Topological Descriptors of Histology Images

Nikhil Singh, Heather D. Couture, J. S. Marron, Charles Perou, and Marc Niethammer

The University of North Carolina, Chapel Hill, USA

Abstract. The purpose of this study is to investigate architectural characteristics of cell arrangements in breast cancer histology images. We propose the use of topological data analysis to summarize the geometric information inherent in tumor cell arrangements. Our goal is to use this information as signatures that encode robust summaries of cell arrangements in tumor tissue as captured through histology images. In particular, using ideas from algebraic topology we construct topological descriptors based on cell nucleus segmentations such as persistency charts and Betti sequences. We assess their performance on the task of discriminating the breast cancer subtypes Basal, Luminal A, Luminal B and HER2. We demonstrate that the topological features contain useful complementary information to image-appearance based features that can improve discriminatory performance of classifiers.

1 Introduction

Clinical diagnosis of cancer is performed by assessing properties of biopsied tissue. For breast cancer, architectural criteria based on the organization and arrangement of cells, form critical cues for a pathologist to assess and grade tissue samples. Methods to automatically and objectively analyze architectural characteristics of human tissue from histology images are therefore needed to aid pathologists and to computationally *quantify* tissue architecture.

A variety of geometric approaches to pattern or shape recognition have been investigated over the last 15 years. Of these, topological data analysis (TDA) enables the investigation of structural characteristics of high-dimensional data [1,3,4]. The strength of TDA lies in its two core ideas: (a) representing objects based on their topology making it invariant to small changes in shapes and hence robust to noise, and (b) considering *a range* of coarse to fine scales of topological changes, thereby, summarizing large *and* small scale objects.

This paper explores to which extent TDA can characterize cell organization and tissue in breast cancer histology images. We study how to analyze nuclear arrangements through TDA to distinguish genetically derived breast cancer subtypes. These subtypes can be used to guide personalized treatments. We propose topological methods for feature extraction and present a method to combine topological summaries with other imaging features thereby demonstrating that topological features can add information over local image-based descriptors. We first review the necessary background of computational topology for TDA in § 2.1

and present its application to the analysis of breast cancer histology images in § 2.2. In § 3, we discuss and evaluate the extracted topological summaries.

2 Methodology

2.1 Background on topological data analysis and homology groups

Topological data analysis uses concepts from algebraic topology [10,3] and provides methods to characterize geometric information in the data. The classical way is to represent the data in the form of combinatorial objects called simplicial complexes to form a topological space. TDA then studies connectivity information and characterizes loops, voids and higher dimensional surfaces within the space [4]. To analyze tissue architecture the simplicial complex is built, for example, based on the center points of segmented cell nuclei, which define a point-cloud. See the section on the Vietoris-Rips filtration on point clouds below. We review the necessary concepts in topological data analysis in what follows.

Simplicial complexes and filtration. A simplicial complex consists of a collection of simplices, such as vertices, edges, triangles or d-dimensional simplices, which is closed under inclusion. More precisely, a simplicial complex is a collection, K of d-dimensional simplices, τ, such that if $\tau \in K$, all its faces, $\sigma \subset \tau$, are also in K. A subcollection L of simplices from K which itself is a simplicial complex, forms a subcomplex of K, denoted as $L \hookrightarrow K$. A nested sequence of simplicial subcomplexes that ascends from an empty set all the way up to K is called a *filtration* of K. An N-step filtration is therefore denoted by the sequence,

$$\emptyset = \mathscr{F}_0 K \hookrightarrow \mathscr{F}_1 K \hookrightarrow \mathscr{F}_2 K \hookrightarrow \ldots \hookrightarrow \mathscr{F}_{N-1} K \hookrightarrow \mathscr{F}_N K = K.$$

Topological summaries using homology groups. The representation of data by a simplicial complex, K, allows for its characterization through homology groups, which we denote as $H_d(K)$. $H_d(K)$ is the collection of d-dimensional holes. Homology groups consist of groups of d-dimensional homology generators, e.g., 1D connected components for $d = 0$, 2D loops for $d = 1$, 3D cavities for $d = 2$, and so on. The rank of $H_d(K)$ is called the d-th *Betti number*. We now discuss a simple example of a filtration of a 2D simplicial complex formed by point cloud data entities, which will form the basis of our analysis of tissue data.

