Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer

Javier Arsuaga · Nils A. Baas · Daniel DeWoskin · Hideaki Mizuno · Aleksandr Pankov · Catherine Park

Received: 15 February 2011 / Revised: 20 January 2012 / Accepted: 22 January 2012 / Published online: 28 March 2012 © Springer-Verlag 2012

Abstract Genomic technologies measure thousands of molecular signals with the goal of understanding complex biological processes. In cancer these molecular signals have been used to characterize disease subtypes, signaling pathways and to identify subsets of patients with specific prognosis. However molecular signals for any disease type are so vast and complex that novel mathematical approaches are required for further analyses. Persistent and computational homology provide a new method for these analyses. In our previous work we presented a new homology-based supervised classification method to identify copy number aberrations from comparative genomic hybridization arrays. In this work we first propose a theoretical framework for our classification method and second we extend our analysis to gene expression data. We analyze a published breast cancer data set and find that that our method can distinguish most, but not all, different breast cancer subtypes. This result suggests
that specific relationships between genes, captured by our algorithm, help distinguish between breast cancer subtypes. We propose that topological methods can be used for the classification and clustering of gene expression profiles.

**Keywords**  Computational homology · Breast cancer subtypes · Gene expression

1 Introduction

One of the major goals of genomic technologies is the identification of molecular signatures that can characterize disease. In cancer some of these molecular signatures can be found as changes in the structure of chromosomes (called chromosome aberrations) and/or in the expression of particular genes. Identification of these changes using microarrays has been used for the search of oncogenes and tumor suppressor genes, for the classification of tumor subtypes, and for patient stratification (e.g. reviewed in [3,10,25,32]). However despite the progress made, the data are so vast and complex that new mathematical methods are required for further analyses. One approach is to use computational algebraic topology and/or persistent homology (reviewed in [6,14,39,9,17]). In persistent homology a simplicial complex is constructed from a point cloud and the homological properties of this simplicial complex are investigated to uncover relevant properties of the data. Examples of applications of persistent homology are found across physics, engineering and biology (e.g. [18,12,22,27]).

In [13] we introduced a new computational homology method to analyze array comparative genomic hybridization (CGH) with the goal of identifying recurrent copy number aberrations (CNAs) in cancer. This method uses a sliding window algorithm to associate a set of point clouds to each array CGH. Once the data are represented by point clouds we use techniques from persistent homology to classify the data. In particular we proposed that $\beta_0$ is a measure of the copy number aberrations in the cancer genome and can be used to identify CNAs specific to a population of patients.

Our aims in this paper are to first provide a theoretical framework for our sliding window algorithm and second to extend the analysis to gene expression data. In our first aim we propose to use Takens’ theorem to show that under idealized conditions the point cloud defined by the sliding window algorithm is a faithful representation of the original data and therefore findings obtained when analyzing the point cloud are applicable to the original data. Second we apply our method to a breast cancer gene expression data set. Our approach assumes that gene expression is to some extent also a measure of the underlying copy number changes [16,23]. We therefore first order genes according to their physical location in the genome; second we apply the sliding window algorithm to generate a set of point clouds associated to the data and third perform topological and statistical analysis of the point clouds. We show that this algorithm, when only $\beta_0$ is used, can distinguish between breast cancer subtypes. In particular we observe remarkable differences between the less aggressive subtypes (i.e. normal-like and luminal A) and the more aggressive ones (i.e. luminal B, basal-like and Her2). Our results also show that luminal B can not be distinguished from Her2 and basal-like suggesting that luminal B has characteristics similar to both subtypes. These findings suggest that breast cancer subtypes can be characterized, not only by