Clinical trajectories estimated from bulk tumoral molecular profiles using elastic principal trees

*

Alexander Chervov
Institut Curie
Paris, France
alexander.chervov@curie.fr

Andrei Zinovyev
Institut Curie
Paris, France
andrei.zinovyev@curie.fr

27 January 2021

Abstract

Clinical trajectory is a clinically relevant sequence of ordered patient phenotypes representing consecutive states of a developing disease and leading to some final state. Extracting trajectories from large scale medical data is of great interest for dynamical phenotyping of various diseases but remains a challenge for machine learning methods, especially in the case of synchronic (with short follow up) observations. Here we describe an approach for trajectory-based analysis of cancer data using elastic principal trees and test it on a large collection of molecular tumoral profiles for breast cancer. We show that the disease progress quantified with pseudotime (the geodesic distance from the root) along a particular trajectory can serve as a significant prognostic factor, not redundant with gene expression-based predictors. We conclude that application of the elastic principal trees to transcriptomic data can be of interest for clinical applications.

clinical trajectories, breast cancer, transcriptome, principal tree, survival analysis

1 Introduction

Machine learning-based analysis of Big Data in medicine promises to accelerate rationalizing and optimizing treatment of various diseases. For unsupervised type of analysis of medical data, we can distinguish two basic types of approaches for achieving this goal: static analysis based on clustering and the analysis of clinical trajectories. This distinction takes its roots back to the very origin of medicine as a field. Thus, the father of medicine, Hippocrates, was considering any patient with a particular disease as a process that can be characterized by the rules of its dynamics [1]. For example, he introduced the term 'crisis' in a disease as a decisive bifurcation point, determining the patient's fate at rather well defined moments of time. This notion of disease as a specific and complex dynamics is the opposite of the naive idea of a discrete diagnosis that defines disease as a condition or state. Following the ideas of the Koan's school founded by Hippocrates, during many centuries the work of many thousands of best medical experts had the aim to transform the scrupulous observations on the millions of individual patients' disease trajectories into standard medical protocols. Thanks to these efforts, today we can sometimes make a prognosis for a patient, and not merely diagnose his current condition. But we need much more: to identify typical crises and to make a reliable individual forecast.

Today we can investigate the patient's organism more systematically than ever, including the molecular level. Sometimes we can record everything that happens with the whole nations in terms of health into billions of clinical records. For example, a dataset containing an electronic health registry collected during 15 years and covering the whole population of Denmark, with 6.2 million individuals, have been analyzed with an objective to determine previously unreported disease co-morbidities [2]. An ambitious 'Data Health Hub' (`https://www.health-data-hub.fr/`) project has been recently launched in France with the aim to

*This work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the Ministry of Science and Higher Education of the Russian Federation (Project No. 14.Y26.31.0022) and by European Union's Horizon 2020 program (grant No. 826121, iPC project).

make available for machine learning-based analysis the collection of several decades-long population-wide anonymized health insurance records [3].

The Big Data point of view on medicine has been connected for long with the principle of clustering. The state of a patient can be represented as a vector in a multi-dimensional space combining all bits of information collected, including very precise molecular measurements. Collecting many such vectors, one can think of applying an unsupervised machine learning approach and define patient subgroups with similar states. In cancer research, this approach being applied to omics measurements, leads to the definition of molecular cancer subtypes [4]. The hypothesis is that each such a subgroup of patients requires a specific treatment. Quite strikingly, this approach appears to be in direct contradiction with the Hippocrates legacy, since by classifying the snapshot states of the patients, we tend to neglect the dynamical nature of the disease as a process.

The instantaneously observed state of a patient does not tell how the patient arrived there, the process that could take years and be preceded by other diagnoses. Mere classifying the current patient state also does not immediately allow predicting the probable future states. Therefore, today we are facing the paradigm change from defining a disease as a static snapshot of the organism's state towards the notion of an individualized clinical or disease trajectory. The disease trajectory represents the history of a patient with interplay and mutual influences of multiple diagnoses. Millions of such trajectories can be grouped into dynamical phenotypes, representing fewer number of major stereotypical pathological scenarios [2]. Dynamical phenotyping is the conceptual paradigm underlying such studies which can be applied at organismal and cellular scales [5–7]. It states that distinguishing various dynamical types of progression of a disease or a cellular program is more informative than classifying biological system states at any fixed moment of time. From the machine learning point of view, this dictates different choices of methods, with clustering more adapted to the synchronic (snapshot) data [8] while more specific methods for trajectory analysis are needed in the case of diachronic (having important temporal aspect) data [9–13]. Modern methods of artificial intelligence based on collecting large amounts of clinically relevant data can help us to perform such dynamical patient phenotyping.

There is still one big problem in this respect. Reconstructing precise individual clinical trajectory requires long-term follow up of a patient, with systematic collection of the information about the state of the patient's organism. These data (called longitudinal or diachronic observations) remain very difficult and expensive to collect. We have in possession much more synchronic (snapshot) data by observing patients within a relatively short period of time (for example, during the stay in a hospital).

We hypothesize that if the number of synchronic observations is large, then they will map the structure of clinical trajectories because each patient would represent a different state of a progressing disease, along a particular disease trajectory. Then the large-scale clinical data can be modeled as a bouquet of diverging clinical trajectories, even if the nature of the data is synchronic and none of the patients is followed for a long time. The root of this bouquet corresponds to the least complicated disease state, the onset of many possible pathological scenario.

