Protein Folding Characterization via Persistent Homology

Marcio Gameiro

University of São Paulo @ São Carlos
Rutgers University

Workshop on Topology: Identifying Order in Complex Systems

IAS - April 7, 2018

Collaborators

Konstantin Mischaikow (Rutgers)
Shaun Harker (Rutgers)

Protein Folding

Protein folding is the process by which a protein (a chain of amino acids) folds in its 3-dimensional structure.

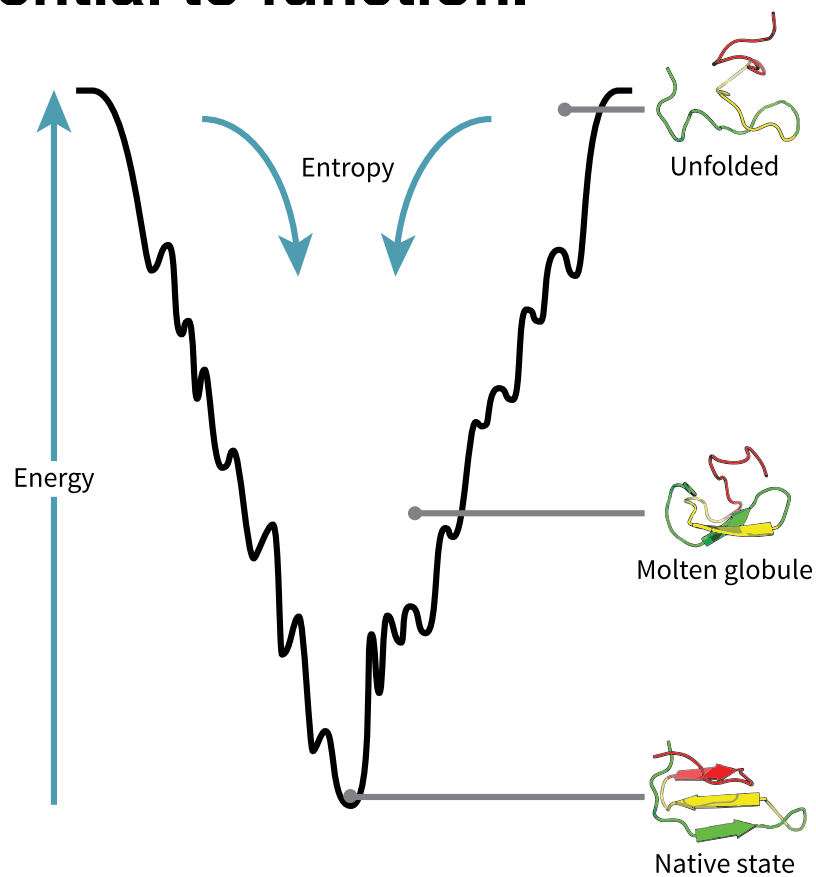
This folded protein is known as the native state of the protein.

The correct 3-dimensional shape is essential to function.

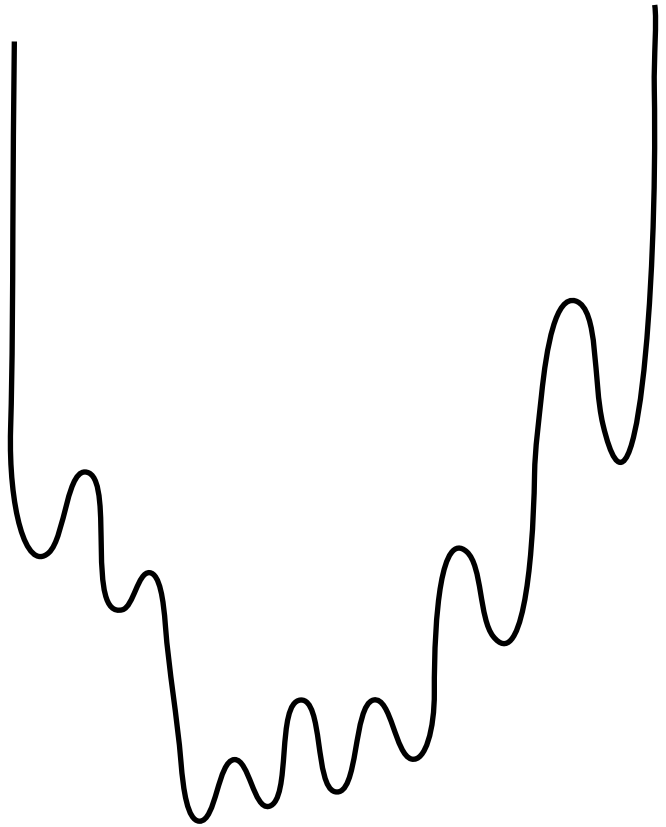
Two very challenging problems in synthetic protein design is to predict stability and the native state of a protein from its amino acid sequence.

