Adaptation through the lense of single-cell multi-omics data Comment on "Dynamic and thermodynamic models of adaptation" by A.N. Gorban et al.

Andrei Zinovyev[1,2,3]

[1]Institut Curie, PSL Research University, F-75005 Paris, France
[2]INSERM, U900, F-75005 Paris, France
[3]CBIO-Centre for Computational Biology, Mines ParisTech, PSL Research University, 75006 Paris, France

Keywords: single-cell data, omics data, adaptation, stress, cancer

A.N.Gorban and his colleagues in their inspiring review described several theoretical models of adaptation and highlighted multiple examples convincing us in the existence of surprising at first thought phenomenon: the pattern of dynamical changes of basic statistical measures (correlation between features, their variance) can diagnose and prognose crises in the populations of objects exposed to stress (1). The surprise is caused by the universality of this observation. Firstly, it is manifested in many different situations. Secondly and even more surprising, the features used do not have to be specifically designed to measure stress, even though the feature selection is still important. This suggests that the proposed models can serve as an insightful approach for Big Data analysis and interpretation.

One can note that all provided examples deal with macroscopic objects (people, patients, mice, plants, companies at the stock market). What if we will change the focus of the middle-out approach to a significantly smaller scale, e.g. from a tumor or a patient to a single cell? Would the suggested principles of adaptation thermodynamics still hold and what specific problems will arise?

Modern biotechnologies revolutionized biology and medicine by providing single-cell tools able to collect many thousands or even millions of quantitative traits per cell (typically, counts of RNAs transcribed from each gene in the genome, large-scale genomic modifications, profiling the chromatin state) (4). Today we can take a piece of biological tissue (e.g, a tumor), dissociate it into millions of cells each of which is quantitatively characterized, for example, by expression (as counts of RNAs) of about 20000 distinct genes. Therefore, each biological sample can be represented as a multi-dimensional data point cloud in R^{20000} which properties can be characterized by gene-gene correlation graph and variance together with other characteristics.

Single-cell measurements are frequently used to study the effect of a stressful perturbation applied to cell populations. It can be a toxic treatment of cancer cells but also such general stresses as hypoxia or starvation. The stress can be extrinsic or intrinsic, e.g, resulting from non-physiological modifications of the cell genome. The main question here consists in how the cell population adapts to the caused stress by modifying the cell states and their distribution, and is the picture of high-dimensional omics measurements informative for characterizing the adaptation process.

The examples of such studies are numerous. We can briefly mention two recent ones from cancer biology. Early (in the first three days) adaptation of melanoma cell population was investigated after application of a potent targeted anticancer drug dabrafenib, belonging to the family of so-called BRAF inhibitors (2). Most of the cells adapt to such treatment by entering quiescence but some "escapees" manage to maintain their cell cycle and eventually outproliferate other cells. Second example comes from Ewing sarcoma, a rare pediatric cancer, which is caused by a mutation translocating two chromosomes and leading to the appearance of a chimeric (non-physiological) transcription factor called EWS-FLI1. EWS-FLI1 can bind to short microsatellite sequences which are distributed more or less randomly in the human genome and perturbs the expression of thousands of genes. The effect of this perturbation is usually lethal for most cell types, but some (such as mesenchymal stem cells in bone marrow) can adapt, survive and even start to actively proliferate, giving rise to Ewing sarcoma cell populations. Experimentally we could modulate the expression of EWS-FLI1 and observe how the cancer cell population adapts (3).

It looks very tempting to apply the thermodynamic models described in (1), to study the adaptation in single cell populations. Indeed, similarly to human beings, individual cells can possess variable levels of adaptation energy and

resources: therefore, most of the assumptions look valid. Moreover, the interplay between correlation and variance is commonly investigated in single-cell data analysis with such statistical concepts as overdispersion, differential correlation or differential variance being widely exploited (4). Network entropy or dynamical covariation pattern analysis in genomic profiles have been used to anticipate crises that can manifest themselves in initiating, for example, carcinogenesis (5). We showed that the local intrinsic dimensionality of data point clouds (measure related to the weight of the correlation graph) is an informative and biologically meaningful feature of single-cell datasets (6).

However, from our experience, interpreting the weight of the correlation graph and the variance changes computed from single-cell data, remains tricky. Let us just briefly describe some of the challenges that are faced in this way.

The first evident observation is that single-cell data are noisy and contain a large number of empty measures (so called technical drop-outs) (4). These data are much noisier than the so-called "bulk" (using averages over millions of cells) measurements discussed in (1). Various tricks such as using micro-pooling or variance stabilization are frequently used to improve the between-gene correlation structure in single-cell data, which, of course, impacts the quantification of the correlation graph.

Secondly, most omics-based single-cell measurements are destructive, which means that even if we track a cell population in time, we are not able to track the state of an individual cell. Here one deals with synchronic (as opposite to diachronic) data and usually not synchronized cell population snapshots. In single cell data science, a special family of computational methods, based on the concepts of pseudo-time, cell trajectories and data manifold alignment have been developed in order to match such snapshots from different timepoints and making them quasi-diachronic (4; 7). This way of treating the data might affect the conclusions of the analysis based on the weight of the correlation graph.