We suggest to model the clinical data using the method of elastic principal trees, based on the idea of topological grammars [14–16]. Recently we developed a Python package, ClinTrajAn, which can be conveniently used in order to extract trajectories from various types of clinical data, using principal trees [17]. This approach has been applied to the analysis of several large-scale clinical datasets such as the database of complications of myocardial infarction and the diabetes management dataset [18]. In these case studies, the datasets contained various measurements of the patient physiology or the details precising the context of the patient's stay in hospital. Such data are extremely valuable but remain difficult to collect and standardize, especially in the context of increasing stringency of the personal data protection legal rules.

With the arrival of high-throughput molecular profiling technologies to clinics, it is tempting to use these data in order to extract trajectories directly from a sufficiently large set of molecular profiles. Indeed, it was observed that in some diseases such as cancer, the molecular measurements can be ordered in clinically relevant fashion by applying linear and non-linear unsupervised machine learning approaches [19, 20]. One of the first attempts to define trajectories in transcriptomic breast cancer datasets was undertaken in [21] using topological data analysis. Principal curves initially suggested by Hastie [22] have been used to model the transcriptomic data in colon and breast cancer [20, 23]. In these studies existence of a single global trajectory without branching was assumed. This approach was further extended to the case of branching trajectories, assembled into a principal tree, leading to refining the breast cancer classification into subtypes and detecting new rare subtypes [24]. In order to extract branching trajectories, the authors used either cluster-wise assembly of principal curves into a global tree or a structure learning approach denoted as reversed graph embedding, similar in spirit to the elastic principal trees but based on simplified heuristics [25].

Here we apply the ClinTrajAn Python pack) age, developed by us earlier, to the analysis of large tran-

scriptomic breast cancer dataset METABRIC, which was previously analysed by clustering approaches [26] and by detecting clinical trajectories [23, 25]. Complementing the previous studies, we focus on the clinical significance of the branching pseudotime quantified with ClinTrajAn. We stored the data and code at kaggle datasets (https://www.kaggle.com/alexandervc/breast-cancer-omics-bulk-data/code), one can use it to make further analysis.

2 Material and methods

2.1 Method of elastic principal trees

Principal tree is a set of principal curves assembled in a tree-like structure, characterized by branching topology [14, 27]. Principal trees can be constructed using ElPiGraph computational tool, which has been previously exploited in determining branching trajectories in various genomics datasets (in particular, in single cell omics data) [16, 28, 29].

Elastic principal graph (ElPiGraph) is an undirected graph with a set of nodes $V = \{V_i\}$ and a set of edges $E = \{E^i\}$. The set of nodes V is embedded in the multidimensional space. In order to denote the position of the node in the data space, we will use the notation $\phi(V_j)$, where $\phi(V_j)$ is a map $\phi : V \to R^m$. The optimization algorithm search for such $\phi(\cdot)$ that the sum of the data approximation term and the graph elastic energy is minimized. The details of the optimization-based approach can be found elsewhere [16, 27, 30, 31].

In this study we used freely available ClinTrajAn Python package built on top of ElPiGraph and facilitating the extraction and the analysis of clinical trajectories [18]. The principal tree inference with ElPiGraph was performed here using the following parameters: $R_0 = \infty$, $\alpha = 0.01$, $\mu = 0.1$, $\lambda = 0.05$. Parameters have the following meaning: R_0 is the trimming radius such that points further than R_0 from any node do not contribute to the optimization of the graph, λ is the edge stretching elasticity modulo regularizing the total length of the graph edges and making their distribution close to equidistant in the multidimensional space, μ is the star bending elasticity modulo controlling the deviation of the graph stars from harmonic configurations (for any star S^j, if the embedding of its central node coincides with the mean of its leaves embedding, the configuration is considered harmonic). α is a coefficient of penalty for the topological complexity (existence of higher-order branchings) of the resulting graph. (The detailed explanations on parameters can be found in [18], section "Method of Elastic Principal Graphs (ElPiGraph)" formulas 2,3,4,5).

The ElPiGraph and ClinTrajAn packages are freely available from `https://github.com/sysbio-curie/ElPiGraph.P` and `https://github.com/sysbio-curie/ClinTrajAn`.

2.2 Projecting data points onto principal tree

An artbitrary vector x – not necessary belonging to the dataset X – can be projected onto the principal tree and receive a position in its intrinsic geodesic coordinates. The projection is achieved by finding the closest point on the principal graph as a piecewise linear manifold, composed of nodes and edges as linear segments connecting nodes. Therefore, the projection can end up in a node or on an edge. In further we define a projection function $\{p, \epsilon\} = Proj(x, G)$, returning a couple containing the index of the edge which is the closest one to x and the position of the projection from the beginning of the edge $E^p(0)$ as a fraction of the edge length $\epsilon \in [0, 1]$. Therefore, if $\epsilon = 0$ then x is projected into $E^p(0)$ and if $\epsilon = 1$ then the projection is in $E^p(1)$. If $\epsilon \in (0, 1)$ then the projection is on a linear segment, connecting $E^p(0)$ and $E^p(1)$.

2.3 Quantifying pseudotime along clinical trajectories

After computing the principal tree, a root node V_{root} has to be defined by the user, accordingly to the application-specific criteria. For example, it can correspond to the node of the graph closest to a set of data points enriched with those having the least of disease severity.