In this talk we will focus on protein folding.

Computer simulations use energy landscapes.



Protein Folding



Dataset of predicted protein structures from Sagar Khare's lab @ Rutgers

Computations with Rosetta and Amber software packages

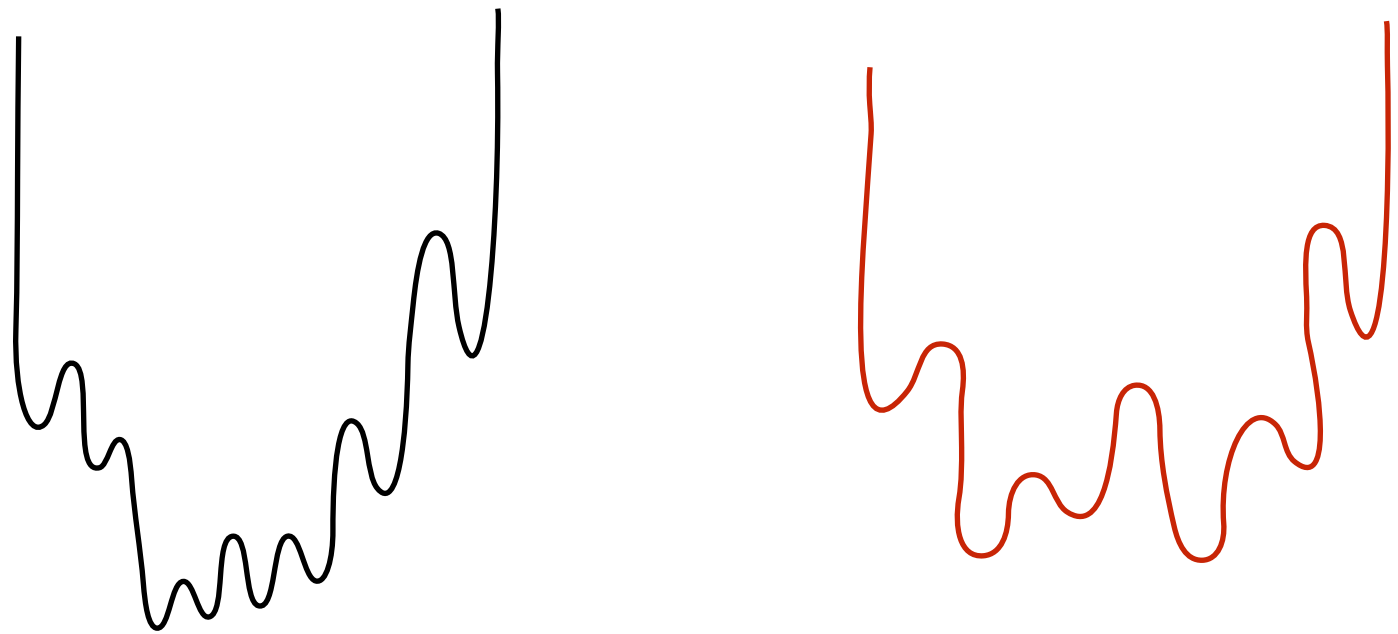
Root mean square deviation (RMSD) is computed by comparing predicted structure with native protein

Many computations fail to produce a minimum energy configuration (local minima?)