Thirdly and quite fundamentally, the role of selection and inheritance can be much more important when studying adaptation to stress at the level of cells rather than whole organisms. Thus, cancer cells in culture are rapidly dividing at the time scale of experimental perturbation, which is not the case in the examples described in (1). The adaptive changes are inheritable to some extent after cell division. The difference between genetic and epigenetic inheritance conceptually consists only in the time scale for which the cell state modifications persist. Therefore, part of the changes in the structure of variance and correlations in a single cell population can be attributed to positive or negative selection. Whether to consider selection a part of adaptation, and also potential antagonism between instantaneous and population-based Darwinian fitness, need to be carefully considered in the analysis.

Forthly, defining cellular functional subsystems, an ingredient necessary for building thermodynamic adaptation models, can be not obvious. Methods based on matrix factorization and, in particular, Independent Component Analysis (ICA) seem to work well for splitting the set of genes into biologically meaningful groups with coordinated gene expression changes. In case of ICA, these collective changes are forced to be as statistically independent as possible (8), which reproducibly identifies such functional subsystems as DNA replication, mitotic programs (performing and exiting from mitosis), metabolic programs such as oxphos and glycogenesis, cell migration, response to hypoxia and others (3). Scoring these functional subsystems defines a biologically tractable space of cell states, where one can observe dynamical coupling between functional subsystems under stress, increase of variance in cell state distribution, etc. Such representation of single-cell data might be more instructive than the use of the raw count data.

Finally, the most important in cancer biology cellular functional subsystem is the cell cycle machinery. Since the individual cells are usually not synchronized, the progression through the cell cycle creates a strong signal which can dominate the pattern of variance and correlation. We observed that in slowly proliferating cell populations, the correlations between cell cycle genes can be stronger as well as their total variance compared to rapidly dividing cells. This is probably caused by that in rapidly dividing cells, some of the cell cycle-related transcripts do not possess enough time to be fully degraded after exiting from mitosis, which leads to their smaller expression amplitude. Interestingly, slowing down the cell cycle is a known cellular mechanism of adaptation to stress (9).

To conclude, we are convinced that the models of adaptation, based on the application of "middle-out" ideas suggested in (1) must become a working tool for making sense of single-cell omics datasets. However, further work, both theoretical and related to practical data analysis, is needed to take into account the specificity of the data and the studied systems, which should lead to verifiable and reproducible conclusions.

Acknowledgements

References

[1] Gorban AN, Tyukina TA, Pokidysheva LI, Smirnova EI, Dynamic and thermodynamic models of adaptation. Phys. Life Rev. 2021;37:17–64. `https://doi.org/10.1016/j.plrev.2021.03.001`

[2] Yang C, Tian C, Hoffman TE, Jacobsen NK, Spencer SL, Melanoma subpopulations that rapidly escape MAPK pathway inhibition incur DNA damage and rely on stress signalling. Nat. Comm. 2021;12:1–14. `https://doi.org/10.1038/s41467-021-21549-x`

[3] Aynaud MM, Mirabeau O, Gruel N, Grossetete S, Boeva V, Durand S, Surdez D, Saulnier O, Zai?di S, Gribkova S, Fouche A, Kairov U, Raynal V, Tirode F, Gruunewald TG, Bohec M, Baulande S, Janoueix-Lerosey I, Vert JP, Barillot E, Delattre O, Zinovyev A., Transcriptional programs define intratumoral heterogeneity of Ewing sarcoma at single-cell resolution. Cell Rep. 2020;30:1767–1779.e6. `https://doi.org/10.1016/j.celrep.2020.01.049`

[4] Laehnemann D, Koster J, Szczurek E, McCarthy D, Hicks S, Robinson M, Vallejos C, Campbell K, Beerenwinkel N, Mahfouz A, Pinello L, Skums P, Stamatakis A, Attolini C, Aparicio S, Baaijens J, Balvert M, de Barbanson B, Cappuccio A, Schoonhuth A, Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):31. `https://doi.org/10.1186/s13059-020-1926-6`

[5] Teschendorff AE, Liu X, Caren H, Pollard SM, Beck S, Widschwendter M, Chen L, The Dynamics of DNA Methylation Covariation Patterns in Carcinogenesis. PLoS Comp. Biol. 2014;10:e1003709. `https://dx.plos.org/10.1371/journal.pcbi.1003709`

[6] Albergante L Bac J Zinovyev A. Estimating the effective dimension of large biological datasets using Fisher separability analysis. In Proceedings of the International Joint Conference on Neural Networks. IEEE. 2019, N-19814. `https://doi.org/10.1109/IJCNN.2019.8852450`

[7] Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, Zhou S, Gorban AN, Bauer DE, Aryee MJ, Langenau DM, Zinovyev A, Buenrostro JD, Yuan GC, Pinello L, Single-cell trajectories reconstruction exploration and mapping of omics data with STREAM. Nat. Comm. 2019;10(1):1903. `https://doi.org/10.1038/s41467-019-09670-4`

[8] Sompairac N, Nazarov PV, Czerwinska U, Cantini L, Biton A, Molkenov A, Zhumadilov Z, Barillot E, Radvanyi F, Gorban A, Kairov U, Zinovyev A, Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int J Mol Sci 2019;20(18):4414. `https://dx.plos.org/doi:10.3390/ijms20184414`

[9] Min M, Spencer SL, Spontaneously slow-cycling subpopulations of human cells originate from activation of stress-response pathways. PLoS Biology 2019;17:e3000178. `https://dx.plos.org/10.1371/journal.pbio.3000178`