Topological Features in Cancer Gene Expression Data

PDB, 2015; arXiv: 1410.3198

Washington State University

Bala Krishnamoorthy

Svetlana Leekwood
Cancer Gene Expr. Data

Expression values, \( \geq 0 \)

Genes, 10Ks

Samples
CANCER GENE Expr. DATA

samples

expression values

~50,000 genes

10s to a few 100 samples

0, 10K's, 10K's, 10K's
Select a few biologically relevant genes
10s to a few 100 samples
~50,000 genes

Cancer Gene Expr. DATA

Expression values, $\geq 0$
Genes, 10Ks
too many degrees of freedom

* The Challenge
The Challenge

(Beyer et al., 1999)

* Concentration of measure in high dim.
* Too many degrees of freedom
The Challenge

- might not work
- PCA, ... clustering, PC
- (Behar et al., 1999) concentration of measure in high dim.
- too many degrees of freedom
- *
 Higher order method

 "... might not work"

 (Beher et al., 1999)

 * concentration of measure in high dim.

 * too many degrees of freedom

 The Challenge
Our Result — genes in sample space

dualize the data

\[\text{Our Result} \]
Use persistent homology to find loops (≡ "holes") in the data. Our results suggest that genes in samples space dualize the data.
Our Result

- "dualize" the data
  - genes in samples space

- use persistent homology to find loops (≡ "holes")

- genes forming loops implicated in cancer
* a method for data exploration...

"dualize" the data

- genes in sample space

use persistent homology to find loops (\( \equiv \) "holes")

genes forming loops implicated in cancer
2D illustration
Higher-Order Structures

Local Structure

Traditional Approach

2D Illustration
Higher Order Structures

- Local Structure
- Higher Order (Hop)
- Miss Higher Order
- 2D Illustration
expression values, $\geq 0$  

<table>
<thead>
<tr>
<th>Samples</th>
<th>Genes, IOK's</th>
</tr>
</thead>
</table>

instead of *  

HI-DIM : DUAL SPACE
HI-DIM: DUAL SPACE
| Samples | Genes | Gene expressions considered | across patients | use | * | * | * |

HI-DIM: DUAL SPACE
HI-DIM: DUAL SPACE

- use
- gene expressions considered across patients
- pairwise distances are much more meaningful
Persistent Homology

Characterizes significant, multidimensional "holes"
PERSISTENT HOMOLOGY

- Characterizes significant high-dimensional "holes"

- Points → pairwise distances → simplicial complex → filtration → persistence diagrams/barcodes
- dim 0: connected components
- dim 1: loops (around holes)
- dim 2: enclosed voids

Persistence diagrams/persistent homology
- filtration
- persistence diagrams/persistence complexes
- points
- simplicial complexes

Characterizes significant h-dimensional "holes"

Persistent Homology
Persistent homology

* characterizes significant multidimensional "holes"

* points -> pairwise distances -> simplicial complex

* filtration -> persistence diagrams/barcodes

(from Ghrist, 2008)
- edge \( [i, j] \) is in complex if \( E \subseteq D \) 

s.t. \( l_i, l_j \) are 2 nearest neighbors of \( w \)

- size of complex grows fast w.r.t. points in data \( D \)

**Witness Complex**

de Silva & Carlsson, 2004
Witness Complex

de Silva & Carlsson, 2004
- \( n \) is the witness for the p-simplex
- p-simplex \([x_0, \ldots, x_p]\) is in complex \( \mathcal{F} \subseteq \mathcal{E} \)
  - s.t. \( \forall e \in \mathcal{E} \), \( \forall e' \in \mathcal{E} \) \( e \neq e' \), \( p \) and \( p' \) are 2 nearest neighbors of \( v \) and \( v' \), respectively.
- \( \mathcal{F} \subseteq \mathcal{E} \) is in complex \( \mathcal{F} \subseteq \mathcal{E} \)
- define complex on \( \mathcal{L}_C \) subset of landmarks
* witness complex grows fast w.r.t. points in data