Many computations produce a minimum energy configuration with large error (false minima)

Protein Folding

Is the predicted structure a valid folding not observed or is the error due to imprecisions in the energy functional?



We consider are small (around 100 amino acids). So errors are more likely due to imprecisions in the energy functional.

How can we try to identify errors in the energy functional?

Protein Folding

Local energy between amino acids are well understood.

Major sources of imprecisions are likely due to global structures of the protein.

If so, can we identify particular global geometric structures that give rise to these imprecisions?

We use topology to measure global structure (shape) of the proteins.

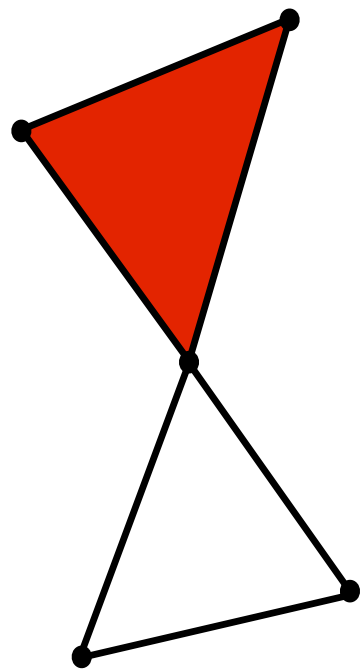
Use persistent homology to search for correlations between global geometric structures and the failure to predict the desired folding

This could help guide the energy minimization techniques to the correct minimum

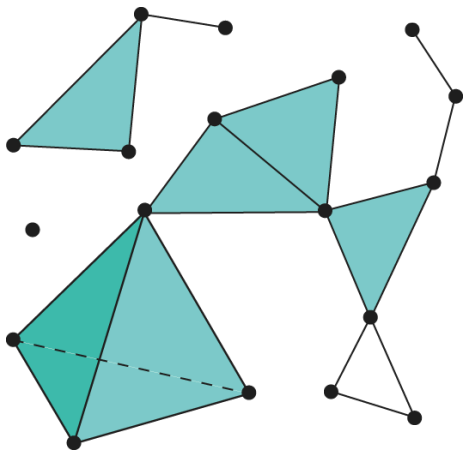
Homology

- Homology: $H_\ell(X) \longrightarrow \beta_\ell =$ number of ℓ-dim "holes"
- $H_\ell(X) = Cycles/Bounday$
- β_ℓ are called Betti numbers

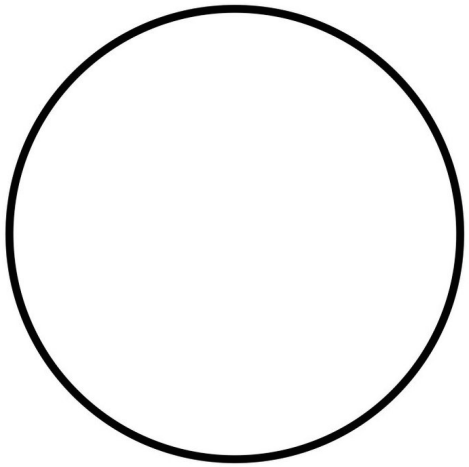
- $\beta_0 =$ number of connected components
- $\beta_1 =$ number of cycles (or tunnels in 3D)
- $\beta_2 =$ number of connected cavities



