TOPOLOGICAL FEATURES IN CANCER GENE EXPRESSION DATA

S. LOCKWOOD

School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164, U.S.A.
E-mail: svetlana.lockwood@email.wsu.edu

B. KRISHNAMOORTHY

Department of Mathematics, Washington State University, Pullman, WA 99164, U.S.A.
E-mail: bkrishna@math.wsu.edu

We present a new method for exploring cancer gene expression data based on tools from algebraic topology. Our method selects a small relevant subset from tens of thousands of genes while *simultaneously* identifying nontrivial higher order topological features, i.e., holes, in the data. We first circumvent the problem of high dimensionality by *dualizing* the data, i.e., by studying genes as points in the sample space. Then we select a small subset of the genes as landmarks to construct topological structures that capture persistent, i.e., topologically significant, features of the data set in its first homology group. Furthermore, we demonstrate that many members of these loops have been implicated for cancer biogenesis in scientific literature. We illustrate our method on five different data sets belonging to brain, breast, leukemia, and ovarian cancers.

Keywords: Persistent homology; Cancer; High-dimensional data.

1. Introduction

During the past few decades, science has made great progress in understanding the biology of cancer [1, 2]. The latest technological tools allow assaying tens of thousands of genes simultaneously, providing large volumes of data to search for cancer biomarkers [3, 4]. Ideally, scientists would like to extract some qualitative signal from this data in the hope to better understand the underlying biological processes. At the same time it is desirable that the extracted signal is robust in the presence of noise and errors while effectively describing the dataset [5, 6].

One suite of methods which have enjoyed increased level of success in recent years is based on concepts from the mathematical field of algebraic topology, in particular, *persistent homology* [7–9]. The key benefits of using topological features to describe data include coordinate-free description of shape, robustness in the presence of noise and invariance under many transformations, as well as highly compressed representations of structures [10]. Analysis of the homology of data allows detection of high-dimensional features based on *connectivity*, such as loops and voids, which could not be detected using traditional methods such as clustering [8, 11]. Further, identifying the most *persistent* of such features promises to pick up the significant shapes while ignoring noise [8]. Such analysis of the stable topological features of the data could provide helpful insights as demonstrated by several studies, including some recent ones on cancer data [12, 13].

1.1. *Our contributions*

We present a new method of topological analysis of various cancer gene expression data sets. Our method belongs to the category of exploratory data analysis. In order to more efficiently handle the huge number of genes whose expressions are recorded in such data sets (typically in the order of tens of thousands), we transpose the data and analyze it in its *dual space*, i.e., with each gene represented in the much lower dimensional (in the order of a few hundred) sample space. We then sample critical genes as guided by the topological analysis. In particular, we choose a small subset (typically 120–200) of genes as *landmarks* [14], and construct a family of nested simplicial complexes, indexed by a proximity parameter. We observe topological features (loops) in the first homology group (H_1) that remain significant over a large range of values of the proximity parameter (we consider small loops as topological noise). By repeating the procedure for different numbers of landmarks, we select stable features that persist over large ranges of both the number of landmarks and the proximity parameter. We then further analyze these loops with respect to their membership, working under the hypothesis that their topological connectivity could reveal functional connectivity. Through the search of scientific literature, we establish that many loop members have been implicated in cancer biogenesis. We applied our methodology to five different data sets from a variety of cancers (brain, breast, ovarian, and acute myeloid leukemia (AML)), and observed that in each of the five cases, many members of the significant loops in H_1 have been identified in the literature as having connections to cancer.

Our method is capable of identifying geometric properties of the data that cannot be found by traditional algorithms such as clustering [15, 16]. By employing tools from algebraic topology, our method goes beyond clustering and detects connected components around holes (loops) in the data space. The shown methodology is also different from techniques such as graph [17, 18] or manifold learning [19–23]. Graph algorithms, while identifying connectivity, miss wealth of information beyond clustering. Manifold learning algorithms assume that the data comes from an intrinsically low-dimensional space and their goal is to find a low-dimensional embedding. We do not make such assumptions about the data.

1.2. *Related work*

Several applications of tools from algebraic topology to analyze complex biological data from the domain of cancer research have been reported recently. DeWoshkin et al. [12] used computational homology for analysis of comparative genomic hybridization (CGH) arrays of breast cancer patients. They analyzed DNA copy numbers by looking at the characteristics of H_0 group. Using $Betti_0$ (β_0) numbers, which are the ranks of the zeroth homology groups (H_0), their method was able to distinguish between recurrent and non-recurrent patient groups.

Likewise, Seeman et al. [13] applied persistent homology tools to analyze cancer data. Their algorithm starts with a set of genes that are preselected using the *nondimensionalized standard deviation* metric [13]. Then, by applying persistent homology analysis to the H_0 group, the patient set is recursively subdivided to yield three subgroups with distinct cancer types. By inspecting the cluster membership, a core subset of genes is selected which allows sharper differentiation between the cancer subtypes.