The pseudotime $Pt(x)$ of an arbitrary vector x is defined as the total geodesic distance in the principal tree from V_{root} to the projection $\{p, \epsilon\} = Proj(x, G)$ of x on the graph. Algorithmically, we need to define which node of the edge E^p is the closest to the V_{root} and add the ϵ accordingly, i.e.

$$Pt(x) = \begin{cases} |V_{root} \to E^p(0)| + \epsilon, \\ \qquad \text{if } |V_{root} \to E^p(0)| < |V_{root} \to E^p(1)| \\ \\ |V_{root} \to E^p(0)| - \epsilon, \\ \qquad \text{if } |V_{root} \to E^p(0)| > |V_{root} \to E^p(1)| \end{cases} \tag{1}$$

where $|V_i \to V_j|$ signifies the number of edges (length) of the trajectory $V_i \to V_j$.

2.4 METABRIC transcriptomic breast cancer dataset

METABRIC dataset is the largest publicly available collection of transcriptomic profiles of breast tumors, accompanied by their clinical annotations, including the molecular subtype, follow up time, overall survival and relapse free status that allows application of survival analysis and identification of prognostic biomarkers [26]. METABRIC was used in 2012 DREAM Breast Cancer Challenge for predicting breast cancer patient survival using machine learning methods. Using traditional clustering approach, breast cancer transcriptomes in METABRIC can be classified into molecular subtypes called normal-like, luminal A, luminal B, HER2-enriched, basal-like and more recently introduced 'claudin-low' breast cancer subtype [4]. METABRIC dataset contains 1981 tumoral transcriptomic profiles each of which contained 24360 genes. All the samples underwent transcriptional profiling on the Illumina HT-12 v3 platform and data were normalized as described in [26].

3 Results

3.1 Extracting trajectories in the breast cancer transcriptome

For the trajectory-based analysis we used 1000 most variable genes, for which 30 principal components have been computed, so the trajectories were computed in the reduced data space.

By exploring several preprocessing schemes and parameter values for ElPiGraph, we concluded that considering 20 nodes in the principal tree, using tree pruning (reducing branches with just one node) and extending end nodes by additional edge produces biologically and clinically relevant definition of trajectories, as described below.

We identify the root node of the tree to be node #20 (see Figure 1) on the basis of the following reasoning. This node is a terminal node of the graph located in the region of the data space where data points labeled as 'Luminal A' and 'normal-like' dominate, thus it is reasonable to suggest that it corresponds to tumors which are most similar to normal cells, and so having good prognosis. This choice of the root node defines three trajectories in the omics data: trajectory #20-#19, passing through the luminal breast cancer tumors, trajectory #20-#18 leading to the basal-like breast tumor profiles, trajectory #20-#21, leading to HER2+ breast tumors (Figure 3). These trajectories are connected through two branching points, one of which corresponds to the divergence between basal-like and HER2+ molecular profiles, while the second one corresponds to the split between luminar profiles. Thus distance from the root node along one of the three trajectories serves a measure how malignant the tumor is. The further analysis confirms the reasonable choice of the root node.

3.2 Partitioning breast cancer patients accordingly to the segments of the principal tree

Having trajectory defined for the data, it automatically implies splitting data into certain clusters defined by trajectory segments. It is done in the following way: first one splits the graph into non-intersecting segments which connect branching or end nodes to other branching or end nodes (and not containing any branching points inside), second one assigns to each point of the data the nearest segment of the graph. Thus the data is split into partitions enumerated by graph segments. We refer to [18] (section "Partitioning (clustering) the data according to the principal graph segments" and Figure 9 therein) for more details. Such clusters are natural from the point of view of introduced pseudotime. Indeed, one can imagine data points representing some objects changing in time. If a branching point occurs then it is natural to think that objects after this branching point are significantly different from those before the branching (otherwise it would not be a branching point). Thus, it is natural to think of branching points as points where new clusters emerge.

Figure 1 shows data points partitioned in two ways: using the standard PAM50(+cluadin-low) definition of clusters provided with METABRIC annotation and the clusters defined by the principal graph segments. Let us argue that new clusters in many respects are at least not worse then the original ones from the clinical

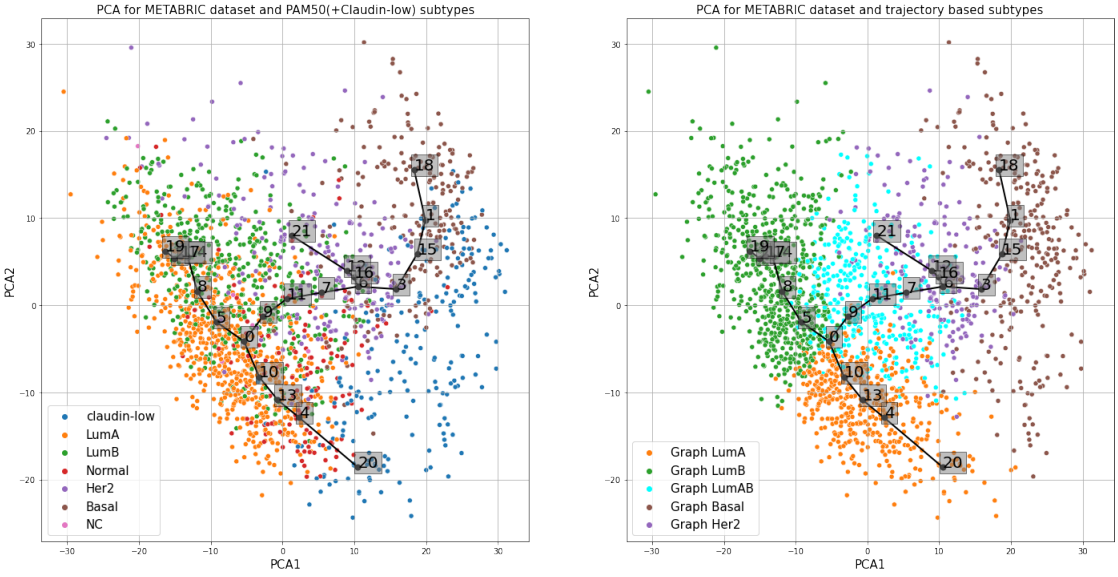

Figure 1: PCA plots for METABRIC dataset with trajectory visualisation. Colored by PAM50 subtypes **(left)** and trajectory based subtypes **(right)**.