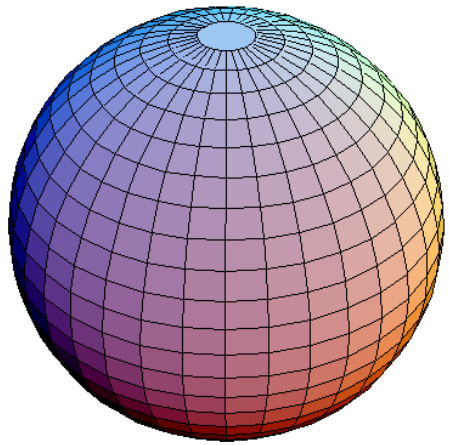
Homology



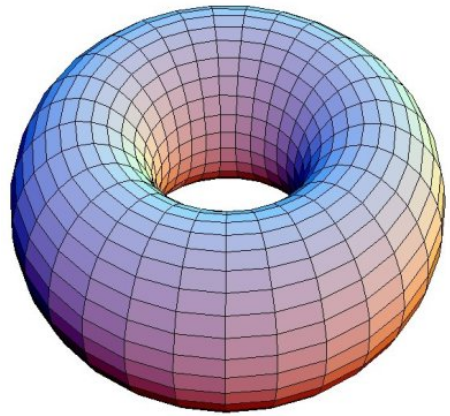
$\beta_0 = 3$
$\beta_1 = 1$
$\beta_2 = 0$



