SCIENTIFIC REPORTS
natureresearch

OPEN

Molecular phenotyping using networks, diffusion, and topology: soft tissue sarcoma

Received: 27 November 2018
Accepted: 6 September 2019
Published online: 27 September 2019

James C. Mathews[1], Maryam Pouryahya[1], Caroline Moosmüller[2], Yannis G. Kevrekidis[2], Joseph O. Deasy[1] & Allen Tannenbaum[3]

Many biological datasets are high-dimensional yet manifest an underlying order. In this paper, we describe an unsupervised data analysis methodology that operates in the setting of a multivariate dataset and a network which expresses influence between the variables of the given set. The technique involves network geometry employing the Wasserstein distance, global spectral analysis in the form of diffusion maps, and topological data analysis using the Mapper algorithm. The prototypical application is to gene expression profiles obtained from RNA-Seq experiments on a collection of tissue samples, considering only genes whose protein products participate in a known pathway or network of interest. Employing the technique, we discern several coherent states or signatures displayed by the gene expression profiles of the sarcomas in the Cancer Genome Atlas along the TP53 (p53) signaling network. The signatures substantially recover the leiomyosarcoma, dedifferentiated liposarcoma (DDLPS), and synovial sarcoma histological subtype diagnoses, and they also include a new signature defined by activation and inactivation of about a dozen genes, including activation of serine endopeptidase inhibitor *SERPINE1* and inactivation of TP53-family tumor suppressor gene *TP73*.

Modern biological investigations often result in dense, high-dimensional datasets describing genes, proteins, mutations, or other variables. A near universal problem arises as the dimensionality of the data grows: how can the data be investigated in a relatively unbiased manner, to expose underlying clusters and relational structure? To date, there is a lack of robust techniques for exposing the structure of biological data in an unbiased, agnostic fashion. Biological relevance is maintained in this work by considering known pathways as an underlying guide. The technique we describe involves network geometry via the Wasserstein distance[1,2], global spectral analysis in the form of diffusion maps[3], and topological data analysis using the Mapper algorithm[4]. We apply the technique to gene expression profiles along gene sets participating in known pathways. We discern several coherent states or signatures displayed by the gene expression profiles of The Cancer Genome Atlas (TCGA) sarcoma project along the TP53 signaling network. The signatures substantially recover the leiomyosarcoma, dedifferentiated liposarcoma (DDLPS), and synovial sarcoma histological subtype diagnoses, and they also include a new signature defined by activation and inactivation of about a dozen genes, including activation of serine endopeptidase inhibitor *SERPINE1* and inactivation of TP53-family tumor suppressor gene *TP73*.

The mechanisms that intervene between DNA sequence genotype and overall cell phenotype are complex, including the presence of transcription factors, the chemistry of the cell microenvironment, and epigenetic factors like phosphorylation and methylation. We focus on the determination of transcriptomic molecular phenotypes. The simplest molecular phenotypes are defined by single marker genes, like the estrogen receptor (ER), progesterone receptor (PR), or human epidermal growth factor receptor 2 (HER2/ERBB2) status of breast carcinomas. In general a comparatively large number of genes must be considered simultaneously.

Methods falling under the heading of Genome Wide Association Studies (GWAS, typically concerning mutational profiles) or Gene Set Enrichment Analysis (GSEA, typically concerning gene expression profiles) take into account data concerning a large number of genes to ascertain statistical significance with respect to given known outcomes or endpoints such as disease states. In general, they do not attempt to discern coherent states in gene expression quantification profiles in an "unsupervised" manner. That is, these methods do not ascertain existing

[1]Department of Medical Physics, Memorial Sloan-Kettering Cancer Center, New York, USA. [2]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, USA. [3]Departments of Computer Science and Applied Mathematics & Statistics, Stony Brook University, Stony Brook, USA. Correspondence and requests for materials should be addressed to J.C.M. (email: mathewj2@mskcc.org)

apparent molecular phenotypes, but rather impose or design molecular phenotypes specifically to serve as predictors for variables of ultimate interest like prognosis.

Seemann, Shulman, and Gunaratne[5] employ degree 0 persistent homology towards the end of unsupervised analysis, leveraging the robustness of Topological Data Analysis (TDA) techniques for unsupervised clustering. One could also use various established unsupervised clustering algorithms such as hierarchical clustering or *k*-mean optimization methods, optionally preceded by dimensional reduction techniques like Principal Component Analysis (PCA), *t*-distributed Stochastic Neighbor Embedding (t-SNE), or Multi-Dimensional Scaling (MDS). Note, however, that hierarchical clustering has the drawback that the output of the algorithm strongly underdetermines the usual heat map visual representation. Every branch of the hierarchy tree creates an ambiguity in the order in which the samples are displayed.