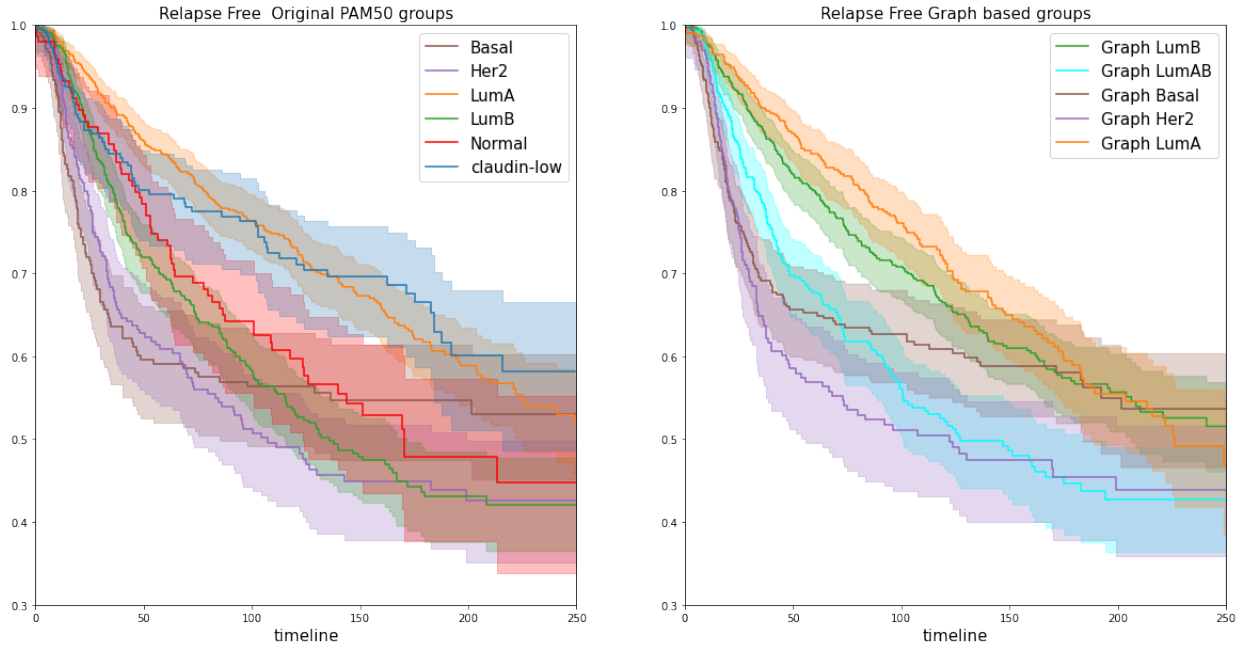

Figure 2: Kaplan-Meier survival curves for subtypes. PAM50 subtypes **(left)** and trajectory based subtypes **(right)**.

point of view, also being defined in a completely unsupervised way. Indeed, the structure of clusters is similar to the original PAM50 classification, e.g. HER2 and Basal clusters in our and PAM50 classifications strongly overlap and fit to each other. Our luminal A, B clusters are split in a way different from PAM50. We also suggest a new cluster intermediate between luminal A,B and HER2+ subtypes (see Figure 1).

To elucidate the clinical significance of the breast tumor clustering based on the principal tree, we performed Kaplan-Meier survival analysis for each cluster and compare with analysis for PAM50 clusters (see Figure 2). To be more precise we calculated the Kaplan-Meier curves for relapse free survival, that is standard practice in breast cancer analysis.

We see that the definition of principal tree-based clusters is consistent from the survival point of view - luminal A is the best, luminal B is the second (which is different from PAM50), Her2 is the worst (similar to

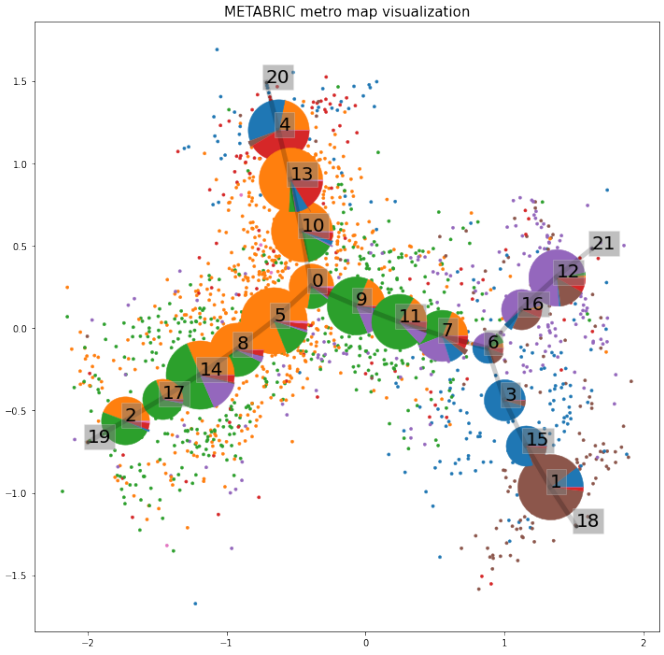

Figure 3: Principal tree metro map visualisation. Circles colored according to proportion of corresponding PAM50 subtype in neighbourhoods of the point. The colors shown in the pie-charts correspond to the ones used in Figure 1.