$\beta_0 = 1$
$\beta_1 = 1$
$\beta_2 = 0$

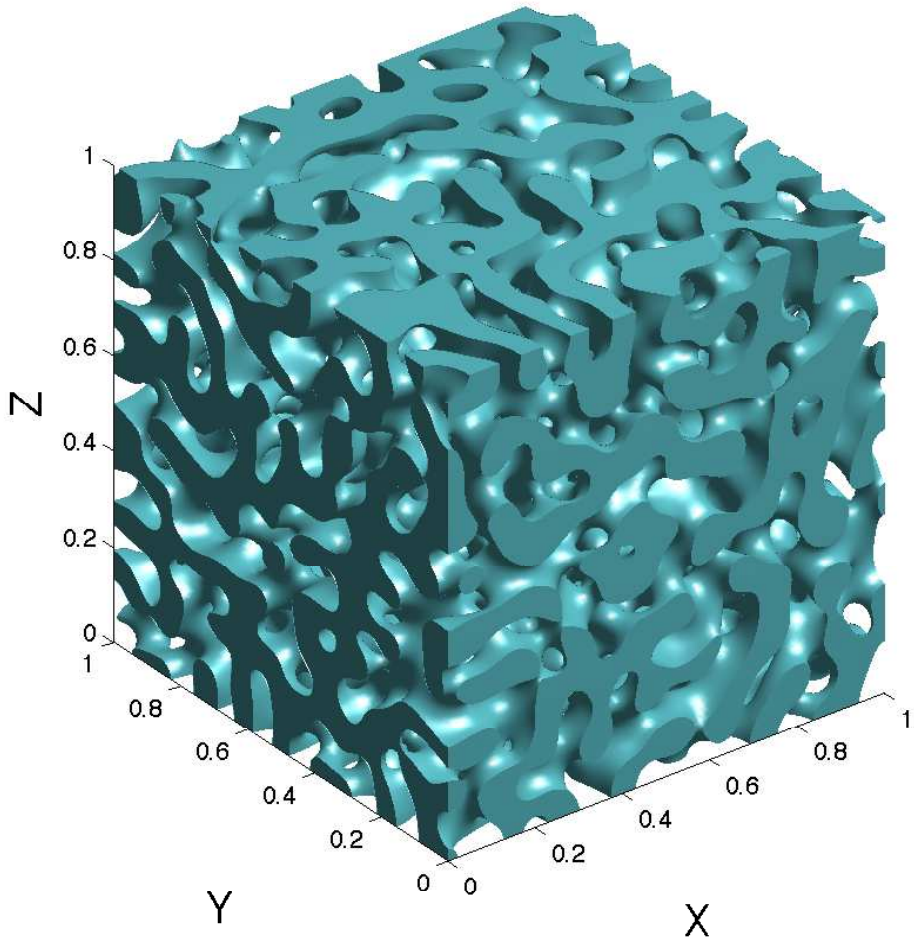


$\beta_0 = 1$
$\beta_1 = 0$
$\beta_2 = 1$



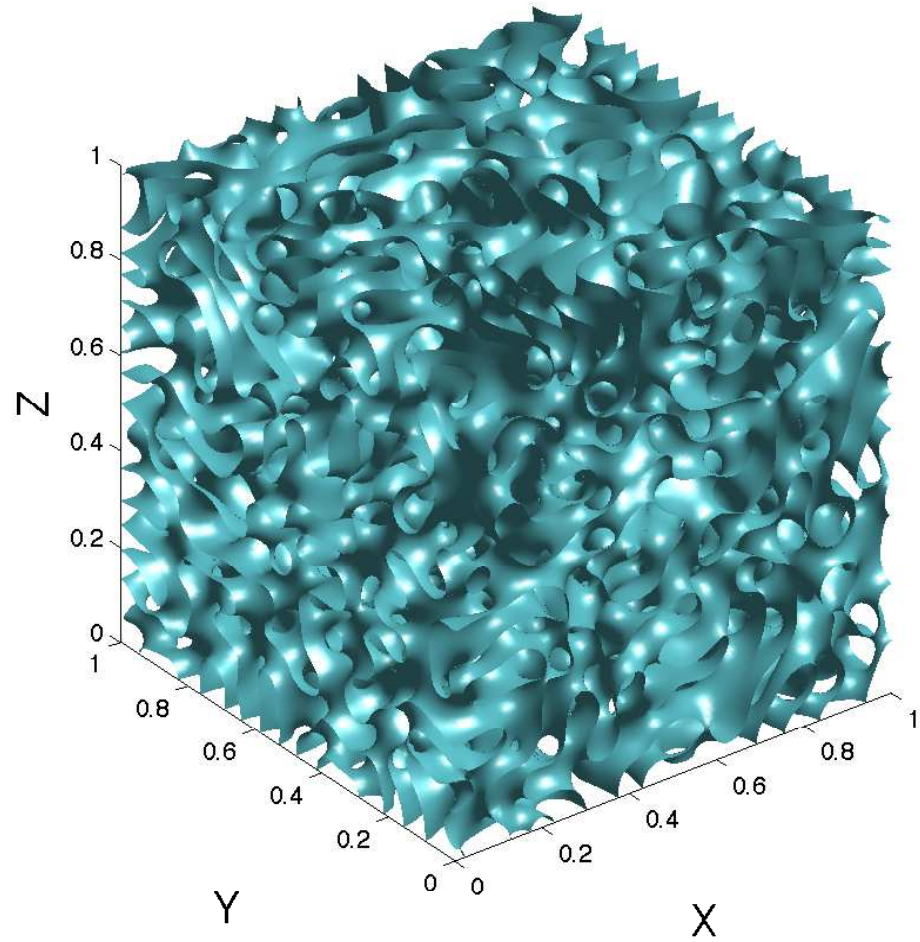
$\beta_0 = 1$
$\beta_1 = 2$
$\beta_2 = 1$