Another example of topological data analysis is the work of Nicolau et al. [24]. Their method termed *progression analysis of disease* (PAD) is applied to differentiate three subgroups of breast cancer patients. PAD is a combination of two algorithms – disease-specific genome analysis (DSGA) [25] and the topology-based algorithm termed *Mapper* [26]. First, DSGA transforms the data by decomposing it into two components – the disease component and the healthy state component, where the disease component is a vector of residuals from a *Healthy State Model* (HSM) linear fit. A small subset of genes that show a significant deviation from the healthy state are retained and passed on to Mapper, which applies a specified filter function to reveal the topology of the data. Mapper identified three clusters corresponding to ER+, ER−, and normal-like subgroups of breast cancer. This work is somewhat different from the previous two papers mentioned above because it does not explicitly analyze features of any of the homology groups.

All studies mentioned above utilized β_0 numbers, thus performing analyses that are topologically equivalent to clustering. In contrast, our method relies on β_1 numbers (ranks of H_1 groups). One can think of β_1 numbers characterizing the loops constructed from connected components (genes) around "holes" in the data. The underlying idea is that connections around holes may imply connections between the participating genes and biological functions. Also, most of other methods use some data preprocessing to limit the initial pool of candidate genes. Our method selects the optimal number of genes as part of the analysis itself.

2. Mathematical background

We review some basic definitions from algebraic topology used in our work. For details, refer one of the standard textbooks [27, 28]. Illustrations of simplices, persistent homology, and identification of topological features from landmarks are available in the literature [8, 14, 34].

2.1. *Simplices and simplicial complexes*

Topology represents the shape of point sets using combinatorial objects called simplicial complexes. Consider a finite set of points in $\mathbb{R}^n$. More generally, the space need not be Euclidean. We just need a unique pairwise distance be defined for every pair of points. The building blocks of the combinatorial objects are *simplices*, which are made of collections of these points.

Formally, the convex hull of $k+1$ affinely independent points $\{v_0, v_1, \ldots, v_k\}$ is a k-simplex. The dimension of the simplex is k, and v_js are its vertices. Thus, a vertex is a 0-simplex, a line segment connecting two vertices is a 1-simplex, a triangle is a 2-simplex, and so on. Observe that each p-simplex σ is made of lower dimensional simplices, i.e., k-simplices τ with $k \leq p$. Here, τ is called a *face* of σ, denoted $\tau \subset \sigma$. A collection of simplices satisfying two regularity properties forms a *simplicial complex*. The first property is that each face of every simplex in a simplicial complex K is also in K. Second, each pair of simplices in K intersect in a face of both, or not at all. Due to these properties, algorithms to study shape and topology run much more efficiently on the simplicial complex than on the original point set.

To construct a simplicial complex on a given point set, one typically considers balls of a given diameter ϵ (called ϵ-ball) centered at each point. The two widely studied complexes of this form are the *Čech* and the *Vietoris-Rips* complexes. A k-simplex is included in the Čech

complex if there exists an ϵ-ball containing all its $k+1$ vertices. Such a simplex is included in the the Vietoris-Rips complex R_ϵ if each pair of its vertices is within a distance ϵ. As such, Vietoris-Rips complexes are somewhat easier to construct, since we only need to inspect pairwise, and not higher order, distances.

However, both the Čech and the Vietoris-Rips complexes have as vertex set all of the points in the data. Such complexes are computationally intensive for datasets of tens of thousands of points. The feasible option is to work with an approximation of the topological space of interest [14]. The key idea is to select only a small subset of points (landmarks), while the rest of points serve as *witnesses* to the existence of simplices. Termed *witness complexes*, such complexes have a number of advantages. They are easily computed, adaptable to arbitrary metrics, and do not suffer from the curse of dimensionality. They also provide a less noisy picture of the topological space. We use the *lazy Witness complex*, in which conditions for inclusion are checked only for pairs and not for higher order groups of points [14], analogous to the distinction between the constructions of Vietoris-Rips and Čech complexes.

We employ the heuristic landmark selection procedure called *sequential maxmin* to select a representative set of landmark points [14, 29, 30]. The first landmark is selected randomly from the point set S. Then the algorithm proceeds inductively. If L_{i-1} is the set of the first $i-1$ landmarks, then the i-th landmark is the point of S which maximizes the function $d(x, L_{i-1})$, the distance between the point x and the set L_{i-1}. We vary the total number of landmarks, exploring each of the resulting lazy witness complexes. The final number of landmarks is chosen so that the resulting witness complex maximally exposes topological features.