From the topological point of view, however, any method within the clustering paradigm is order 0 in the sense that it summarizes a dataset in terms of a finite/discrete set of disjoint categories, a "space" of dimension 0. Lockwood and Krishnamoorthy[6] advocate "higher order" methods, e.g. degree 1 persistent homology, extracting 1-dimensional features in the space charted by the data points, roughly in order to take account of the relations between categories and not just the categories themselves. One major difficulty with this approach is that homology classes are defined by *cycles*; topological features which are not cycles, such as *relative cycles* or branches, are not detectable with existing tools (see[7] for general background on topology, or[8] for background on TDA methods). A second major difficulty is that homology classes do not have canonical representative cycles. This means that in theory an almost arbitrary subset of the points of a point cloud can appear along the path of a cycle belonging to an observed persistent 1-homology class. In other words, while persistent homologies are certainly evidence of important dataset-specific global features, there is an unsolved problem of interpretability of such features.

Camara, Emmett, and Rabadan[9] calculate persistent 1-homologies in evolutionary/phylogenetic data, surmounting both of these difficulties simultaneously by interpreting the presence of non-trivial cycles (closed loops), and not the internal structure of their representative cycles *per se*, as an indication of the presence of genetic recombination events.

We largely follow Nicolau *et al.*[4] in that we use the Mapper algorithm to map our point clouds onto summary spaces of dimension 1, graphs or networks. This algorithm can be regarded as a discrete version of the Morse-theoretic analysis of a smooth manifold with respect to a height function (called the filter function). Nicolau *et al.* heavily de-sparsify the point clouds, in order to avoid the normal preprocessing step of dimensional reduction (virtually always required for biological datasets), and employ a carefully designed deviation-from-normal filter function in accordance with what they call the Progression Analysis of Disease paradigm. We take a slightly different tack: First, we perform a biologically-motivated intermediate-scale dimensional reduction by considering only those genes participating in well-known pathways (we use the Kyoto Encyclopedia of Genes and Genomes). Next, we replace the ordinary Euclidean distance metric between gene expression profiles with alternative metrics, especially a version of the Wasserstein 1-metric which takes account of curated knowledge of the network structure linking the genes (coordinates). Then we employ the dimensional reduction and analysis technique of diffusion maps[3] to regularize the point cloud with respect to intrinsic or characteristic global geometry. We have found that this process results in datasets with favorable properties for the application of Mapper and the interpretation of its resultant graph summaries.

Methods

We take as our primary input a gene expression quantification sample set, as a point cloud $S \subset \mathbb{R}^N$, and an influence, regulation, or pathway network G relating the N genes which label the coordinates. Optionally, we include an additional control dataset $C \subset \mathbb{R}^N$, or a function $f: S \to \mathbb{R}$ with the interpretation as an experimentally-determined "degree of progression" with respect to some process (e.g., a disease process).

The output is a list of coherent states or molecular phenotypes, characterized by activated, inactivated, and equivocally-activated genes. We now enumerate the steps of our pipeline. Details will be given in subsequent sections.

1. **Normalize the values of *S***
2. **Restrict/project *S* to the genes appearing in the gene network *G***
3. **Calculate a network-based distance metric between samples**
4. **Evaluate a diffusion map**
5. **Perform the Mapper algorithm on the re-mapped $S \subset \mathbb{R}^M$**
6. **Extract and process the state graph**
7. **Plot heat maps and discern coherent states**

Normalize the values of *S*.

We must ensure that the values of S represent gene expression quantification, for example, FPKM (Fragments Per Kilobase Million) or TPM (Transcripts Per Kilobase) values as a result of a high-throughput sequencing pipeline. These values correspond roughly to the concentration of RNA transcripts in the tissue samples, typically across many cells for each sample (bulk sequencing) though sometimes for single cells.

Optionally, for comparison between genes and for the purposes of image-rendering, for each of the N genes, we replace the values of S in the corresponding coordinate by a truncated translated z-score, $x \mapsto (x - \mu)/(3\sigma) + 0.5$, where μ and σ are the mean and standard deviation of the values for this coordinate

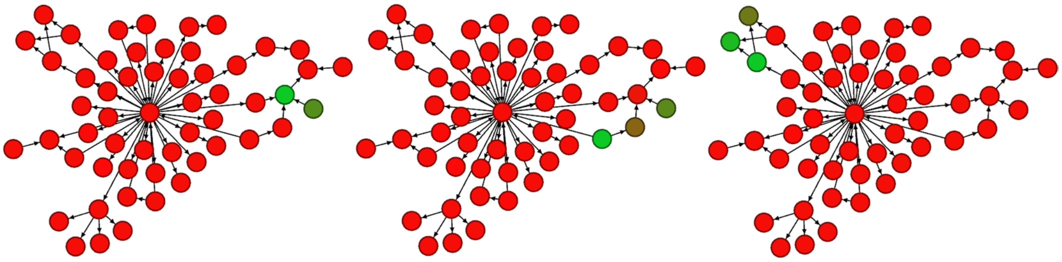

Figure 1. Left to right, distributions s_1, s_2, s_3, for illustration. Red represents values close to zero, and green more positive. The Wasserstein 1-distance $d(s_1, s_2) = 1.10$ is much less than $d(s_2, s_3) = 3.08$, while the corresponding Euclidean distances are approximately equal to each other.

and x is a typical value of this coordinate. The resulting values will be substantially normalized to lie on a scale from 0 to 1 with mean 0.5.