PAM50), our new cluster has survival intermediate between Her2 and luminal B, which corresponds to what we can expect from its position. Figure 2 also demonstrates that the principal tree segments are separating survival status at least not worse than the original PAM50 clusters, so it gives another evidence that our clusters are clinically relevant.

We also present metro map visualization of the principal graph and the dataset [32]. In this visualization, the graph is embedded into 2D using Kamada-Kawai algorithm and data points plotted according to their relative position to the nearest graph point, Figure 3). It shows how the original PAM50 clusters are recombined into neighbourhoods of nodes.

3.3 Genes associated with transcriptomic trajectories in METABRIC

Figure 4 shows the expressions of some of the selected genes along the trajectories defined by the constructed principal tree. The analysis shows that behavior of key genes correspond to known biological facts. Let us briefly summarize it. ESR1 is the gene coding one of two main types of estrogen receptor, it is well-known to be elevated for luminal tumors and suppressed in HER2/basal tumors, that agrees with the pseudotemporal expression plots. FOXA1 acts as a pioneer factor for ERa in ERa+ breast cancer, it is well-known to be highly correlated with ESR1 gene (for METABRIC data correlation is 0.72), so we see that its expression plot is reasonably similar to ESR1 plot. ERBB2 gene is a known marker for HER2+ breast tumors. One can see it is most highly expressed at the end of the trajectory 20-21 which ends at HER2 region. CDC20 is one of the key cell cycle genes and thus typically elevated for cancer cells which proliferate a lot, high expression typically correspond to poorer outcomes. So we can conclude that behaviour of key genes along trajectories is biologically relevant.

3.4 Properties of pseudotime as a prognostic biomarker in the standard survival analysis

Pseudotime can be quantified along each trajectory, and provides a measure how far the data point is from the root node, which presumably represents the least advanced stage of the tumorigenesis. In our situation we provide a biological interpretation of growing pseudotime as progression towards more malignant tumor phenotype (at least for the first and the third trajectories). According to that interpretation one might

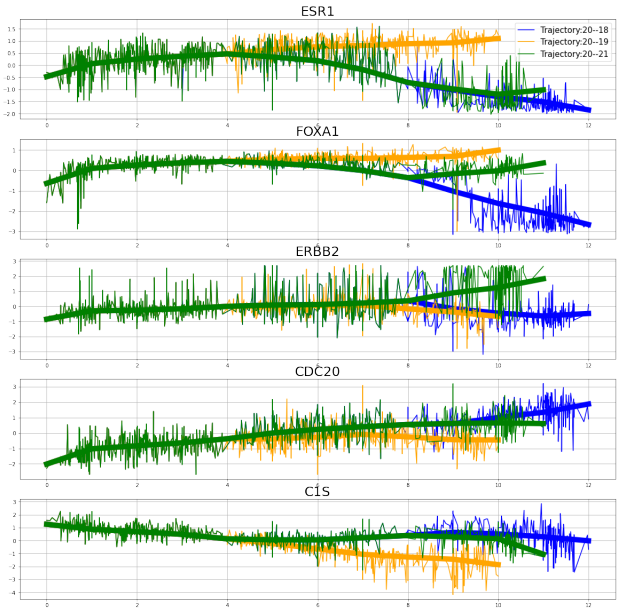

Figure 4: Visualization of several selected gene expressions along the clinical trajectories as a function of pseudotime.

expect that pseudotime would be a prognostic factor for these trajectories. To confirm that proposal we construct Kaplan-Meier curves for the groups of patients splitted by pseudotime, see Figure 5. Indeed, one can conclude that the pseudotime-based tumor groups correspond to different survival curves for the corresponding trajectories. Moreover, we compare pseudotime performance as predictor with performance of the best genes used for prognostic purposes. As one can see from the same Figure, pseudotime provides comparable performance. We provide picture only for CDC20 gene, but the results for other highly prognostic genes are similar.

In order to better characterize the predictive power of pseudotime compared with individual gene-based predictors, we determined 38 genes whose p-value in the univariate Cox survival regression was smaller than the one of pseudotime. As expected, approximately half of these genes (including CDC20, UBE2C, CCNB2) were related to cell cycle molecular mechanism or regulation of proliferation, and many indeed represented well-known breat cancer survival prognosis markers (such as ESR1, PGR, FGD3, SUSD3). We asked if the measure of pseudotime was redundant with respect to these known prognostic factors, and constructed multivariate Cox regression using these 38 gene expression levels and the value of pseudo-time as co-variates. We found that for one of the trajectories, #20-#21 leading to HER2+ tumoral profiles, pseudotime remained as a significant predictive factor, selected among the first three best predictive features. In several other computational experiments we observed that including pseudotime as a feature in the logistic regression predicting absence or presence of relapse in the 5 years following the cancer treatment systematically improved the area under the curve (AUC) value. As a result, we have concluded that the value of pseudotime can complement the traditional gene expression-based linear predictors of breast cancer survival.