2.2. *Persistent homology*

Homology is the concept from algebraic topology which captures how space is *connected*. Thus, homology can be used to characterize interesting features of a simplicial complex such as connected clusters, holes, enclosed voids, etc., which could reveal underlying relationships and behavior of the data set. Homology of a space can be described by its *Betti numbers*. The k-th Betti number β_k of a simplicial complex is the rank of its k-th homology group. For $k = 0, 1, 2$, the β_k have intuitive interpretation. β_0 represents a number of connected components, β_1 the number of holes, and β_2 the number of enclosed voids. For example, a sphere has $\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$, as it has one component, no holes, and one enclosed void.

Consider the formation of a simplicial complex using balls of diameter ϵ centered on points in a set. For small ϵ, the simplicial complex is just a set of disjoint vertices. For sufficiently large ϵ, the simplicial complex becomes one big cluster. What value of ϵ reveals the "correct" structure? Persistent homology [7, 8] gives a rigorous response to this question. By increasing ϵ, a sequence of nested simplicial complexes called a filtration is created, which is examined for attributes of connectivity and their robustness. Topological features appear and disappear as ϵ increases. The features which exist over a longer range of the parameter ϵ are considered as signal, and short-lived features as noise [7, 31]. This formulation allows a visualization as a collection of barcodes (one in each dimension), with each feature represented by a bar. The longer the life span of a feature, the longer its bar. In the example barcodes in Figs. 1–7, the x-axis represents the ϵ parameter, and the bars of persistent loops of interest are circled.

Research questions

Our approach could address several critical questions in the context of cancer data analysis. First, could we select a small subset of relevant genes while *simultaneously* identifying robust nontrivial structure, i.e., topology, of the data? Most previous approaches require the selection of a subset of genes *before* exploring the resulting structure, and hence limiting the generality. Second, could we elucidate higher order interactions (than clusters) between genes that could have potential implications for cancer biogenesis? Higher order structures such as loops could reveal critical subsets of genes with relevant nontrivial interactions, which together have implications to the cancer. Third, could this method work even when data is available from only a *subset* of patients?

3. Data

We analyzed five publicly available microarray datasets of gene expression from four different types of cancer – breast, ovarian, brain, and acute myeloid leukemia (AML). Four of the datasets have the same protocol, GPL570 (HG_U133_Plus_2, Affymetrix Human Genome U133 Plus 2.0 Array). The fifth dataset has a different protocol, HG_U95Av2, which has a fewer number of genes (see Table 1). By including data sets from different protocols, we could verify that the topological features identified are not just artifacts of a particular protocol.

The number of genes represents the number of unique gene id tags defined by a protocol excluding controls. While the brain dataset is of the same protocol as the breast and ovarian datasets, the former one has fewer genes – 46201 vs. 54613. This variability, however, did not affect our procedure to find topological features.

Table 1. Datasets used in the study.

Dataset	Series	Protocol	# Genes	# Samples
Brain	GSE36245	GPL570	46201	46
Breast	GSE3744	GPL570	54613	47
Ovarian	GSE51373	GPL570	54613	28
AML188	GSE10358	GPL570	54613	188
AML170	willm-00119	HG_U95Av2	12558	170

All datasets, except for AML170, were obtained from NCBI Gene Expression Omnibus `http://www.ncbi.nlm.nih.gov/geo/` in October 2013. AML170 was retrieved at the same time from the National Cancer Institute caArray Database at `https://array.nci.nih.gov/caarray/project/willm-00119`.

4. Methods

We work with the raw gene expression values. In particular, we do not log-transform them.

4.1. *Dual space of data*

Traditionally, gene expression data is viewed in its gene space, i.e., the expression profile of patient i with m genes is a point in $\mathbb{R}^m$, $\mathbf{x}_i = [x_{i_1}\ x_{i_2}\ \cdots\ x_{i_m}]$. Each x_{i_j} is an expression of gene j in the patient i. For example, each patient in the Brain dataset is a point in $\mathbb{R}^{46201}$.

We analyze expression data in its dual space, i.e., in its *sample space*. Hence a gene j is represented as a point in $\mathbb{R}^n$, $\mathbf{x}_j = [x_{j_1}\ x_{j_2}\ \cdots\ x_{j_n}]$, where n is the number of samples or

patients. Each x_{j_i} is the expression of the gene j in the patient i. For example, in the same Brain dataset, the points now sit in $\mathbb{R}^{46}$ space. Hence we study gene expression across the span of all patients.

The key motivation for this approach is to handle the high dimensionality in a meaningful way for analyzing the *shape* of data. Given the small size of patients, one could efficiently construct a Vietoris-Rips complex using the set of pairwise distances between patients in the gene space (no need to choose landmarks). But such distances become less discriminatory when the number of genes is large [32, 33]. Working in the dual space, we let our topological method select a manageable number of genes as landmarks to construct the witness complexes, which potentially capture interesting topology of the data. Hence we do not preselect a small number of genes before the topological analysis, as is done by some previous studies [13, 24].