Restrict/project *S* to the genes appearing in the gene network *G*. We select a network G whose nodes correspond to genes whose presence or absence constitutes participation in a coordinated function or process of interest, and whose edges represent the coordinating relationships. Omit the expression values for genes not participating in G. The networks we consider are the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways concerning cell cycle regulation, senescence, proliferation, apoptosis, and TP53 signaling.

Calculate a network-based distance metric between samples. For each sample $s \in S$, define a probability distribution p_s on the set of nodes of G by interpreting the values of s (divided by their sum) as a probability density function. Alternatively, use a distribution on the nodes which is the invariant measure for a Markov chain stochastic process inferred from the values of the sample s, in the manner of[10]. It can happen that G is disconnected into several components, with no edges/links between components, in which case one should define separate distributions for each component. Note that this disconnection may be either genuine or an indication of missing biological knowledge, so that a network-connection inference method may be useful. For each component c, calculate the Wasserstein 1-metric or Earth Mover's Distance $d_c(s, s')$ between each pair p_s and $p_{s'}$ with respect to the path-length metric on c (weighted by the reciprocal of strength-related edge weights, if present). Intuitively, the Earth Mover's Distance measures the total effort needed, in the best case scenario, to displace one distribution into another, taking account of the ground point-to-point distance. This distance is illustrated for networks in Fig. 1. Technically, the Earth Mover's Distance between two mass distributions on a common metric space is defined as the infimum of the total mass-weighted displacement among displacement functions from the space to itself which map the first mass distribution onto the second. Classically it is only defined if the total masses of both distributions are equal[1]. We use the "direct sum" formula to amalgamate these distances across components c into a single distance for each pair of samples (s, s'):

$$d(s, s') = \sqrt{\sum_c d_c(s, s')^2}.$$

The Wasserstein 1-metric employed in this way is perhaps the simplest alternative to the standard Euclidean metric for which a network or pathway structure relating the coordinates is in some way taken into account. The principal benefit of this metric is that it greatly increases the distance between two samples in comparison with the Euclidean metric in case the main activity of one sample takes places in an area of the network very far from the area of main activity of the other sample. One conceivable disadvantage is that isolated changes to a given sample, say in the expression of a single gene, can have an outsized effect on the Wasserstein 1-distance of the displacement. We also caution that since the KEGG database networks are enriched with nodes for various compounds and macromolecules in addition to protein gene products, an analysis which considers only gene expressions will not take advantage of the full KEGG pathways and may have some misleading consequences. To compute the Wasserstein distance, we used the Hungarian algorithm[2].

Evaluate a diffusion map. In order to reduce the dimension and complexity of the dataset S, while preserving key information for subsequent analysis, we apply diffusion maps[3,11]. This manifold learning technique provides a global parametrization of a low-dimensional, possibly nonlinear manifold on which the high-dimensional data is assumed to lie. Such an embedding is obtained by spectral properties (eigenvalues and eigenvectors) of the graph Laplacian on a certain weighted graph with nodes S. Its eigenvectors can be used as a coordinate system on the dataset S, which is justified by the fact that they approximate the eigenvectors of the Laplace-Beltrami operator of the underlying manifold[3,12]. See also some recent work[13] for employing diffusion map techniques for data whose "points" are weighted graphs.

Following[11], for data samples $s_i, s_j \in S$ we define a connectivity matrix W using a Gaussian kernel:

$$W_{ij} = \exp\left(-\frac{d^2(s_i, s_j)}{\varepsilon}\right), \tag{1}$$

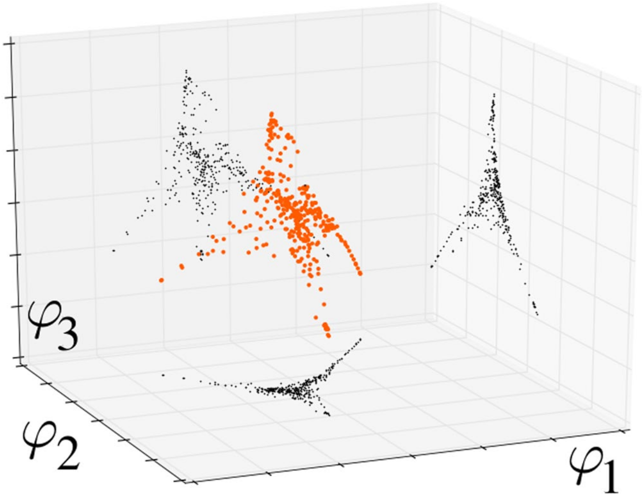

Figure 2. The diffusion re-mapped images of the gene expressions of 355 adipose visceral omentum tissue samples from the GTEx database. The first three eigenfunctions of the diffusion operator are used, to make a three-dimensional plot.

where d is the Wasserstein 1-metric as defined in[1] and ε is the kernel scale parameter. The kernel is intended to capture the features of the underlying dataset and it is therefore reasonable to choose the metric d and the scale parameter ε based on the application. The parameter ε defines a local connectivity scale, in the following sense: If s_j is in the ε-ball around s_i, the kernel induces high weight between s_i and s_j. Otherwise the weights are negligible. We can choose ε to be almost any value between the minimum and maximum among the pairwise squared distances $(d(s_i, s_j))^2$.