To provide yet another demonstration for the performance of pseudotime as a prognostic biomarker, we analysed the behaviour of an average relapse free time along each trajectory, Figure 6. The plots are splitted into two groups: red points and mean value curves correspond to those patients with relapse occurred while blue points and mean curves correspond to those where relapse was not observed.

We can see that red curves for the first and the third trajectories are decreasing, thus average relapse free time decreases along the first and the third trajectories, which confirms our interpretation of pseudotime for these trajectories: higher pseudotime corresponds to more malignant tumors. From these plots one may also see that density of red points is drastically higher for relapse free time less than 60 months especially near the end of trajectory 20-18, that corresponds to the known fact: the risk of relapse is much higher for the first 3–5 years, but afterwards drops sharply and substantially for the basal subtype of breast cancer.

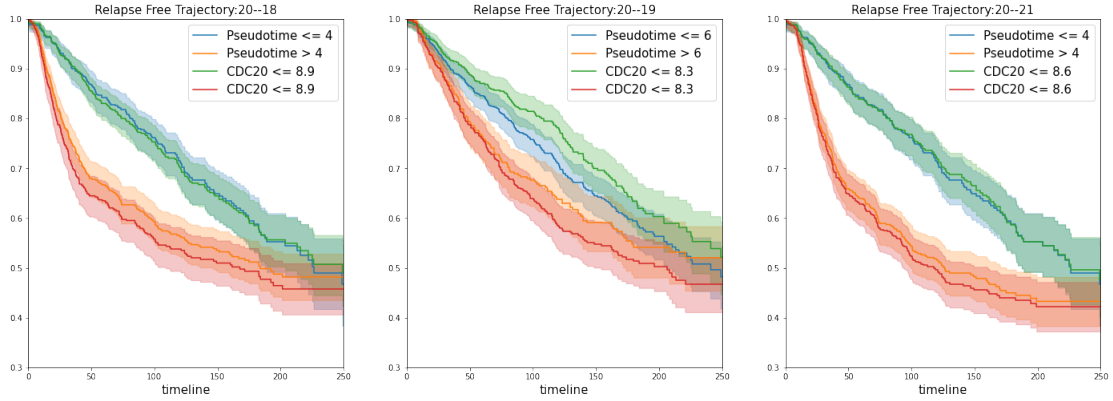

Figure 5: Kaplan-Meier curves for groups of patients. Red and orange colored lines correspond to split by pseutotime. Green and red colored lines correspond to split by the cell cycle 'CDC20' gene. One can see the pseudotime is good prognostics factor the first and the third group, however it is not more effective than 'CDC20' gene (similar results holds for several other specially selected genes).

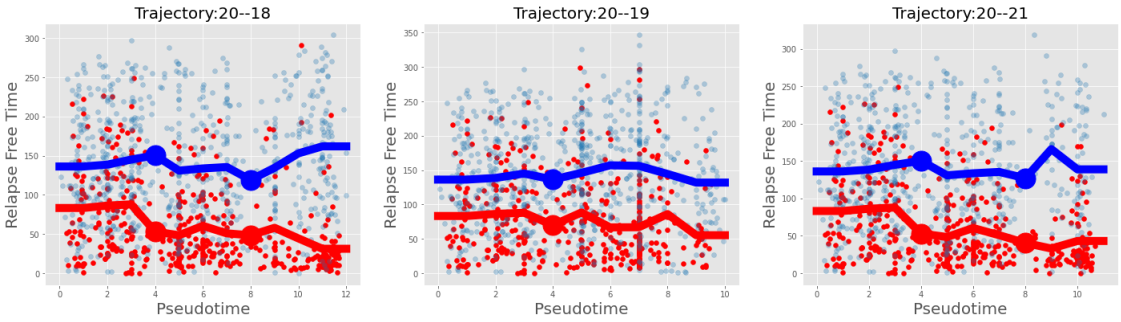

Figure 6: Average relapse free time along trajectories, red curves - patients with relapse occurred, blue curves - relapse was not observed. For the first and the third trajectories red curves decreases which fully correspond to the intuition that: further from the root, more malignant is the tumor.

4 Discussion

To conclude, we demonstrated that trajectory-based approach makes sense for the analysis of bulk transcriptomic tumoral data. It provides a clinically meaningful alternative to the standard clustering analysis and also provides more "contiunous" representation of the molecular changes in the tumors from the tumors closest to normal tissue samples (presumably least advanced) to the most malignant tumoral phenotypes. Indeed, we demonstrated that clusters defined by the trajectory-based analysis in a completely unsupervised manner are overlapping and not worse than the standard molecular subtypes defined by PAM50 classification. However, the trajectories and pseudotime quantified from them provides a measure to describe gradual change from one cluster to another, rather than a discrete picture of molecular changes represented by cluster labels, which does not correspond to the geometry of the data point cloud where the clusters are not clearly separated. We have demonstrated that for two out of three trajectories pseudotime has clear biological interpretation: it corresponds to the degree of how malignant the tumor is. For the remaining trajectory, the pseudotime summarizes important heterogeneity within luminal breast tumors which, however, is not associated with poorer prognosis.

There exists two possible interpretations of the nature of the clinical trajectories extracted from the bulk omics profiles. The pragmatic one simply suggests that the trajectories and pseudotime define the metrics of deviation of a transcriptomic profile from the one of the matching normal tissue, taking into account the geometry of the low-dimensional data manifold. More challenging interpretation consists in claiming that the extracted trajectories do reflect the biological processes underlying tumorigenesis, such that a tumor characterized by a larger value of the pseudotime corresponds to a more advanced observed stage in the tumor development. Indeed, transformation between cancer subtypes in the process of tumorigenesis seems to be a feasible scenario [24,33]. On the other hand, molecular subtypes can be connected to different cells of origin:

in this case, between-type transformation is more difficult to justify. We suggest that the interpretation of the trajectories as representations of accumulation of malignant molecular changes for a hypothetical tumor is valid at least in some cases: for example, for the case of transformation between luminal breast cancer subtypes.