Example of Vietoris-Rips filtration on point cloud simplicial complex. Consider a set of points, $C \subset R^d$ (Fig. 1, left) and define the largest possible simplicial complex, K_C, consisting of all subsets of C. We construct a subcomplex by using a threshold on the pairwise distances between any two points. We define a simplicial subcomplex as a function of filtration scale, s. The subcomplex $\mathscr{F}_s K_c$ consists of a subcollection of points with pairwise distance between them less than s. It is helpful to think of this subcollection of points obtained when the balls of radius, s, centered at each point intersect (Fig. 1, center). Increasing the ball radii results in a chain of subcomplexes defining a filtration of K_c. For a given s, the 0-dimensional homology group consists of the set of independent components. The 1-dimensional group consists of the loops. The rank

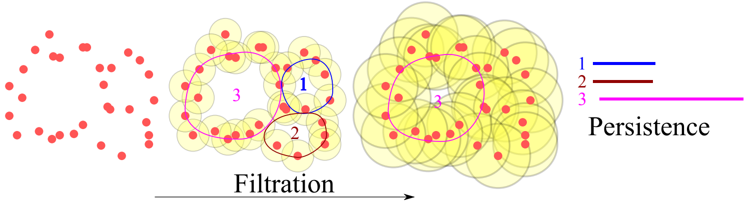


Fig. 1: Vietoris-Rips filtration of a point cloud. Two steps of filtration that depict topological features forming and disappearing (left). Three loops as H_1 formed during filtration persist for different length of filtration (right).

of $H_0(K_c)$ is the count of connected components and the rank of $H_1(K_c)$ is the count of loops (Fig. 1, right). These topological objects can be summarized in terms of their persistence during the filtration steps. This results in a signature representation of topology of the data in the form of persistent diagrams or bar charts [2].

2.2 Cell architecture, nuclei arrangement and topology

Clinical diagnosis of breast cancer is usually performed by analyzing H&E-stained histology images. The arrangement of cells and other structures are some of the cues guiding a pathologist to characterize tissue and assess prognosis. Hence, TDA as described in Sec. 2.1 seems a natural choice to quantify such arrangements. In previous work, TDA has been applied to microarray data. Nicolau et al. [12] propose cluster analysis of persistence charts derived from the simplicial complex of microarray data to identify breast cancer subtypes. However, the topological characterizations of nuclear arrangements in tumor tissue has not yet been investigated. Fig. 2 suggests that such an approach could capture architectural characteristics, using an example H&E stained histology image. The patterns in the organization of cells is evident simply by observing the nuclei in a region (Fig. 2, b). Distinct topological object characterizations

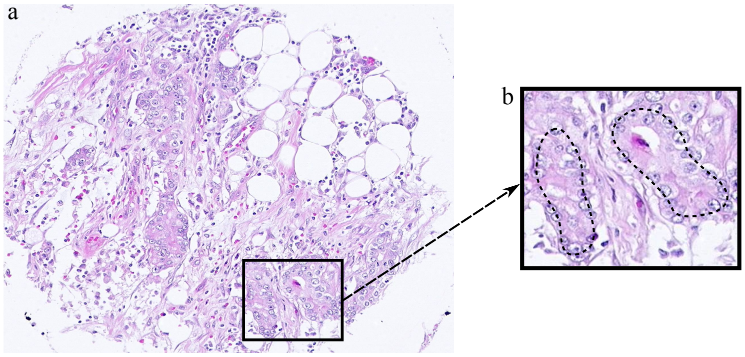


Fig. 2: Example of a histology image for a tumor of subtype, Basal. The highlighted loops formed due to the arrangement of nuclei are an example of architectural feature.

such as nuclear connectivity and loops based on the Vietoris-Rips filtration of nuclei centers look promising as summaries of the arrangement of nuclei in tissue.