Homology



Region $u \geq 0$

$\beta_0 = 1,\ \beta_1 = 847,\ \beta_2 = 0$



Isosurface $u = 0$

$\beta_0 = 1,\ \beta_1 = 1705,\ \beta_2 = 0$

Persistent Homology

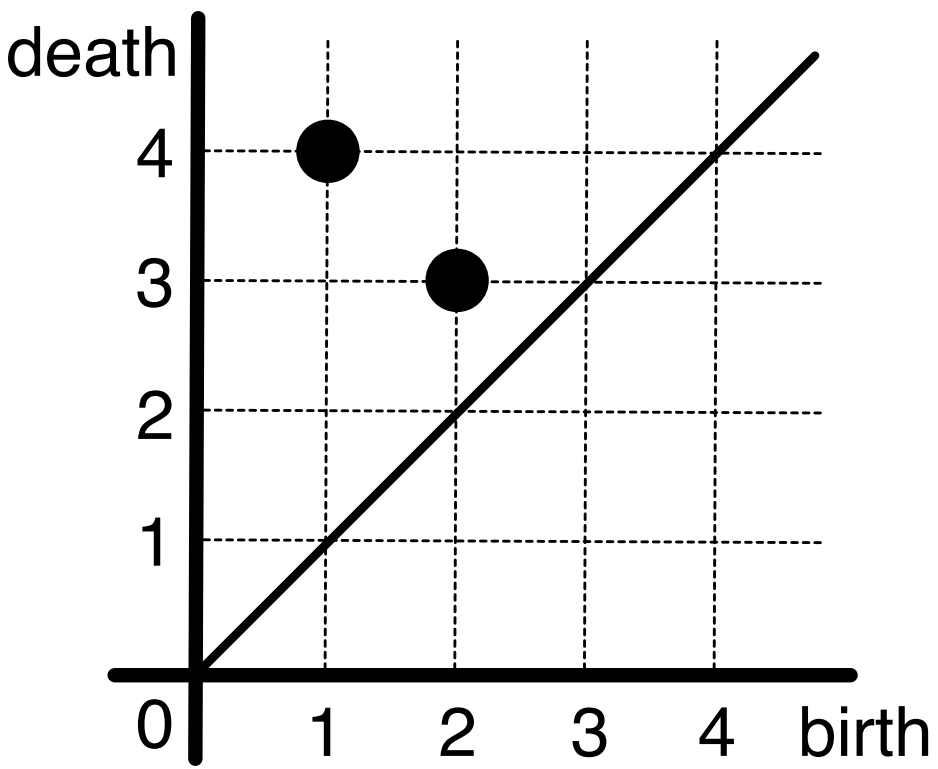
- Filtration $\mathcal{X}$: $X_1 \subset X_2 \subset \cdots \subset X_n$

- Persistent Homology: $PH_\ell(\mathcal{X}) \longrightarrow$ Metric info. about ℓ-dim holes
- Information about when holes appear and for how long they persist

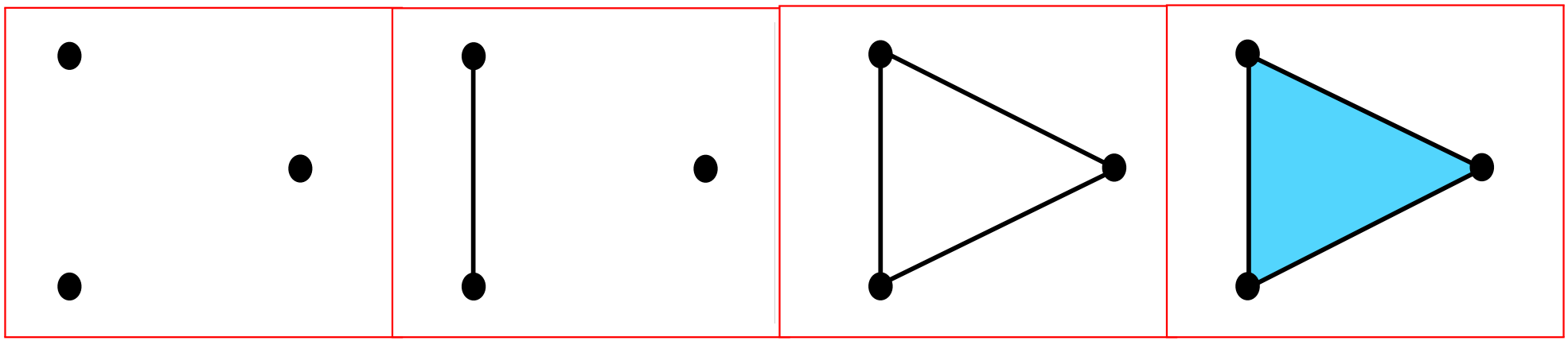
`Perseus` and `PHAT` provide fast software to compute persistence