We also note that the trajectory-based model of breast cancer transcriptome suggested here is significantly different from the one previously introduced in [24]. The previously suggested model was characterized by only one major branching between HER2+ and Basal-like subtypes, with other minor branchings representing less important divergencies in the transcriptomic profiles. Therefore, the previous model basically defined two successions of states: Normal-like → Luminal-A → Luminal-B → HER2+ and Normal-like → Luminal-A → Luminal-B → Basal-like. By constrast, in our study, we suggest that there exists a particular set of luminal breast cancer tumors (denoted as "Graph LumAB" in Figure 1) diverging relatively early from the main trend of expression changes along the luminal tumors and seeding the transition to more aggressive subtypes. Indeed, we can see from Figure 2, right panel, that this group of tumors is characterized by significantly poorer relapse-free survival with increase of relapse probability appearing very early in pseudotime (Figure 6). Therefore, the trajectory-based model of breast cancer transcriptome suggested here can be abstracted as existence of three paths: Normal-like → Luminal-A → Luminal-B (good prognosis), Normal-like → Luminal-A → LumA/B → HER2+ (bad prognosis) and Normal-like → Luminal-A → LumA/B → Basal-like (bad prognosis). The discrepancy between two suggested models can be in part explained by that in our case the method of principal trees has been applied in pure unsupervised setting, while in [24] a selection of features based on survival analysis have been performed prior to the trajectory reconstruction, using an ad-hoc approach, based on computation of local first principal components.

From the general machine learning perspective for supervised (prognostic) tasks (e.g. survival analysis) both clusters and trajectories can be thought as a kind of engineered features. Both cluster labels and pseudotime provide a non-linear combination of original features into a new prognostic factor (new feature). We have argued that pseudotime is a good prognostic factor on the one hand, and has a clear biological interpretation on the other hand. Even though the performance of pseudotime in terms of relapse prediction in METABRIC dataset is comparable to several individual gene-based predictors, we can suggest that it can better generalize to independent datasets (of course, this has to be demonstrated). Indeed, pseudotime value is based on combination of expression of many individual genes into a feature reflecting the geometrical structure of the data point cloud which can have better reproducibility than the connection of any individual gene with clinical annotation. Pseudotime is constructed in a pure unsupervised fashion and is free from the usual multiple testing problematics.

Summarising biological and clinical insights from the trajectory-based analysis performed in the present work, we argue that key genes expressions along the trajectories have clear biological interpretation which agrees with known facts. Survival analysis also confirms the clinical relevance of trajectories and pseudotime. First part of our analysis shows that clusters defined by trajectories gives good separation by survival status, the second part shows that pseudotime is a good prognostic factor for survival analysis, and the third part shows reasonable behavior of the average survival time along trajectories.

The present study provides arguments to a claim that trajectory-based analysis of omics datasets can be a general tool which can be applied together with a wide class of machine learning tasks, improving omics data analysis in many respects, in particular its interpretability and explainability.

References

[1] F. Adams *et al.*, *The genuine works of Hippocrates.* Sydenham society, 1849, vol. 17.

[2] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, and S. Brunak, "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients," *Nature Communications*, vol. 5, no. 1, pp. 1–10, 2014.

[3] G. Moulis, M. Lapeyre-Mestre, A. Palmaro, G. Pugnet, J. L. Montastruc, and L. Sailler, "French health insurance databases: What interest for medical research?" *Rev Med Interne*, vol. 36, no. 6, pp. 411–417, 2015.

[4] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, and B. Shi, "Breast cancer intrinsic subtype classification, clinical use and future trends," *Am J Cancer Res.*, vol. 5, no. 10, pp. 2929–43, 2015.

[5] D. J. Albers, E. Tabak, A. Perotte, and G. Hripcsak, "Dynamical Phenotyping : Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations," *PLoS ONE*, vol. 9, no. 6, p. e96443, 2014.

[6] D. Ruderman, "The emergence of dynamic phenotyping," *Cell Biology and Toxicology*, vol. 33, pp. 507–509, 2017.

[7] W. Wang, B. Zhu, and X. Wang, "Dynamic phenotypes : illustrating a single-cell odyssey," *Cell Biology and Toxicology*, vol. 33, pp. 423–427, 2017.

[8] R. Xu and D. C. Wunsch, *Clustering*. IEEE Press, Piscataway, NJ, 2008.

[9] T. Jung and K. A. S. Wickrama, "An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling," *Social and Personality Psychology Compass*, vol. 2, no. 1, pp. 302–317, 2008.

[10] D. S. Nagin and C. L. Odgers, "Group-Based Trajectory Modeling in Clinical Research," *Annual Review of Clinical Psychology*, no. 6, pp. 109–138, 2010.

[11] D. Rizopoulos, "Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data," *Biometrics*, vol. 67, no. 3, pp. 819–829, 2011. [Online]. Available: https://www.jstor.org/stable/41242530

[12] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 2956–2964.

[13] P. Schulam and R. Arora, "Disease trajectory maps," in *Proceedings of the Thirtieth Conference on Neural Information Processing Systems*, 2016.