3 Experiments

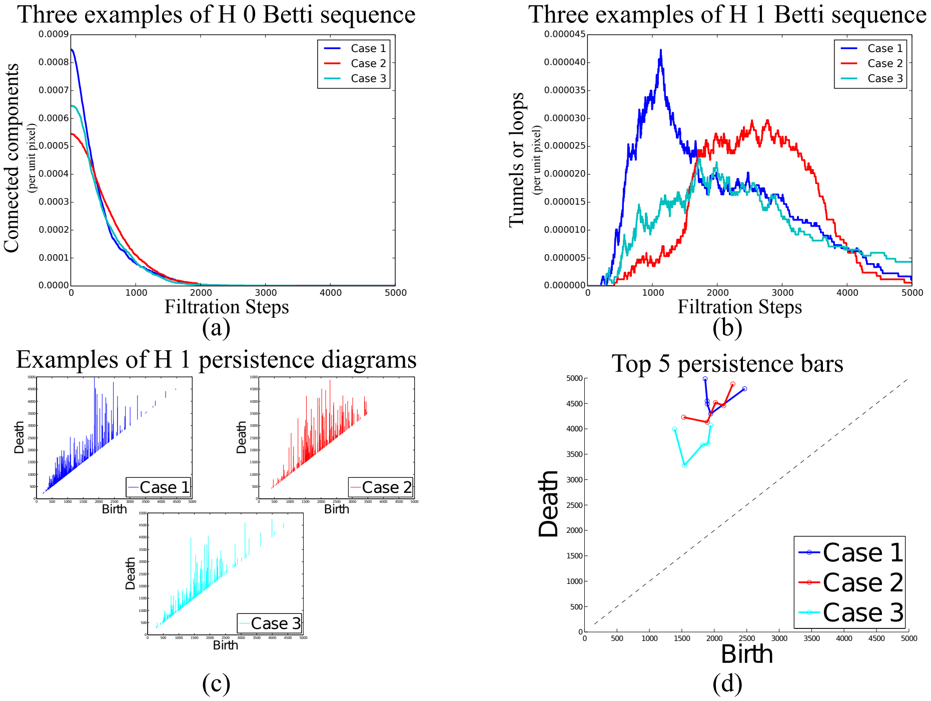
We present our topological analysis of nuclei arrangements using a dataset of breast cancer microarray tissue samples, imaged at the University of British Columbia from a Washington University cohort of patients [13]. The dataset consists of 111 subjects with two images each. Subtypes of Basal, Luminal A, Luminal B, and HER2 have been assigned to each sample by molecular means. The ensemble has 38 Basal, 35 Luminal A, 18 Luminal B and 17 HER2 cases and our goal is to assess whether these subtypes differ in terms of their topological characteristics. We combine the features extracted from two images to construct a single patient level representation.

Constructing topological summaries of homology images. As discussed in § 2.1, we define the simplicial complex by representing the collection of nuclei as point clouds such that the center of mass of each nucleus denotes a vertex. We perform the Vietoris-Rips filtration of this complex by growing balls centered at each vertex. The initial start radius of a ball is proportional to the mass of its nucleus. Since each nucleus has a different size, such an initialization ensures that at the first step, the balls approximately encircle the respective nuclei. We successively increase the radii of all balls with equal rates and stepsizes. Beginning at the start scale, where the number of connected components is equal to the number of nuclei, this filtration computes the generators of zero (H_0: connected components) and one dimensional (H_1: loops) homology groups. We use the Perseus software [11,9] to perform the filtration on the Rips complex.

We summarize the resulting topological objects into a sequence of Betti numbers, e.g., Fig. 3 a and b. We convert the Betti numbers into densities, by dividing them by the area of the tissue in the image making the representation invariant to tissue size. Fig. 3 b suggests that loops exhibit the most dynamics with changing filtration scale. Thus, in another representation, we consider the bar chart representation, called the persistence diagram, based on birth and death of loops during filtration (Fig. 3 c and d). Small bars can be considered as noise artifacts in imaging and segmentation. For robustness, we consider the top few persistent bars (lengthwise), arranged in the order of their birth, as features.

3.1 Evaluating topological features

We perform leave-one-out cross-validation experiments to demonstrate the discriminatory capabilities of the topological features to classify tissue images into subtypes. We use distance weighted discrimination (DWD) [8] as a classifier. For each pair of subtypes, we evaluated the prediction accuracy using the two classifiers on Betti densities, top 5 and top 75 persistent bars. The best results were obtained for the Basal vs Luminal A classification using Betti density features and for Luminal B vs HER2 classification using the top 5 persistent bars.