4.2. *Choosing the number of landmarks*

For construction of the witness complexes, the number of landmarks has to be defined *a priori* (Sec. 2.1). Hence the question becomes how many landmarks do we select? We let the data itself guide the selection of genes used as landmarks. If there is a significant loop feature in the data, it would persist through a range of landmarks in H_1 of the complexes. We reconstruct the topological space incrementing the number of landmarks while observing appearance and disappearance of the topological features. Initially, there would be very few small, noisy features because of insufficient number of points. As the number of landmarks increases, some features stabilize, i.e., do not change much either in size or membership. Then they reach their maximal size, and start to diminish once some critical number of landmarks is exceeded (when the "holes" are all filled in). The "optimal" number of landmarks is chosen when the length of the bar representing the topological feature is maximal.

Table 2. Numbers of landmarks selected in each dataset.

Dataset	# landmarks
Brain	120
Breast	110
Ovarian	200
AML188	150
AML170	130

A typical example of such behavior is seen in the Breast dataset (see Fig. 1). A small loop appears when the number of landmarks is $L = 50$. It stabilizes around $L = 90$, reaches its maximum span at $L = 110$, and then decreases as L grows.

4.3. *Composition of loops*

One of the goals of our method is to determine the genes which participate in H_1 features, which could indicate potential implications for cancer biogenesis. Since the first landmark is chosen randomly in the sequential maxmin procedure, the composition of the loops identified may differ based on this first choice. To circumvent this effect, we do 20 different runs in each case to collect possible variations in loop formation. Members of the loops are then pooled together for further analysis. Due to the almost deterministic nature of sequential maxmin selection (apart from the first landmark being selected randomly), we observed very little variation over the 20 runs in most cases. The recovered members of loops are then queried in scientific literature for cancer-related reports.

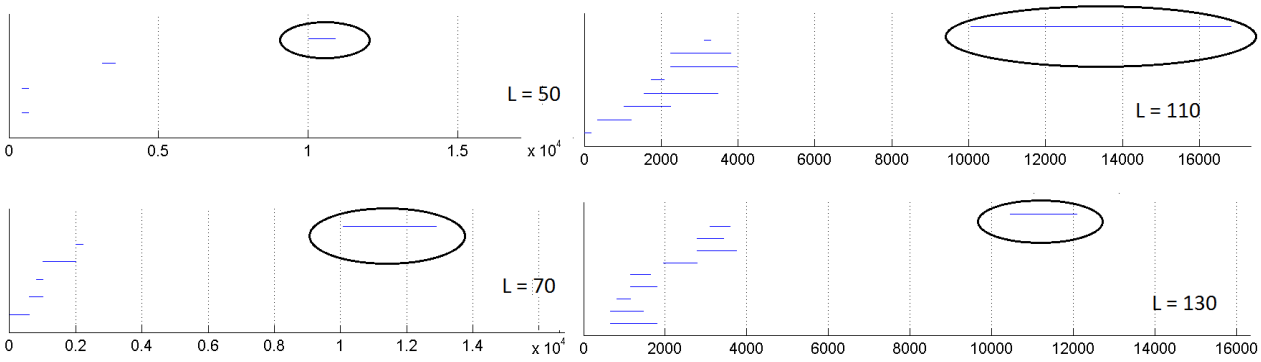

Fig. 1. Evolution of the loop of interest (circled) for varying number of landmarks from L=50 to L=130 in the breast dataset. Here and in Figs. 2–7, the x-axis represents the ϵ parameter.

We implemented our computations using the package JavaPlex [34]. We explored the barcodes for $\mathsf{H}_0, \mathsf{H}_1$, and H_2, but interesting persistent features with members related to cancer biogenesis were detected only for H_1.

5. Results

Persistent topological features in the homology group H_1 were observed in every cancer dataset we analyzed. Representative examples are shown in Figs. 2–6. The AML datasets both had two persistent loops, while the other datasets have one loop each. The Ovarian dataset had a few medium length bars in the H_1 barcode, but we investigated only the longest loop. Once the persistent loops were identified, they were inspected with respect to their composition and relation to cancer through search of scientific literature. Below is a brief description of results for each of the datasets (the full list of all loop members is also available [35]).

Table 3. Selected representatives of loops in different datasets.

Gene	Dataset	Description	References
CAV1	Brain	tumor suppressor gene	[36]
RPL36	Brain	prognostic marker in hepatocellular carcinoma	[37]
RPS11	Breast	downregulation in breast carcinoma cells	[38]
FTL	Breast	prognostic biomarkers in breast cancer	[39]
LDHA	Ovarian	overexpressed in tumors, important for cell growth	[40]
GNAS	Ovarian	biomarker for survival in ovarian cancer	[41]
LAMP1	AML170	regulation of melanoma metastasis	[42]
PABPC1	AML170	correlation with tumor progression	[43]
HLF	AML188	promotes resistance to cell death	[44]
DTNA	AML188	induces apoptosis in leukemia cells	[45]

5.1. *Brain dataset*

The lifespan of the longest loop in this dataset was about 820 (bar between $\approx [480, 1300]$). The loop was consistent over different choices of the first landmark. We identified

13 loop members, out of which 9 were found in cancer literature. Some cancer-related members include EGR1 and CAV1, which have genes been characterized as cancer suppressor genes [36, 46], A2M, which has been identified as a predictor for bone metastases [47], and RPL36 which has been found to be a prognostic marker in hepatocellular carcinoma [37].