Define an adapted kernel

$$\widetilde{W} = D^{-1}WD^{-1} \tag{2}$$

with D the diagonal matrix $D_{ii} = \sum_{j=1}^{N} W_{ij}$, $N = \#S$. We use the adapted kernel (2) instead of (1), corresponding to the choice $\alpha = 1$ in the family of kernel normalizations presented in[3,14], to recover the Riemannian geometry of the underlying data independently of the data sampling.

We build a weighted graph with node set S and weights $\widetilde{W}_{ij}$ of the edge connecting s_i to s_j. Now apply weighted graph Laplacian normalization to $\widetilde{W}$

$$\widetilde{D}^{-1}\widetilde{W}, \tag{3}$$

with $\widetilde{D}$ the diagonal matrix $\widetilde{D}_{ii} = \sum_{j=1}^{N} \widetilde{W}_{ij}$. The associated graph Laplacian is given by

$$L = I - \widetilde{D}^{-1}\widetilde{W}. \tag{4}$$

We compute the eigenvalues $1 = \lambda_1 \geq |\lambda_2| \geq \ldots \geq |\lambda_N|$ and eigenvectors $\varphi_1, \ldots, \varphi_N$ of $\widetilde{D}^{-1}\widetilde{W}$. These eigenvectors provide an embedding of the data into a space of dimension $M < N$ (for example, as shown in Fig. 2):

$$\Phi(s_i) = [\varphi_1(i), \ldots, \varphi_M(i)]. \tag{5}$$

We reiterate that rather than using the Euclidean distance between samples s_i and s_j, we select the "more informed" 1-Wasserstein network metric. Unlike dimensional reduction techniques like Principal Component Analysis or Local Linear Embedding, but in common with t-SNE or MDS, the technique of diffusion maps can function on arbitrary intrinsic-metric representations of the data of the point cloud and does not require this point cloud to be presented in some Euclidean space. We prefer diffusion maps over t-SNE or MDS because as far as we know the latter are not guaranteed to recover the intrinsic manifold degrees of freedom of the original dataset, while diffusion maps are so guaranteed in principle. In practice biological datasets of present interest seem to represent processes of sufficient complexity that precise quantitative accounting for the relationships between all of the variables is rarely proposed, so such theoretical considerations are arguably premature.

We remark that dimensional reduction of gene expression data via diffusion maps is also suggested e.g. in[15], where the authors combine diffusion maps with a neural network clustering method to differentiate between different types of small round blue-cell tumors.

Perform the Mapper algorithm on the re-mapped $S \subset \mathbb{R}^M$. This algorithm results in a simplicial complex, in some sense modeling the mesoscopic-scale topology of the support space for the collection of samples or states S. It works by (1) dividing the point cloud into overlapping "slices" by binning the values of a chosen filter function $f\colon S \to \mathbb{R}$ into overlapping bins, (2) clustering the points of each slice (e.g. with single-linkage

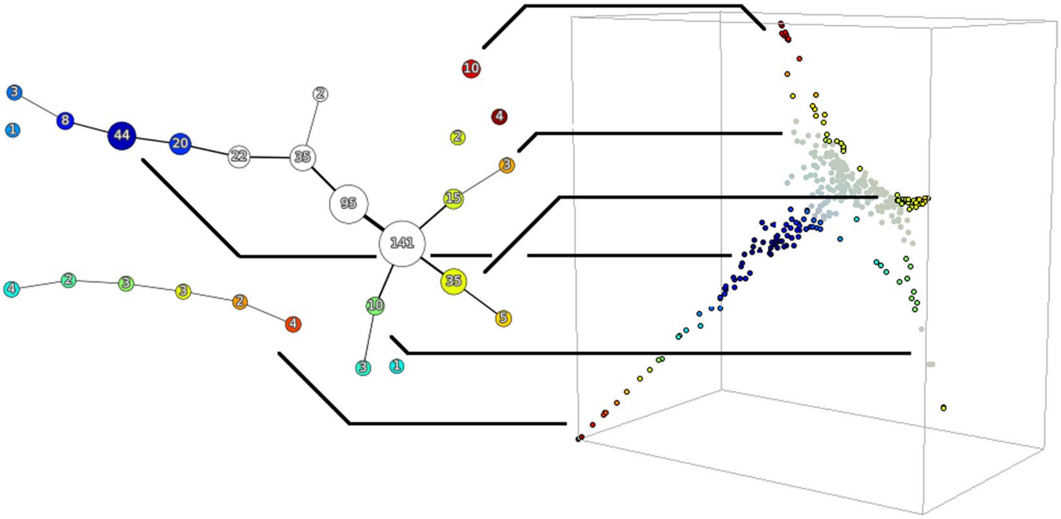

Figure 3. The Mapper state graph of the diffusion re-mapped 355-sample GTEx dataset of Fig. 2, with branches highlighted. For illustration, the "core" is not highlighted. The numerical labels indicate the number of samples in a cluster. The color indicates the value of the filter function which was used to seed the Mapper algorithm (a nearest-neighbor network closeness centrality in this case).

clustering), and (3) linking pairs (or tuples) of clusters by edges (or higher-dimensional simplices) depending on the amount of overlap between clusters.