[14] A. N. Gorban, N. R. Sumner, and A. Y. Zinovyev, "Topological grammars for data approximation," *Applied Mathematics Letters*, vol. 20, no. 4, pp. 382 – 386, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893965906001856

[15] A. N. Gorban and A. Zinovyev, "Principal manifolds and graphs in practice: from molecular biology to dynamical systems." *International Journal of Neural Systems*, vol. 20, no. 3, pp. 219–232, 2010. [Online]. Available: http://arxiv.org/abs/1001.1122

[16] L. Albergante, E. Mirkes, J. Bac, H. Chen, A. Martin, L. Faure, E. Barillot, L. Pinello, A. Gorban, and A. Zinovyev, "Robust and scalable learning of complex intrinsic dataset geometry via ElPiGraph," *Entropy*, vol. 22, no. 3, p. 296, 2020.

[17] S. E. Golovenkin, J. Bac, A. Chervov, E. Mirkes, Y. Orlova, E. Barillot, A. Gorban, and A. Zinovyev, "Supporting data for "Trajectories, bifurcations and pseudotime in large clinical datasets: applications to myocardial infarction and diabetes data"," 2020. [Online]. Available: http://dx.doi.org/10.5524/100819

[18] S. E. Golovenkin, J. Bac, A. Chervov, E. M. Mirkes, Y. V. Orlova, E. Barillot, A. N. Gorban, and A. Zinovyev, "Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data," *GigaScience*, vol. 9, no. 11, 11 2020, giaa128. [Online]. Available: https://doi.org/10.1093/gigascience/giaa128

[19] L. Martignetti, L. Calzone, E. Bonnet, E. Barillot, and A. Zinovyev, "Roma: Representation and quantification of module activity from target expression data," *Frontiers in Genetics*, vol. 7, p. 18, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2016.00018

[20] Y. Drier, M. Sheffer, and E. Domany, "Pathway-based personalized analysis of cancer," *Proceedings of the National Academy of Sciences*, vol. 110, no. 16, pp. 6388–6393, 2013. [Online]. Available: https://www.pnas.org/content/110/16/6388

[21] M. Nicolau, A. J. Levine, and G. Carlsson, "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 7265–7270, 2011. [Online]. Available: https://www.pnas.org/content/108/17/7265

[22] H. T and S. W, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[23] A. Livshits, A. Git, G. Fuks, C. Caldas, and E. Domany, "Pathway-based personalized analysis of breast cancer expression data," *Molecular Oncology*, vol. 9, no. 7, pp. 1471 – 1483, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574789115000940

[24] Y. J. N. N. G. S. Sun, Y., "Cancer progression modeling using static sample data," *Genome Biology*, vol. 15, no. 11, p. 440, 2014.

[25] Q. Mao, L. Wang, I. W. Tsang, and Y. Sun, "Principal graph and structure learning based on reversed graph embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2227–2241, 2017.

[26] C. Curtis, S. Shah, S.-F. Chin, G. Turashvili, O. Rueda, M. Dunning, D. Speed, and A. e. a. Lynch, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, p. 346–352, 2012.

[27] A. N. Gorban and A. Y. Zinovyev, "Principal Graphs and Manifolds," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques (ed. E.Olivas)*. Information Science Reference, Hershey, PA, 2008. [Online]. Available: http://arxiv.org/abs/0809.0490{%}0Ahttp://dx.doi.org/10.4018/978-1-60566-766-9

[28] H. Chen, L. Albergante, J. Y. Hsu, C. A. Lareau, G. Lo Bosco, J. Guan, S. Zhou, A. N. Gorban, D. E. Bauer, M. J. Aryee, D. M. Langenau, A. Zinovyev, J. D. Buenrostro, G. C. Yuan, and L. Pinello, "Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM," *Nature Communications*, vol. 10, no. 1903, 2019.

[29] R. G. Parra, N. Papadopoulos, L. Ahumada-Arranz, J. E. Kholtei, N. Mottelson, Y. Horokhovsky, B. Treutlein, and J. Soeding, "Reconstructing complex lineage trees from scRNA-seq data using MERLoT," *Nucleic Acids Research*, vol. 47, no. 17, pp. 8961–8974, 2019. [Online]. Available: https://academic.oup.com/nar/article/47/17/8961/5552070

[30] A. N. Gorban, E. Mirkes, and A. Y. Zinovyev, "Robust principal graphs for data approximation," *Archives of Data Science*, vol. 2, no. 1, p. 1:16, 2017.

[31] A. Chervov, J. Bac, and A. Zinovyev, "Minimum spanning vs. principal trees for structured approximations of multi-dimensional datasets," *Entropy*, vol. 22, no. 11, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/11/1274

[32] A. N. Gorban, N. R. Sumner, and A. Y. Zinovyev, "Beyond the concept of manifolds: Principal trees, metro maps, and elastic cubic complexes," in *Principal manifolds for data visualization and dimension reduction (eds. Gorban A.N, Kegl, B., Wunsch D., Zinovyev A.)*. Lecture Notes in Computational Science and Engineering, Springer, 2008, pp. 219–237.

[33] L. Cantini, G. Bertoli, C. Cava, T. Dubois, A. Zinovyev, M. Caselle, I. Castiglioni, E. Barillot, and L. Martignetti, "Identification of microRNA clusters cooperatively acting on epithelial to mesenchymal transition in triple negative breast cancer," *Nucleic Acids Research*, vol. 47, no. 5, pp. 2205–2215, mar 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30657980/