Topological Approach to Data Sets of Proteins

Peter Petrov, Anastas Pashov
IMI, BAS, IMicB BAS

The Plan

- Proteins , antibodies, epitopes
- Persistent homology in two variants
- Point cloud persistent homology
- Longest common subsequence distance

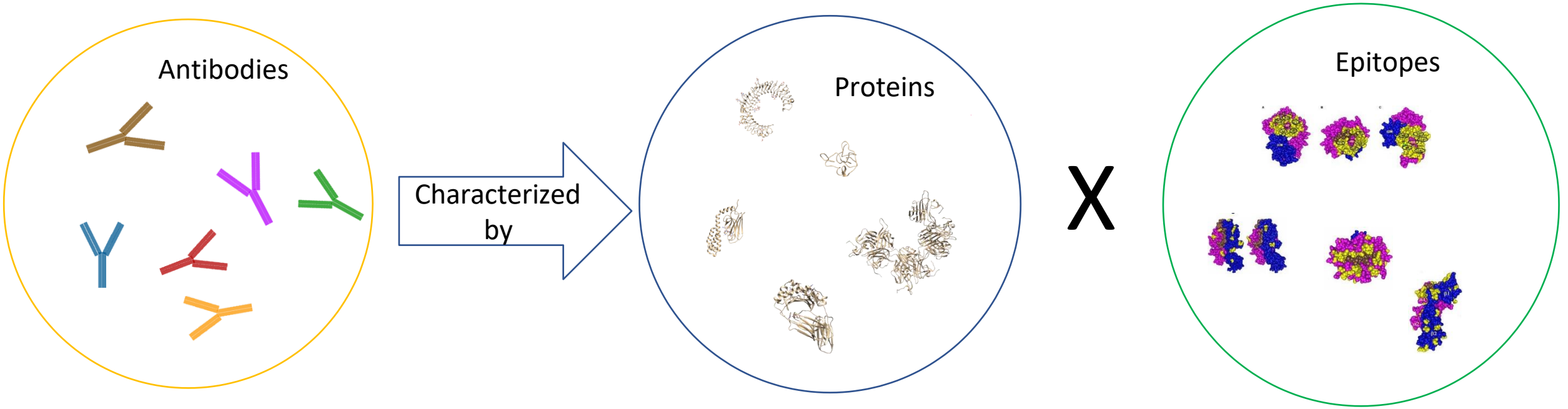
Why Topological Data Analysis?

- Describes the general shape of multidimesional data clouds
- Persistent homology is metric insensitive
- Does not require specific parametrization

Big data is typically multidimensional.
TDA helps reduce the dimension which helps the visualization.

Proteins , antibodies, epitopes

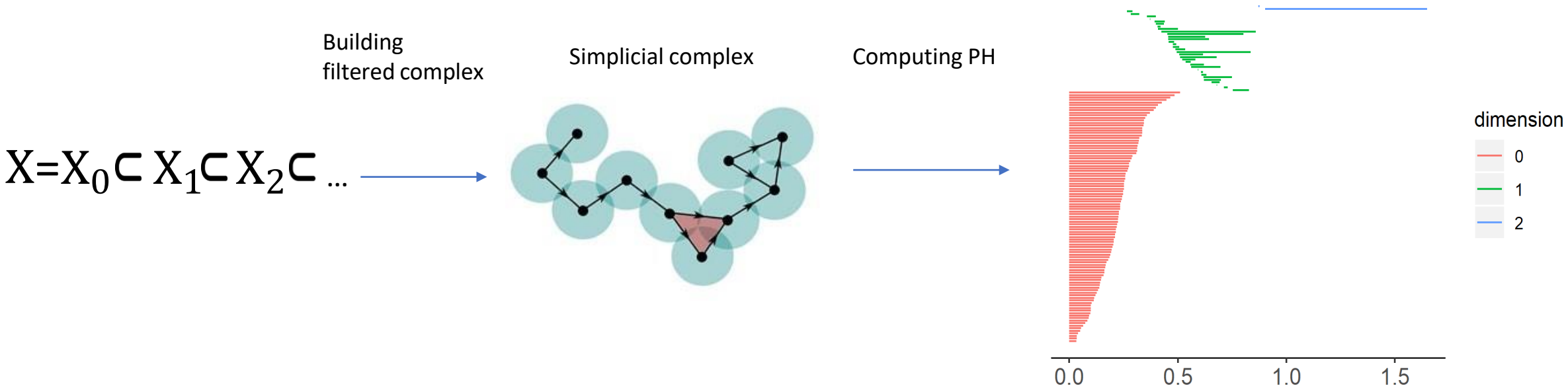
Igome – A global view of the antigen-antibody interactions through the geometry of the “epitope space”.