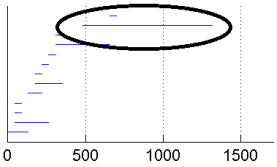

Fig. 2. Representative loop in brain dataset.

5.2. *Breast dataset*

The lifespan of the longest loop in this dataset is in the range $[10080.0, 16684.2]$. As with the brain dataset, this loop is very consistent. However, there were only 10 members of this loop, and 8 of which were found in cancer literature. An interesting feature of this loop is that it had five ribosomal proteins which are known to play a critical role in tightly coordinating p53 signaling with ribosomal biogenesis [48].

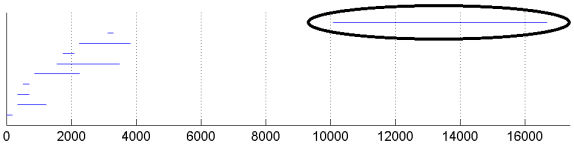

Fig. 3. Representative loop in breast dataset.

5.3. *Ovarian dataset*

The Ovarian dataset had the most variable features in H_1. However, we investigated the loop corresponding to the most consistent and longest bar, which ranged from about 4000 to over 7000. This loop consisted of 17 members, and 9 were mentioned in the cancer-related literature. Among cancer-related members were GNAS, which was identified as "an independent, qualitative, and reproducible biomarker to predict progression-free survival in epithelial ovarian cancer" [41], and HNRNPA1, which has been identified as a potential biomarker for colorectal cancer [49].

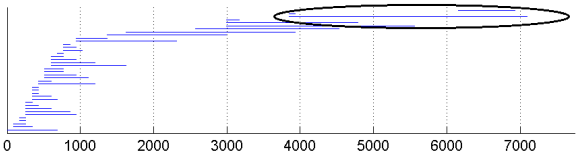

Fig. 4. Representative loop in ovarian dataset.

5.4. *AML188 dataset*

Acute myeloid leukemia 188 (AML188) had two significant loops (as did AML170). The first one occurred at $[25200.0, 102200.0]$ and the second one at $[78400.0, 146219.24]$. The first loop has 27 members while the second one only 6. Altogether, only 14 of these 33 genes were mentioned in cancer literature. Some cancer-related representatives were hepatic leukemia

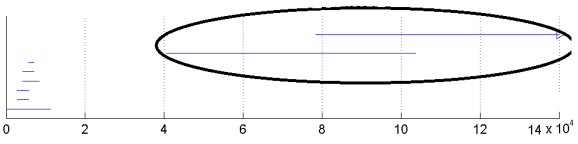

Fig. 5. Representative loops in AML188 dataset.

factor (HLF) which promotes resistance to cell death [44], RPL35A known for inhibition of cell death [50], and GRK5 which regulates tumor growth [51]. A group of zinc finger proteins were present in the first loop, some of which have been reported as novel biomarkers for detection of squamous cell carcinoma [52, 53].

5.5. *AML170 dataset*

AML170 comes from caArray database and its protocol HG_U95Av2 has only 12558 genes. Even though this protocol had a smaller (about 1/4) number of genes compared to the other data sets, we still detected two loops in this dataset. They were relatively shorter than the loops in AML188, and occurred at $[5800.0, 11400.0]$ and $[11000.0, 17600.0]$. The two loops comprised of 19 members, of which 10 were found in cancer-related literature. These relevant members included ubiquitin C (UBC), which was recently identified as a novel cancer-related protein [54], PABPC1 whose positive expression is correlated with tumor progression in esophageal cancer [43], and LAMP1 which facilitates lung metastasis [42].

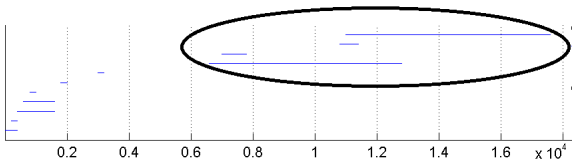

Fig. 6. Representative loops in AML170 dataset.