A reasonable choice for the filter function is a deviation function devised in comparison with a control dataset C, roughly as in Nicolau *et al.*[4], e.g. the Mahalanobis distance function adapted to C in case the size of C is sufficiently large in comparison to the dimension M. We often use the general-purpose network centrality measure available in Daniel Müllner's Python Mapper implementation (http://danifold.net/mapper/).

The algorithm requires the choice of certain resolution or scale parameters: the number n_f of filter-level-set bins and a threshold t for the single-linkage clustering algorithm applied to each slice. One should select these parameter values intermediate between the extreme values which completely divide the sample set into isolated clusters and those which completely merge the sample set into a single cluster. In practice a narrow range of such values exists.

Applying the Mapper algorithm in this way is an *ad hoc* (case by case) form of ascertaining persistent topological features. *Persistence* is meant in roughly the same sense as the technique of *persistent homology*. Though persistent homology would ordinarily determine preferred values for parameters like n_f, existing persistent homology tools are seemingly inapplicable to our setting. This is because it is multi-dimensional in that the simplicial complexes of interest depend on multiple parameters n_f and t, and because a well-defined relation of containment or mapping between the complexes across parameter values is not apparent. Nevertheless appropriate values for n_f and t are normally apparent.

Extract and process the state graph.

Next, we consider the graph which is the 1-skeleton of the simplicial complex resulting from the Mapper algorithm. We decompose it into linear paths, and concatenate these paths for display. An example state graph is shown in Fig. 3.

Plot heat maps and discern coherent states.

We order the samples within each node of the state graph according to the filter values. This ordering is combined with the concatenated linear path structure for an ordering of the samples S along one or both axes of a two-dimensional plot of:

- the expression values
- the correlations with subpopulations defined by discrete covariates
- 1-Wasserstein distance matrix
- diffusion map Euclidean distance matrix

Salient coherent states may appear in the expression heat map defined by patterns of activation and inactivation of particular genes, especially near the two extreme values for the filter function. This may require approximate dichotomization of the expression values (i.e. increasing the contrast, in the terminology of image processing).

From the point of view of topological data analysis, the most interesting states are ones which are not separated by the chosen filter function alone, but are nevertheless distinguished by branching of the state graph. We caution that although the topological aspect of this pipeline has the benefit of insensitivity to dimensionality, functioning well even in very high-dimensional settings, it sometimes provides little insight beyond that already provided by the spectral analysis or diffusion map when the number of samples is small. On the other hand, Mapper seems to have the potential to function well even without diffusion maps preprocessing or gene network analysis, provided that a filter function can be chosen that brings sufficiently rich outside information into the analysis.

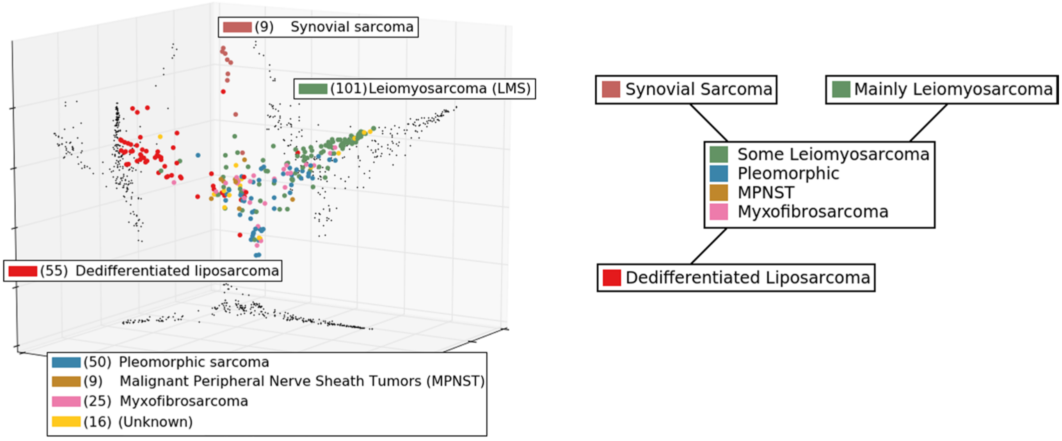

Figure 4. (Left) The diffusion re-mapped images of the 265 gene expression profiles from the TCGA sarcoma project, restricted to the TP53 signaling network defined in the KEGG database[24,25]. The first, second, and fourth eigenvectors were used. (Right) A schematic of the state graph summary produced by the Mapper algorithm.

Results

The application of the network-metric/diffusion-map/Mapper pipeline to the 265 gene expression profiles of the samples of the TCGA sarcoma project demonstrates the basic efficacy of the method. The state graph is shown in Fig. 4, and the heatmaps are shown in Fig. 5, with tissue classification.