(a) (b) (c) (d)

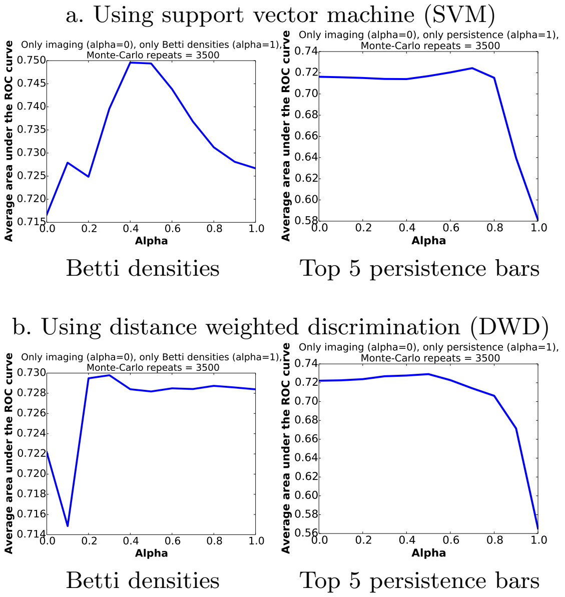
Fig. 3: Different topological summaries for nuclei arrangements in histology images demonstrated for three different examples. (a) and (b) display the Betti densities for H_0 and H_1 homology as a function of filtration steps, while (c) shows the corresponding persistence diagrams, and (d) shows how those are summarized into just the 5 longest bars, in birth order (connecting line segments represent the order of arrangement).

For Basal vs Luminal A, we achieved a classification accuracy of 69.86%, an improvement of 17.80% over the baseline accuracy of predicting the subtype based on the proportion of the samples of the largest class. For Luminal B vs HER 2 subtype classification, the topological features improved the prediction accuracy by 17.14% over the baseline, giving an overall accuracy of 68.57%.

3.2 Joint analysis of topological and other imaging features

Besides topological connectivity and nuclei arrangements, a histology image has other potentially complementary information about tumor tissue. We augment the topological features with those extracted from local image intensities: we construct another set of features learned directly from image patches. A dictionary is learned by modeling 9×9 pixel image patches as sparse linear combinations of dictionary elements [7]. Each patch of an image is encoded with this dictionary. The frequency of usage of each dictionary element is summarized with a 128 bin histogram resulting in a 128-dimensional feature vector for each image.

We define the combined feature space as a product space of topological features, represented as a matrix T, and the patch based image feature space, repre-



a. Using support vector machine (SVM)

Betti densities

Top 5 persistence bars

b. Using distance weighted discrimination (DWD)

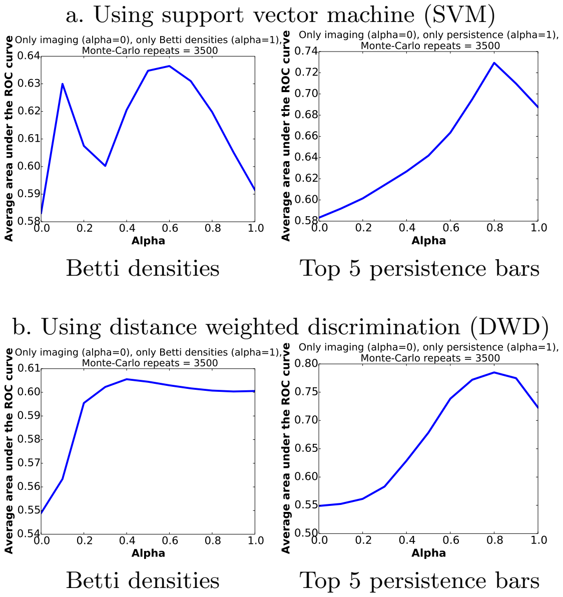
Betti densities

Top 5 persistence bars

Fig. 4: Repeated 5-fold cross validation for Basal vs Luminal A: combining TDA with patch-based image features suggests improvement in performance for certain cases.

sented as a matrix I. In these matrices, let rows represent samples and columns represent features. We construct a convex combination of columnwise concatenated features, to form the augmented feature matrix, $C = \begin{pmatrix}\alpha T & (1-\alpha)I\end{pmatrix}$, where α controls the feature weight; α is a relative weight when both feature matrices are normalized to have unit variance. This is achieved by mean centering and dividing the two matrices by the sum of their eigenvalues. Another possibility is to use a multi-kernel approach to combine complementary features [5].