Persistence Diagrams

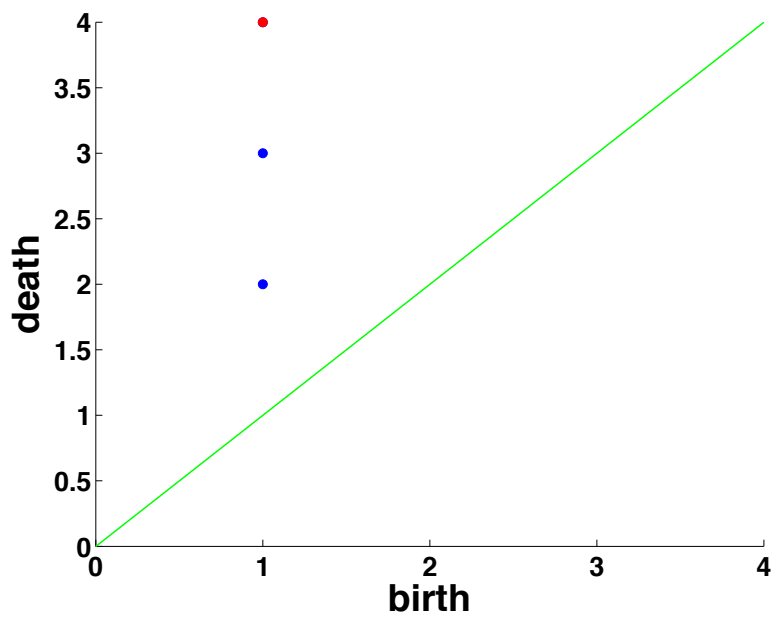
- Persistence Diagram (PD): $PD_\ell = \{(b_i, d_i) \mid i = 1, \ldots, n\}$
- Each pair (b_i, d_i) represents and ℓ-dimensional hole
- (b_i, d_i) is called a birth-death pair