6. Discussion

The Breast, Brain, and Ovarian datasets had only one persistent loop, while AML170 and AML188 had two. Also, the AML datasets had a higher number of patients (samples) than the other three sets (see Table 1). Is this fact just a coincidence or, indeed, does the number of H_1 features (loops) correlate with the number of dimensions? To address this question, we chose samples of random and progressively larger (25–175) subsets of patients from AML170 and AML188 while also increasing the number of landmarks, and studied the evolution of H_1 features. In other words, we repeated our method on smaller subsets of patients from these datasets. Both the AML datasets contained two loops even with only 25 dimensions (see Fig. 7 for AML170), and continued to do so for the progressively larger subsets. Thus, the number of significant H_1 features appears to depend on intrinsic qualities of the data rather than the number of dimensions, demonstrating the robustness of our method to the number of patients in the dataset.

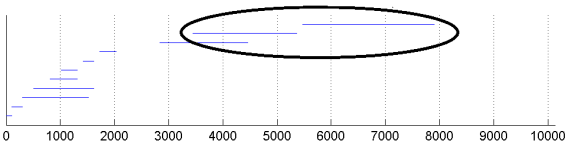

Fig. 7. Two persistent loops in the AML170 dataset detected using only 25 dimensions at the number of landmarks L=130.

An important property of a loop is its lifespan [55]. One may note that the life span of loops for different datasets vary significantly. For example, the lifespan of a significant H_1 feature in the brain dataset is only 820, while for AML188 the lifespan of the first loop is $77,000$. This difference is not only due to the increase in the actual size of a loop as indicated by the number of points comprising the loop (13 vs. 27 in this case), but also in part because

of the different absolute values in microarray expression data. The maximum value for the brain dataset is 24×10^3, while for AML188 is 3×10^6. Therefore, the absolute length of an H_1 feature is not as important as its length relative to other H_1 features.

The crucial step of our method is the choice of landmarks. The goal here is the efficient inference of the topology of data, while selecting a small subset of potentially relevant genes. Landmarks chosen using sequential maxmin tend to cover the dataset better and are spread apart from each other, but the procedure is also known to pick outliers [14]. In our datasets, outliers are typically identified by extreme expression values [56]. We examined the expression values of the chosen landmarks, and found that very few of them had extreme values (Figs. 8 and 9). Similarly, the expressions of the genes implicated in cancer biogenesis (among the loop members) did not have any extreme values, and in fact appear to follow normal distributions. We infer that this observation results because sequential maxmin indeed picks points on the outskirts of the topological features. Further, the distribution of expressions of the loop members suggests that the group as a whole could have potential implications for the disease. More interestingly, the "hole" structure means such groups could not potentially be identified by traditional coexpression or even differential expression analyses [57].

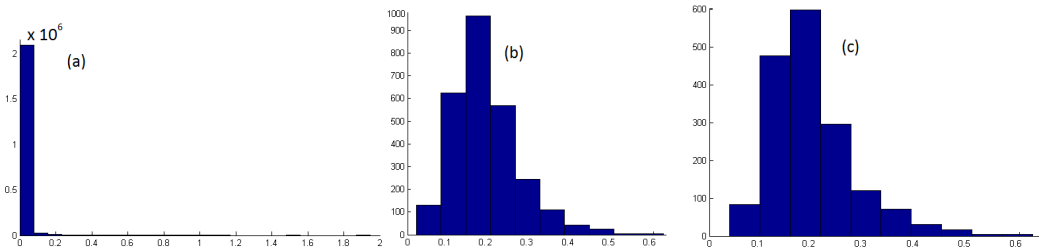

Fig. 8. Histograms for AML170 dataset, x-axis represents gene expression level of scale 10^4. (a) distribution of gene expressions for the whole set; (b) distribution of gene expressions for only the loop members; and (c) distribution of gene expressions for cancer-related loop members.

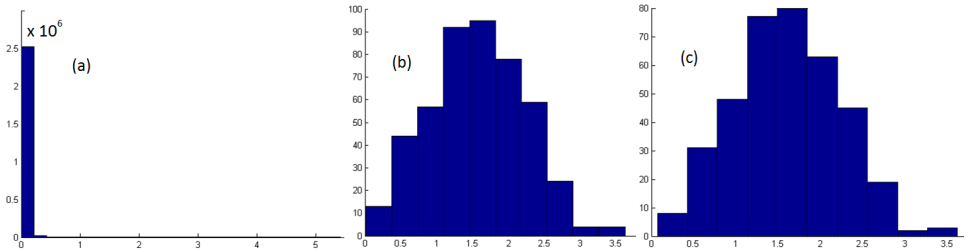

Fig. 9. Histograms for Breast dataset, x-axis represents gene expression level of scale 10^5. (a) distribution of gene expressions for the whole set, (b) distribution of gene expressions for the loop members only, (c) distribution of gene expressions for cancer-related loop members.

The computational complexity of our method is based on the current implementation of JavaPlex [34], where the two main steps are building and filtering simplicial complexes, and computing homology. Up to dimension 2 (β_2), homology could be computed in $O(n^3)$ time [28]. However, building simplicial complexes relies on clique enumeration, which is NP-complete, and has a complexity of $O(3^{n/3})$ [58, 59]. Further, JavaPlex requires explicit enumeration

of simplicial facets appearing at each filtration step, implying the need for large memory resources [34].