To investigate whether topological features and the image based-features provide complementary information relevant to cancer subtypes, we assess the receiver operator characteristics (ROC) of the classifiers over the entire range of $\alpha \in [0, 1]$. Note that ROC analysis is not applicable to leave-one-out crossvalidation since we get test prediction only on a single test sample for each trained model. Hence, we perform Monte-Carlo (MC) repetitions of 5-fold crossvalidation using both SVM and DWD classifiers for 3500 repetitions. We choose the average area under the ROC curve (AUC) as the metric of performance. AUC is a more stable performance measure than accuracy as it considers the whole range of thresholds for a classifier [6]. For each MC iteration, we compute the false positive (FPR) and true positive (TPR) rates for every crossvalidation run for the test data, resulting in an average FPR and TPR to give a mean AUC. We test this for the classification tasks that resulted in the best performance with the leave-one-out classification using only the topology features in § 3.1, i.e., Basal vs Luminal A and Luminal B vs HER2. The trends in AUC as a function of α suggest that, for some cases, the topological Betti features perform better when



a. Using support vector machine (SVM)

Betti densities

Top 5 persistence bars

b. Using distance weighted discrimination (DWD)

Betti densities

Top 5 persistence bars

Fig. 5: Repeated 5-fold cross validation for Luminal B vs HER2: combining TDA with patch-based image features suggests improvement in performance for certain cases.

compared to the patch based image-appearance features for both the classification tasks using either of the classifiers. (Fig. 4 and Fig. 5). In terms of AUC, the Betti features perform better than the persistence summaries, for discriminating Basal from Luminal A. However, the persistence summaries outperform the Betti features, in average AUC metric, for Luminal B vs HER2 discrimination. Another observation is that the AUC peaks in the middle for some of the plots suggesting that a combination of the two features may provide useful information. The results on average accuracy metric as a function of α did not match for all cases with those obtained for the AUC metric. For the top 5 persistence summaries for Luminal B vs HER2 with DWD, both the average AUC and the accuracy analyses suggest that topological features massively outperform the image-appearance features. In particular, using top 5 persistence summaries with DWD improve the AUC by 43% and the accuracy by 22% over imaging features and their combination further adds 13% and 8% improvements, respectively. Additional results are in the supplementary material at `http://www.cs.unc.edu/~nsingh/publications/nsingh2014topology_breast_cancer_supplementary.pdf`.

4 Discussion

We proposed the use of topological methods to summarize architectural features of cancerous tissue. We constructed geometric features that quantitatively capture arrangements of nuclei as seen in histology images. We explored multiple topological features derived from the homology groups resulting from filtrations of simplicial complexes defined using nuclei locations. Our experiments suggest that, for most cases, topological features perform as good as the patch based features on the task of discriminating cancer subtypes. We also demonstrate that for certain combinations, the topological features provide complementary information, which in turn improves the performance of classifiers. Our future work will include exploring more informative features from the persistence diagram and will repeat the analysis on bigger datasets. A possibility could be to use persistent bars from the chart but maintain their order of filtration. This would result in a sparse feature vector of size equal to the number of filtration steps.

We believe that the topological study of histology image data provides complementary information to image-appearance about tissue properties. It holds promise to improve our understanding of cytological and architectural differences in tissues. In the context of cancer a topological characterization of tumor tissue could potentially aid clinicians in cancer diagnosis and treatment planning.

Acknowledgements. Data collection used in this article was funded by the Strategic Partnering to Evaluate Cancer Signatures (SPECS) group. This research is supported by grants, NSF EECS-1148870, NSF EECS-0925875, and NIH P41-EB002025.

References

1. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society 46(2), 255–308 (2009) 1
2. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete Comput Geom 37(1), 103–120 (2007) 3
3. Edelsbrunner, H., Harer, J.: Computational topology: an introduction. American Mathematical Soc. (2010) 1, 2
4. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. Discrete Comput Geom 28(4), 511–533 (2002) 1, 2
5. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. The Journal of Machine Learning Research 12, 2211–2268 (2011) 6
6. Ling, C.X., Huang, J., Zhang, H.: Auc: a statistically consistent and more discriminating measure than accuracy. In: IJCAI. vol. 3, pp. 519–524 (2003) 6
7. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. The Journal of Machine Learning Research 11, 19–60 (2010) 5
8. Marron, J., Todd, M.J., Ahn, J.: Distance-weighted discrimination. Journal of the American Statistical Association 102(480), 1267–1271 (2007) 4
9. Mischaikow, K., Nanda, V.: Morse theory for filtrations and efficient computation of persistent homology. Discrete Comput Geom 50(2), 330–353 (2013) 4
10. Munkres, J.R.: Elements of algebraic topology, vol. 2. Addison-Wesley Reading (1984) 2
11. Nanda, V.: Perseus: The Persistent Homology Software. http://www.sas.upenn.edu/~vnanda/perseus (Accessed 30/04/14) 4
12. 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. Proceedings of the National Academy of Sciences 108(17), 7265–7270 (2011) 3
13. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., et al.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology 27(8), 1160–1167 (2009) 4