probability of belonging to each class. For each new sample, the edge weight is determined as the probability of belonging to a group of cases.

systems. However, one can utilise this approach only if we have information about how features and attributes are connected biologically. If we do not have such information, we must use more sophisticated methods to represent a complex system in the form of the graph in high dimensional space. Such methodology is provided by correlation graphs which are carefully reviewed in fascinating and inciteful detail in the review by Gorban and co-authors [2]. This paper considers two types of correlation graph. Firstly, where edges correspond to the correlations between attributes of different people, with populations represented as a network and secondly, where edges correspond to different time intervals between data belonging to the same person. Most importantly, the authors reviewed and discussed a link between changes in the topology and laws of a general theory of adaptation and thermodynamics. In particular, it is discussed that preceding a network crash or critical state, there is a detectable redistribution of resources within the system resulting in the strengthening of correlations, breaking of topological symmetry and increase of fluctuation variances.

In this comment we would like to draw attention to alternative methods to represent high dimensional data in the form of the graph if *a-priori* we do not have established connections. There are several other methods that we will

briefly mention here.

First of all, correlation-prediction graphs can be used as a marker of survival and have been constructed to represent the gene methylation profiles of individuals [5, 6]. Secondly, there is an algorithm, first described by Zanin and Bocaletti, able to establish links between parameters/nodes without any *a-priori* knowledge of their interactions [3] using residual distances from linear regression models constructed between every pair of analytes to construct a graph. They termed this approach a "parenclitic" network representation, from the Greek term for "deviation". Parenclitic networks have been successfully applied to problems of the detection of key genes and metabolites in different diseases [3, 7–9] see [10] for a review and [11] for a discussion of applications for brain research. In [12] we have applied this methodology to implement machine learning classification to identify signatures of cancer development from human DNA methylation data. Thirdly, based on the understanding that the interactions of two features (at least in biological systems) often cannot be described by a linear model, it was proposed to use 2-dimensional kernel density estimation (2DKDE) to model the control distribution [4]. Finally, in [13] we have introduced a variation of parenclitic networks, that can be called synolytic from the Greek word for "ensemble". In principle, these networks can be considered an ensemble of classifiers in a graph form and thus are a kind of correlation network where the correlation is in the changes between two classes (e.g. disease and non disease). The difference to original parenclitic networks is visualized in the Fig. 1. These networks have been successfully used to detect age related trajectories in Down's syndrome [13] and for prediction of survival for severely ill Covid-19 patients [14].

All these network approaches could be described as variations of correlation networks with different levels of dimensionality. In other words, an ensemble of different correlation graphs can be used to characterise the same complex biological system. This ensemble of graphs as a classifier can provide us with more efficient diagnostic methods than single classifiers. The motivation behind combining the classifiers as an ensemble is based on the classical Condorcet's Jury theorem stating that the probability of a given group of individuals (classifiers) arriving at a correct decision (correct class label) based on a majority vote, increases with the total number of voters. The theorem requires that one of the two outcomes should be correct, and each voter has an independent probability $p > 1/2$ of voting for the correct decision. The assumption that the votes are independent is not always easy to satisfy but a common approach is to train the classifiers on different subsets of the data. A further advantage of combining classifiers into an ensemble is a gain in robustness to noise that may accompany the data.

Designing an optimal aggregation scheme for combining multiple classifiers remains challenging [15] despite a number of methodologies suggested. It is very promising to exploit these different methodologies to construct correlation graphs and use them as an ensemble to develop new diagnostic tools. It would be especially valuable to link topological features of these graphs with universal phenomena, explainable with thermodynamic laws as it was discussed in [2], thus providing a generalisation that can greatly benefit their ability to predict disease.

Acknowledgement

We thanks support from MRC MR/R02524X/1 and the Ministry of Science and Higher Education of the Russian Federation (project no. 075-15- 2020-808).

References

[1] H. J. Whitwell, M. G. Bacalini, O. Blyuss, S. Chen, P. Garagnani, S. Y. Gordleeva, S. Jalan, M. Ivanchenko, O. Kanakov, V. Kustikova, et al., The human body as a super network: Digital methods to analyze the propagation of aging, Frontiers in aging neuroscience 2020;12:136.

[2] A. Gorban, T. Tyukina, L. Pokidysheva, E. Smirnova, Dynamic and thermodynamic models of adaptation, Physics of Life Reviews 2021;37:17–64. https://doi.org/10.1016/j.plrev.2021.03.001.

[3] M. Zanin, S. Boccaletti, Complex networks analysis of obstructive nephropathy data, Chaos: An Interdisciplinary Journal of Nonlinear Science 2011;21:033103.

[4] H. J. Whitwell, O. Blyuss, U. Menon, J. F. Timms, A. Zaikin, Parenclitic networks for predicting ovarian cancer, Oncotarget 2018;9:22717–726. https://doi.org/10.18632/oncotarget.25216.

[5] T. E. Bartlett, A. Zaikin, Detection of epigenomic network community oncomarkers, The Annals of Applied Statistics 2016;10:1373–96. https://doi:10.1214/16-AOAS939.

[6] T. E. Bartlett, S. C. Olhede, A. Zaikin, A dna methylation network interaction measure, and detection of network oncomarkers, PloS ONE 2014:9:e84573.

[7] M. Zanin, E. Menasalvas, P. A. Sousa, S. Boccaletti, Preprocessing and analyzing genetic data with complex networks: An application to obstructive nephropathy, Networks & Heterogeneous Media 2012;7:473.

[8] M. Zanin, E. Menasalvas, S. Boccaletti, P. Sousa, Feature selection in the reconstruction of complex network representations of spectral data, PloS ONE 2013;8:e72045.
[9] M. Zanin, D. Papo, J. L. G. Solís, J. C. M. Espinosa, C. Frausto-Reyes, P. P. Anda, et al., Knowledge discovery in spectral data by means of complex networks, Metabolites 2013;3:155-67.
[10] M. Zanin, D. Papo, P. A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, S. Boccaletti, Combining complex networks and data mining: why and how, Physics Reports 2016;635:1–44.
[11] D. Papo, J. M. Buldú, S. Boccaletti, E. T. Bullmore, Introduction: Complex network theory and the brain, Philosophical Transactions: Biological Sciences 2014;369:1–7
[12] A. Karsakov, T. Bartlett, A. Ryblov, I. Meyerov, M. Ivanchenko, A. Zaikin, Parenclitic network analysis of methylation data for cancer identification, PloS ONE 2017;12: e0169661.
[13] M. Krivonosov, T. Nazarenko, M. Bacalini, C. Franceschi, A. Zaikin, M. Ivanchenko, DNA methylation changes with age as a complex system: a parenclitic network approach to a family-based cohort of patients with down syndrome, bioRxiv `https://doi:10.1101/2020.03.10.986505`.
[14] V. Demichev, P. Tober-Lau, T. Nazarenko, C. Thibeault, H. Whitwell, O. Lemke, et al., A time-resolved proteomic and diagnostic map characterizes covid-19 disease progression and predicts outcome, medRxiv `https://doi:10.1101/2020.11.09.20228015`.
[15] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms, John Wiley & Sons, 2014.