Discussion of sarcoma states. We refer to the KEGG TP53 signaling pathway and indicate in parentheses the gene names appearing there, following official HUGO gene symbols. Italics indicate protein products.

The state PS#1 (TP53/P53 Signaling 1) consists almost entirely of samples tagged for leiomyosarcoma, meaning that tissue pathology determined a derivation from smooth muscle cells. As expected, high levels of CCNG1 (cyclin G), PPM1D (Wip1), and TP73 (P73), as well as *MDM2* are all negatively regulating *TP53*, which is not substantially activated. Arrest of the G1 and G2 phases should be frequently triggered since *CDK2*, *CCNB1* (cyclin B), and *CDK1* (CDC2), are all activated. Although *FAS*, *PIDD1* (PIDD), *PMAIP1* (NOXA), and *SIAH1* (SIAH) are substantially activated, the apoptosis pathway for which they are precursors is not, including low levels of *BAX*, the death receptor protein *TNFRSF10B* (DR5), *BID*, *CYCS* (cytochrome c), and all caspases. Apoptosis seems to have been largely evaded.

PS#3 consists almost entirely of samples tagged for dedifferentiated liposarcoma (DDLPS), and conversely almost all of the dedifferentiated liposarcomas among the 265 samples display state PS#3. We emphasize for clarity that all of the states were determined in an entirely unsupervised manner, with no input from the histological classification. *TP53* is strongly activated, and its negative regulator *MDM2* seems to be repressed by *CDKN2A* (P14ARF). A large number of the elements of the normal apoptosis signaling pathway are activated: *FAS*, *BAX*, *TNFRSF10B*, *BID*, *ZMAT3* (PAG608), and *SIAH1*. The downstream caspase *CASP3* is substantially activated. Apoptosis may occur in dedifferentiated liposarcomas at comparatively high rates. Alternatively, see[16] for a discussion of situations where normally apoptotic caspases are non-lethal to the cell. Substantial activation of *CDKN1A* (P21) is not inhibiting *CDK4*/6 or *CDK2* as expected; rather *CDK4*/6 over-expression is the most salient characteristic of state PS#3. According to Binh *et al.*[17], over-expression of *CDK4 and MDM2* is known to be a reliable diagnostic marker for well-differentiated liposarcoma (not represented in the TCGA sarcoma project).

Note that both leiomyosarcoma and DDLPS subtypes are known to exhibit complex karyotypes, with no apparent characteristic mutation. This seems to be part of the reason why they were selected for inclusion in the TCGA sarcoma project (https://cancergenome.nih.gov/cancersselected/Sarcoma). Nevertheless the coherent states PS#1 and PS#3 show that the expression profiles for these subtypes are more organized than their mutational profiles. We remark that ordinary unsupervised hierarchical clustering substantially reproduces these results, with somewhat less coherence among the apparent states.

On the other hand, synovial sarcoma is known to be well-characterized by a specific translocation resulting in gene fusion of *SYT* with either *SSX1*, *SSX2*, or *SSX4*[18]. So it is not surprising that there is a coherent state, PS#4, displayed by precisely the synovial sarcomas.

Finally, we consider the state PS#2. It does not consist mainly of any one histopathological subtype. The most obvious feature is that *TP53* and almost all of its normal positive regulation targets are inactivated, despite high levels of *CHEK1* (CHK1) potentially indicating DNA damage. Nearly all of the markers for *TP53* negative feedback regulation are strongly inactivated, including *CCNG1*, *PPM1D*, *TP73*, and *MDM2*. *CDKN1A* inactivation is consistent with the appearance of *CCND1* (cyclin D) and *CDK4*/6. With respect to the upstream elements of the apoptosis signaling pathway, we observe in state PS#2 the opposite behavior from the state PS#1, namely that *FAS*, *PIDD1*, *SIAH1*, and possibly *PMAIP1* are absent, but *BAX*, *PERP*, *TNFRSF10B*, and *BID* transcripts are all present.