The cancer-related loop members identified from the two AML datasets were distinct apart from the prominent group of ribosomal proteins. This observation could be explained by two main reasons. First, AML188 has four times the number of genes as AML170 (see Table 1). Second, JavaPlex identifies only *one* representative of each homology class. That is, if a significant topological feature (hole) exists, it will be identified, but only one loop will be found around that hole. There could be other relevant points in proximity to the hole, but are not guaranteed to be included in the loop. If we have prior knowledge of some genes being relevant, we could try to identify loops around the holes that include these genes as members. In this case, the other members of the identified loops could also have potential implications for the cancer biogenesis. Methods to find a member of a homology class that includes specific points could be of independent interest in the context of optimal homology problems [60, 61].

7. Conclusion

We have presented a method to look at cancer data from a different angle. Unlike previous methods, we look at characteristics of the first homology group (H_1). We identify the persistent H_1 features (which are loops, rather than connected components) and inspect their membership. Importantly, our approach finds potentially interesting connections among genes which cannot be found otherwise using traditional methods. This geometric connectedness may imply functional connectedness, however, this is yet to be investigated by oncologists. If such connections are indeed implied, then the genes in the loops could together form a characteristic "signature" for the cancer in question.

Acknowledgment: Krishnamoorthy acknowledges support from the National Science Foundation through grant #1064600.

References

[1] X. Yu, S. Narayanan, A. Vazquez and D. R. Carpizo, *Apoptosis* **19**, 1055 (2014).
[2] J. D. Patel, L. Krilov, S. Adams, C. Aghajanian, *et al.*, *J. Clin. Oncology.* **32**, 129 (2014).
[3] A. Singh and N. Kumar, *International Journal of Current Research and Review* **5**, 1 (2013).
[4] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, *et al.*, *Science* **291**, 1304 (2001).
[5] S. Valarmathi, A. Sulthana, K. Latha, R. Rathan, R. Sridhar and S. Balasubramanian, *Proc. 2nd Intl. Conf. Soft Comput. Prob. Solv (SocProS 2012)* , 1451 (2014).
[6] A. E. Pozhitkov, P. A. Noble, J. Bryk and D. Tautz, *PloS One* **9**, p. e91295 (2014).
[7] H. Edelsbrunner, D. Letscher and A. Zomorodian, *Discret. Comput. Geom.* **28**, 511 (2002).
[8] R. Ghrist, *Bulletin of the American Mathematical Society* **45**, 61 (2008).
[9] G. Carlsson, A. Zomorodian, A. Collins and L. Guibas, *Proc. Eurog./ACM SIGGRAPH Symp. Geom. Proc. SGP '04*, 124 (2004).
[10] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, *et al.*, *Scientific Reports* **3** (2013).
[11] G. Carlsson, *Bulletin of the American Mathematical Society* **46**, 255 (January 2009).
[12] D. DeWoskin, J. Climent, I. Cruz-White, M. Vazquez, C. Park and J. Arsuaga, *Topology and its Applications* **157**, 157 (2010).
[13] L. Seemann, J. Shulman and G. H. Gunaratne, *ISRN Bioinformatics* (2012).
[14] V. de Silva and G. Carlsson, *SPG '04: Proc. Sympos. Point-Based Graphics*, 157 (2004).