Persistent Homology – Version 1

The input is an increasing sequence of spaces: $X_0 \subset X_1 \subset X_2 \subset \dots$

The output is a barcode:



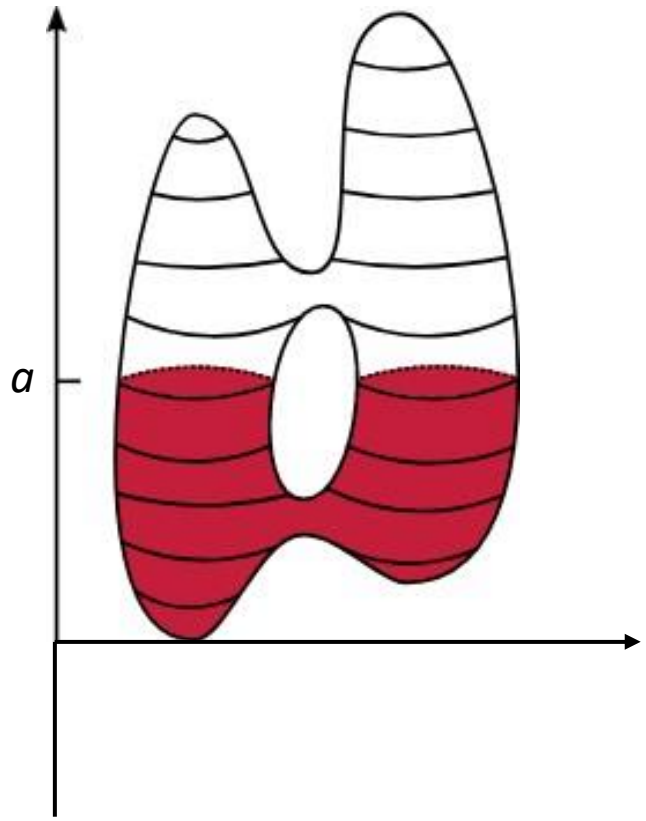
Persistent Homology – Version 2

The input is a space X, a function (energy value) $ff: \mathrm{X} \longrightarrow \mathbb{R}$ and for any **a** $\mathbb{XX}_a \coloneqq \{\mathrm{x} \in \boldsymbol{XX} : f(\mathrm{x}) \leq a\}$**,** the sublevel set
The output is a barcode.

Stability theorem.

$ff, g: \mathrm{X} \longrightarrow \mathbb{R}$, with the corresponding barcodes $U_{ff}, U_{gg} \Rightarrow$
$d_B\left(U_{ff}, U_{gg}\right) \leq \|ff - g\|_\infty \coloneqq max_{x \in \mathbb{X}} \ |ff(x) - g(x)|.$

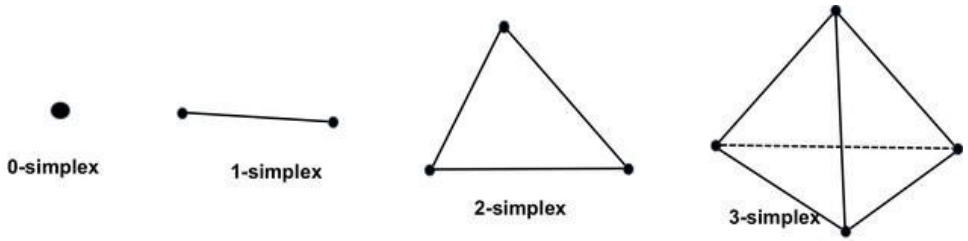
That is, if the spaces are similar then the barcodes are, too.



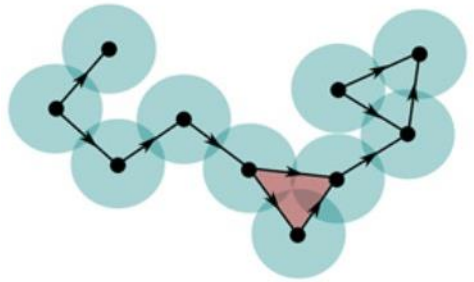
To compare 2 data clouds, we can produce their barcodes and measure the distance between them.

Point cloud persistent homology

Simplices:



$S = \{x_1, \ldots x_n\} \subset R^N$, we thicken *S* by balls of radius *r*,
$B_{ii}(x_{ii}, r)$, for any *i* and in that way we create a simplicial complex,



the homologies of which are given by its barcode.
A choice of *r* is called a resolution: $r = 0, \delta\delta_1, \delta\delta_2 \ldots$.
Usually $\delta\delta_{ii} = iix\delta\delta_1$.

S —*thicken + VR complex*→ Simplicial complex —*Cycles/boundaries*→ Persistent homology

Longest common subsequence distance

LCS is an edit distance with insertion and deletion as the only edit operations both at a cost of 1.
Ex.: BEST -> LEST -> LOST, so LCS(BEST, LOST)=4.
When the length of the strings is fixed the LCS distance is always even.
Therefore, without loss of generality in this case LCS is divided by 2.

Any edit distance with non-negative cost is a metric, so is LCS distance.

THANK YOU FOR YOUR ATTENTION!

Funding from the National Science Fund, Bulgaria (grant #: КР-06 –N 21/14.2018 and КР-06 –N 37/24.2019).