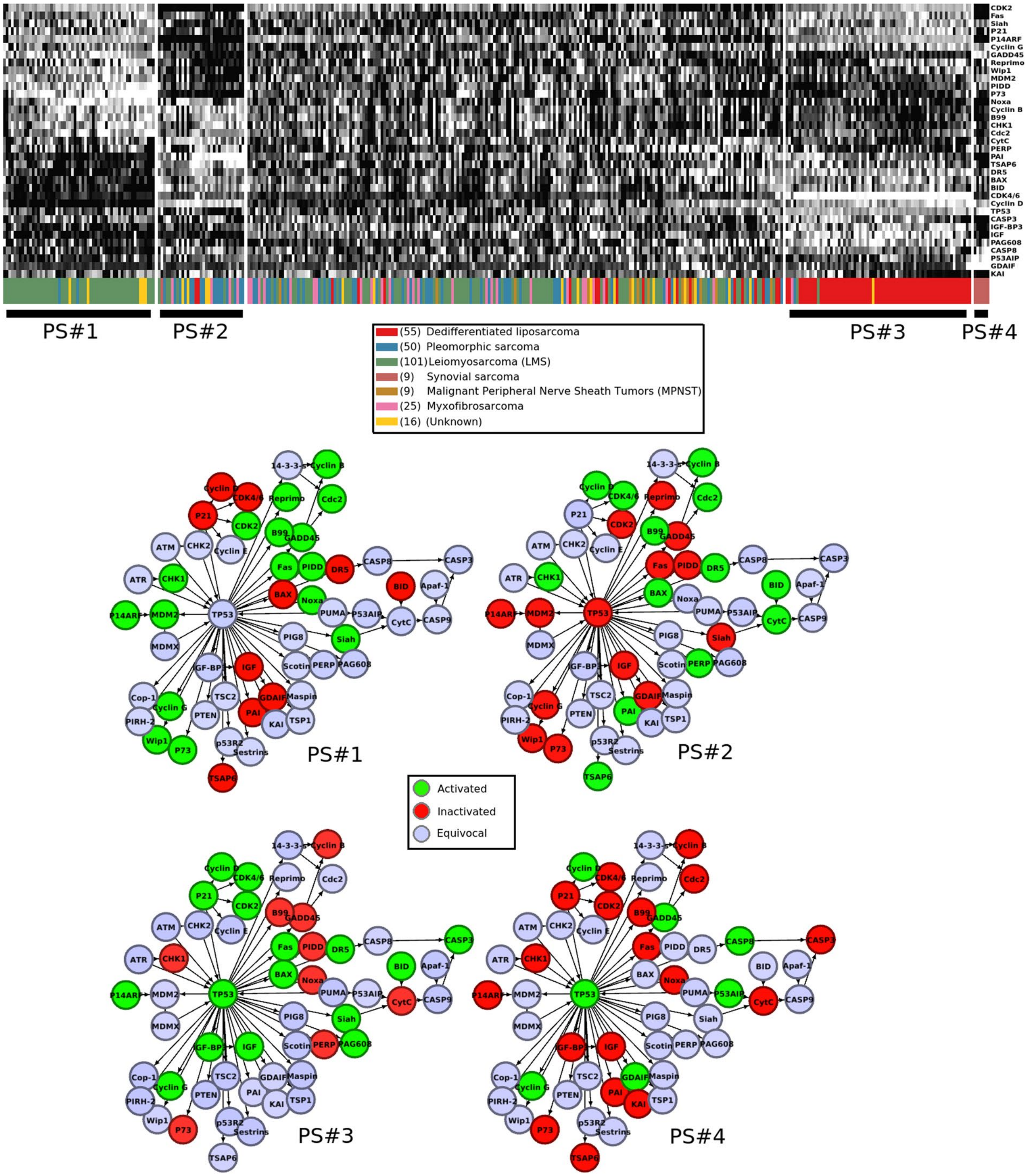

Figure 5. Four coherent states of the KEGG TP53 signaling network displayed by subsets of the TCGA sarcoma samples, shown superimposed on the network. The gene names are shown as they appear in the KEGG network.

A cBioPortal[19] query of the TCGA sarcomas displaying the *TP73-/SERPINE1+* expression pattern of PS#2 reveals high probability of loss of some portion of chromosome 18q. *SERPINB3*, which is located on 18q, and *TP73* are both associated with negative regulation of JUN kinase activity (GO:0043508) according to the human Gene Ontology Annotation database[20].

This suggests the following narrative to explain the molecular mechanisms driving the cancers in state PS#2. Damage to *SERPINB3* on chromosome 18q disrupts serine/cysteine-type endopeptidase inhibitor activity, which is then restored by *SERPINE1* upregulation by some intermediate process. Normal SERPINB3 would in addition inhibit JUN kinase activity, but this inhibitory function is *not* restored by *SERPINE1* upregulation. TP73 also normally inhibits JUN kinase activity. Unchecked JUN kinase activity may then be the main driver of tumor cell proliferation and transformation[21] in state PS#2 since effective *TP73* and *SERPINB3* both seem to be absent.

The high degree of *CCND1* activity of PS#2 is consistent with this hypothesis, since JUN induces transcription of *CCND1*[22].

Note that a proto-oncogenic role for JUN has long been suspected, and its actual function is complex, including alternately pro- and anti-tumor behaviors depending on context[23]. In this specific case, our finding answers the call of Messoussi *et al.*[23] to delineate patients that would potentially benefit from JNK (c-Jun N-terminal kinase) inhibitors. In approximately 10% of soft-tissue sarcomas, largely irrespective of histological subtype and possibly independent of JUN amplification status, JNK inhibitors that can replace the inhibitory function no longer provided by *SERPINB3* may restore JNK activity to normal condition.

Conclusion and future research directions. The network-metric/diffusion-map/Mapper pipeline uncovered some latent features of high-dimensional genomic sarcoma data in a relatively robust way. One promising direction for future research is the inference of phylogenetic trees via mutational data like Single Nucleotide Polymorphism (SNP) calls or gene amplifications and deletions in the evolutionary context. This context could be spatially dense single-tumor samples or single-patient metastasis or micrometastasis samples. Mapper would be especially adapted to the elucidation of branching/inheritance structures when the filter function is a suitable quantification of the deviation of a sample from a founder population. For example, a Hamming-type distance in the case of SNP sequences, which could also be used for the intrinsic metric between SNP sequences. As an alternative to the naive Hamming distance, a network-enriched distance could be obtained by means of linkage disequilibrium calculations.