[15] R. Fa, A. K. Nandi and L.-Y. Gong, *Comm. Control. Sig. Proces. (ISCCSP)*, 1 (2012).
[16] A. K. Jain, *Pattern Recognition Letters* **31**, 651 (2010).
[17] D. A. Spielman and S.-H. Teng, *SIAM Journal on Computing* **42**, 1 (2013).
[18] D. J. Cook and L. B. Holder, *Mining graph data* (John Wiley & Sons, 2006).
[19] J. B. Tenenbaum, V. D. Silva and J. C. Langford, *Science* **290**, 2319 (2000).
[20] S. T. Roweis and K. S. Lawrence, *Science* **290**, 2323 (2000).
[21] D. L. Donoho and C. Grimes, *Proceedings of the National Academy of Sciences* **100**, 5591 (2003).
[22] M. Belkin and P. Niyogi, *NIPS* **14**, 585 (2001).
[23] R. R. Coifman and S. Lafon, *Applied and computational harmonic analysis* **21**, 5 (2006).
[24] M. Nicolau, A. J. Levine and G. Carlsson, *PNAS* **108**, 7265 (2011).
[25] M. Nicolau, R. Tibshirani, A.-L. Børresen-Dale and S. S. Jeffrey, *Bioinformatics* **23**, 957 (2007).
[26] G. Singh, F. Mémoli and G. E. Carlsson, *Symp. Point Based Graphics*, 91 (2007).
[27] J. R. Munkres, *Elements of Algebraic Topology* (Addison–Wesley, 1984).
[28] H. Edelsbrunner and J. L. Harer, *Computational Topology* (American Math. Society, 2009).
[29] H. Adams and G. Carlsson, *SIAM J. Img. Sci.* **2**, 110 (2009).
[30] G. Carlsson, T. Ishkhanov, V. de Silva and A. Zomorodian, *Intl. Jrnl. Comp. Vis.* **76**, 1 (2008).
[31] A. Zomorodian and G. Carlsson, *Discrete Computational Geometry* **33**, 249 (2005).
[32] K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, *Database Theory–ICDT99*, 217 (1999).
[33] M. Ledoux, *The concentration of measure phenomenon* (American Mathematical Soc., 2005).
[34] A. Tausz, M. Vejdemo-Johansson and H. Adams, Javaplex: A software package for computing persistent topological invariants (2011), Available at http://comptop.stanford.edu/programs/.
[35] Supplement: http://math.wsu.edu/faculty/bkrishna/Topo/Cancer/Supplement.pdf.
[36] Cav1 caveolin 1. NCBI Gene Database http://www.ncbi.nlm./nih.gov/gene/857.
[37] M. J. Song, C. K. Jung, C. H. Park, W. Hur, J. E. Choi, *et al.*, *Pathol. Int.* **61**, 638 (2011).
[38] D. Nadano, C. Aoki, T. Yoshinaka, S. Irie and T. A. Sato, *Biochemistry* **40**, 15184 (2001).
[39] G. Ricolleau, C. Charbonnel, L. Lod, D. Loussouarn, *et al.*, *Proteomics* **6**, 1963 (2006).
[40] D. Zhao, S. W. Zou, Y. Liu, X. Zhou, Y. Mo, *et al.*, *Cancer Cell* **23**, 464 (2013).
[41] E. Tominaga, H. Tsuda, T. Arao, S. Nishimura, *et al.*, *Gynecol. Oncol.* **118**, 160 (2010).
[42] A. K. Agarwal, R. P. Gude and R. D. Kalraiya, *Bioch. Biophys. Res. Comm.* **449**, 332 (2014).
[43] N. Takashima, H. Ishiguro, Y. Kuwabara, M. Kimura, *et al.*, *Oncol. Rep.* **15**, 667 (2006).
[44] K. M. Waters, R. L. Sontag and T. J. Weber, *Toxicol. Appl. Pharmacol.* **268**, 141 (2013).
[45] R. Navakauskienė, G. Treigytė, V. V. Borutinskaitė, D. Matuzevičius, D. Navakauskas and K. E. Magnusson, *J. Proteomics* **75**, 3291 (2012).
[46] Egr1 early growth response. NCBI Gene Database http://www.ncbi.nlm.nih./gov/gene/1958.
[47] Y. Kanoh, N. Ohtani, T. Mashiko, S. Ohtani, *et al.*, *Anticancer Res.* **21**, 551 (2001).
[48] A. Artero-Castro, J. Castellvi, A. García, J. Hernández, *et al.*, *Hum. Pathol.* **42**, 194 (2011).
[49] Y. L. Ma, J. Y. Peng, P. Zhang, L. Huang, W. J. Liu, *et al.*, *J. Proteome Res.* **8**, 4525 (2009).
[50] C. D. Lopez, G. Martinovsky and L. Naumovski, *Cancer Lett.* **180**, 195 (2002).
[51] J. I. Kim, P. Chakraborty, Z. Wang and Y. Daaka, *J. Urol.* **187**, 322 (2012).
[52] Y. J. Jou, C. D. Lin, C. H. Lai, C. H. Tang, *et al.*, *Clin. Chim. Acta.* **412**, 1357 (2011).
[53] R. Lovering and J. Trowsdale, *Nucleic Acids Res.* **19**, 2921 (1991).
[54] Y. H. Wong, R. H. Chen and B. S. Chen, *J. Theor. Biol.* (2014).
[55] E. Carlsson, G. Carlsson and V. De Silva, *Intl. J. Comput. Geom. Appl.* **16**, 291 (2006).
[56] V. J. Hodge and J. Austin, *Artificial Intelligence Review* **22**, 85 (2004).
[57] Y. Choi and C. Kendziorski, *Bioinformatics* **25**, 2780 (2009).
[58] J. D. Moon and L. Moser, *Israel journal of Mathematics* **3**, 23 (1965).
[59] J. Binchi, E. Merelli, M. Rucco, G. Petri, *et al.*, *El. Notes in Theor. Comp. Sci.* **306**, 5 (2014).
[60] T. K. Dey, A. N. Hirani and B. Krishnamoorthy, *SIAM Journal on Computing* **40**, 1026 (2011).
[61] T. K. Dey, J. Sun and Y. Wang, *Proc. 26th Symp. Comp. Geom, SoCG '10*, 166 (2010).