\( \) de Silva & Carlsson, 2004

WITNESS COMPLEX
METHOD

* for breast cancer (54,613 genes, 47 samples)
  dim-1 barcode
# Method

* For breast cancer (54,613 genes, 47 samples)

- dim-1 barcode

- # genes = 70
# genes = 110

# genes = 70

dim-1 barcode

* for breast cancer (54,613 genes, ~7 samples)

METHOD
# genes = 130

# genes = 110

# genes = 70

dim. 1 barcode

dim. 1 barcode for breast cancer (54,613 genes, 47 samples)

METHOD
# genes = 110

\[ L = 110 \]

* Pick the longest loop(s)

* Dim-1 barcode (for breast cancer (54,613 genes, 47 samples))

METHOD
* Are genes in loops irrelevant for cancer?

0 genes = 110

* Pick the longest loop(s)

dim. 1 barcode

* For breast cancer (54,613 genes, 47 samples)

METHOD
<table>
<thead>
<tr>
<th>Gene</th>
<th>Relation to Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSP48</td>
<td>Coordinate p53 signaling carcinomma cells</td>
</tr>
<tr>
<td>RPS27A</td>
<td>Downregulated in apoptotic breast cancer</td>
</tr>
<tr>
<td>B2RP51</td>
<td>Prognostic biomarker in breast cancer</td>
</tr>
</tbody>
</table>

**METHOD**

genes in the breast cancer layer:
- dim-1 barcode
- for breast cancer (54,613 genes, 47 samples)
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Genes</th>
<th>Samples</th>
<th>#loops</th>
<th>#samples</th>
<th>#loops</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML170</td>
<td>54613</td>
<td>188</td>
<td>1</td>
<td>46</td>
<td>46201</td>
</tr>
<tr>
<td>AML188</td>
<td>54613</td>
<td>28</td>
<td>1</td>
<td>47</td>
<td>46201</td>
</tr>
<tr>
<td>Ovarian Breast Brain</td>
<td>54613</td>
<td>28</td>
<td>1</td>
<td>47</td>
<td>46201</td>
</tr>
</tbody>
</table>

Results

* Analyzed five different cancer datasets.
The majority of loop genes implicated in cancer in all cases are included.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Genes</th>
<th>Samples</th>
<th>#Loops</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML170</td>
<td>12558</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>AML188</td>
<td>54613</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>Ovarian</td>
<td>54613</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>54613</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Brain</td>
<td>46201</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

**RESULTS**

Five different cancer datasets were analyzed.
Genes do not have extreme expression values selected landmarks (L), as well as loop majority of loop genes implicated in cancer in all cases analyzed five different cancer datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Genes</th>
<th>#Samples</th>
<th>#loops</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML170</td>
<td>122558</td>
<td>170</td>
<td></td>
</tr>
<tr>
<td>AML178</td>
<td>54613</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>Ovarian Breast Brain</td>
<td>54613</td>
<td>28</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>46201</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

RESULTS *
Open Questions

* Small groups (6-13) of genes forming loops cannot be found by other methods
Open Questions

* Does loop connectedness of genes imply functional connectedness?
* Small groups (6-13) of genes forming loops cannot be found by other methods.
Open Questions

* Does loop connectedness of genes imply functional connectedness?

- hard to study compression of multiple genes

* Small groups (6-13) of genes forming loops cannot be found by other methods
Does chronological order affect ability to prove result on structure/stability of data?

Does dualization affect ability to prove

— hard to study coexpression of multiple genes

imply functional connectedness?

Does keep connectedness of genes

loops cannot be found by other methods

small groups (6-13) of genes forming

Open Questions
* a few relevant genes not included in loops

Open Questions
A few relevant genes not included in loops

Open Questions
Open Questions

Can we identify loops with "all critical genes"?

* a few relevant genes not included in loops
* Apply to other classes of data sets?

* Can we identify loops with "all critical genes"?

* A few relevant genes not included in loops

OPEN QUESTIONS