References

1. Rachev, S. T. & Rüschendorf, L. *Mass transportation problems. Vol. II*. Probability and its Applications (New York). Applications (Springer-Verlag, New York, 1998).
2. Rubner, Y., Tomasi, C. & Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* **40**, 99–121 (2000).
3. Coifman, R. R. & Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis* **21**, 5–30, Special Issue: Diffusion Maps and Wavelets (2006).
4. Nicolau, M., Levine, A. J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. USA* **108**, 7265–7270 (2011).
5. Seemann, L., Shulman, J. & Gunaratne, G. H. A robust topology-based algorithm for gene expression profiling. *ISRN Bioinform* **2012**, 381023 (2012).
6. Lockwood, S. & Krishnamoorthy, B. Topological features in cancer gene expression data. http://arxiv.org/abs/1410.3198v1 (2014).
7. Munkres, J. R. *Topology: a first course* (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975).
8. Edelsbrunner, H. & Harer, J. L. *Computational topology*, An introduction (American Mathematical Society, Providence, RI, 2010).
9. Camara, P. G., Rosenbloom, D. I., Emmett, K. J., Levine, A. J. & Rabadan, R. Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination. *Cell Syst* **3**, 83–94 (2016).
10. Chen, Y. *et al.* Pediatric Sarcoma Data Forms a Unique Cluster Measured via the Earth Mover's Distance. *Sci Rep* **7**, 7035 (2017).
11. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS* **102**, 7426–7431, https://doi.org/10.1073/pnas.0500334102 (2005).
12. Jones, P. W., Maggioni, M. & Schul, R. Manifold parametrizations by eigenfunctions of the laplacian and heat kernels. *Proceedings of the National Academy of Sciences* **105**, 1803–1808 (2008).
13. Rajendran, K., Kattis, A., Holiday, A., Kondor, R. & Kevrekidis, I. G. Data mining when each data point is a network, https://arxiv.org/abs/1612.02908 (2016).
14. Nadler, B., Lafon, S., Coifman, R. R. & Kevrekidis, I. G. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis* **21**, 113–127, Special Issue: Diffusion Maps and Wavelets, https://doi.org/10.1016/j.acha.2005.07.004 (2006).
15. Xu, R., Damelin, S., Nadler, B. & Wunsch, D. C. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. *Artificial Intelligence in Medicine* **48**, 91–98, https://doi.org/10.1016/j.artmed.2009.06.001, Artificial Intelligence in Biomedical Engineering and Informatics (2010).
16. Garrido, C. & Kroemer, G. Life's smile, death's grin: Vital functions of apoptosis-executing proteins. *Current opinion in cell biology* **16**, 639–46 (2005).
17. Binh, M. B. *et al.* MDM2 and CDK4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: a comparative analysis of 559 soft tissue neoplasms with genetic data. *Am. J. Surg. Pathol.* **29**, 1340–1347 (2005).
18. Mendelsohn, J., Howley, P. M., Israel, M. A., Gray, J. W. & Thompson, C. B. *Molecular Basis of Cancer*, 4 edn. (Elsevier, 2015)
19. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
20. Huntley, R. P. *et al.* The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.* **43**, D1057–1063 (2015).
21. Johnson, G. L. & Nakamura, K. The c-jun kinase/stress-activated pathway: regulation, function and role in human disease. *Biochim. Biophys. Acta* **1773**, 1341–1348 (2007).
22. Wisdom, R., Johnson, R. S. & Moore, C. c-Jun regulates cell cycle progression and apoptosis by distinct mechanisms. *EMBO J.* **18**, 188–197 (1999).
23. Messoussi, A. *et al.* Recent progress in the design, study, and development of c-Jun N-terminal kinase inhibitors as anticancer agents. *Chem. Biol.* **21**, 1433–1443 (2014).
24. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
25. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

Acknowledgements

We thank Daniel Müllner for making his Python Mapper implementation so readily available (http://danifold.net/mapper/). We also thank Juan M. Bello-Rivas for the use of his diffusion maps software (https://github.com/jmbr/diffusion-maps). C.M. thanks the Department of Chemical and Biological Engineering, Princeton University, for their hospitality. This project was supported by AFOSR grant FA9550-17-1-0435), RO grant (W911NF-17-1-049), grants from National Institutes of Health (1U24CA18092401A1, R01-AG048769), MSK Cancer Center Support Grant/Core Grant (P30 CA008748), and a grant from Breast Cancer Research Foundation. I.G.K. and C.M. thank the National Science Foundation for partial support.

Author Contributions

J.C.M., M.P., C.M., Y.G.K., J.O.D. and A.T. designed research; J.C.M. performed research; J.C.M. analyzed data; J.C.M., M.P. and C.M. wrote the paper; Y.G.K., J.O.D., and A.T. edited the paper; Y.G.K., J.O.D. and A.T. directed research.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.