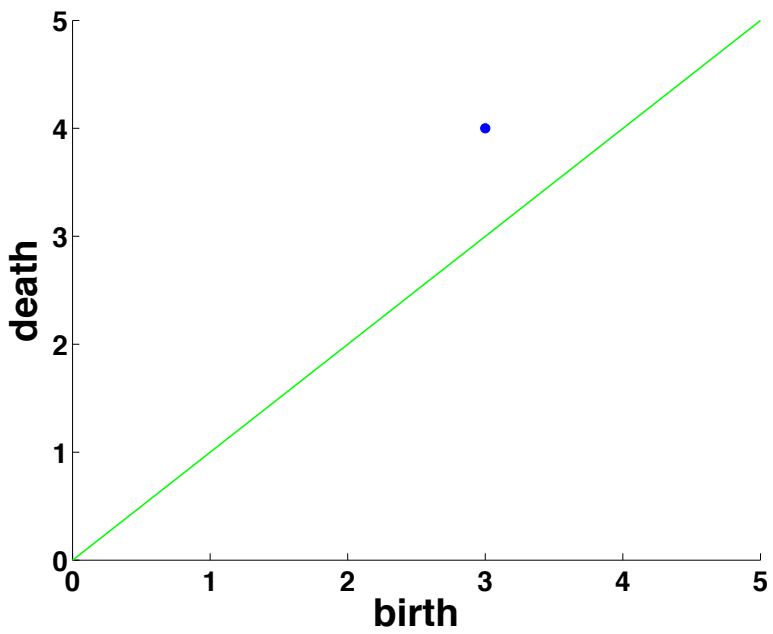
Persistence Diagrams



X_1 X_2 X_3 X_4



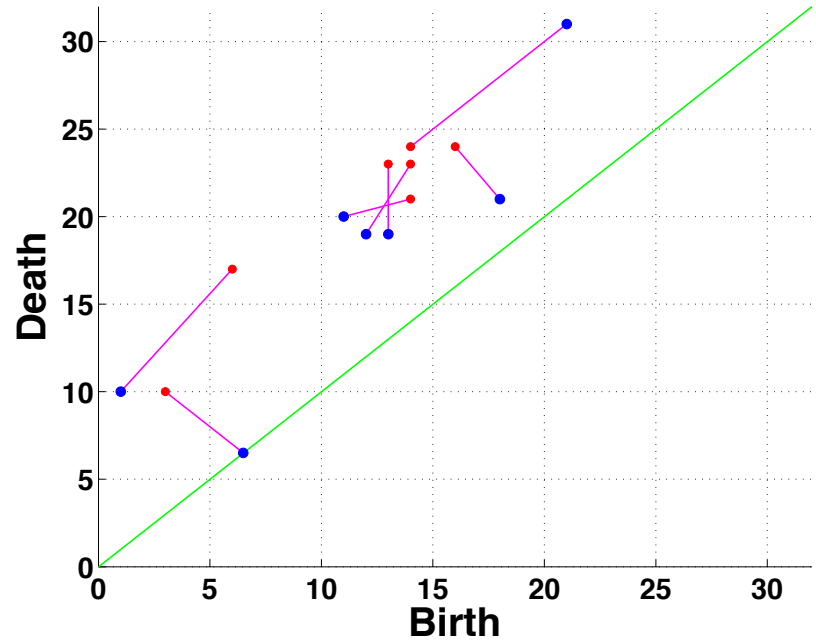
PD_0



PD_1

Stability of Persistence Diagrams

$\gamma : PD \to PD'$ bijection



Bottleneck Distance

$$d_B(PD, PD') = \inf_{\gamma} \sup_{x \in PD} \|x - \gamma(x)\|_\infty$$

PD is a continuous map

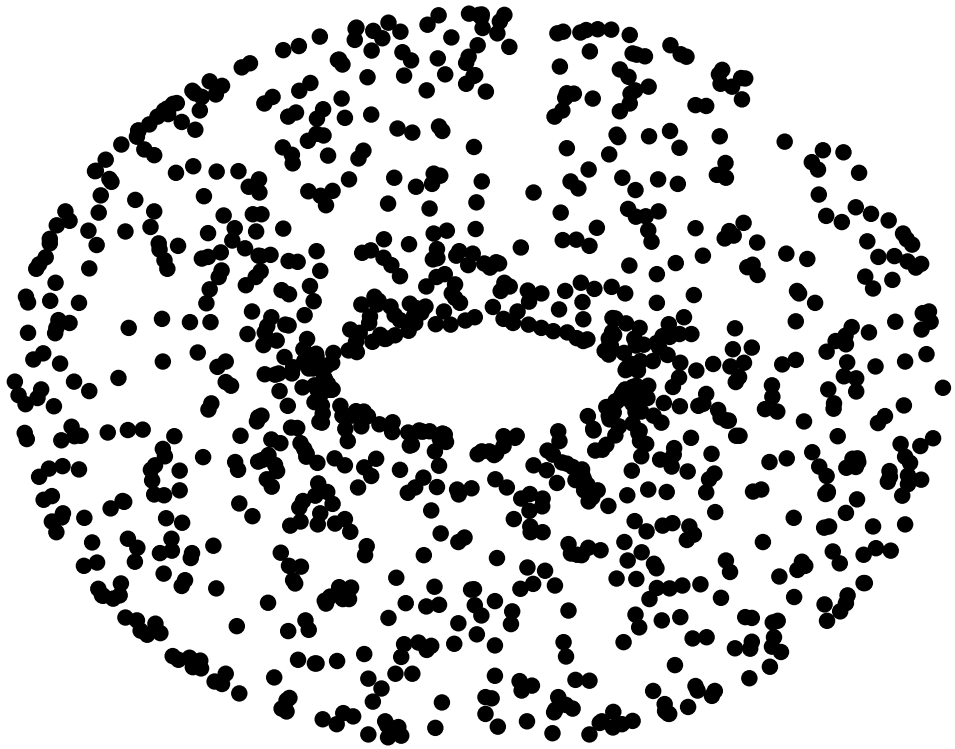
Stability of Persistence Diagrams

David Cohen-Steiner , Herbert Edelsbrunner , John Harer

Discrete & Computational Geometry 2007, Volume 37, Issue 1

Persistence Diagrams of Point Clouds

- How to compute persistence from point cloud data?

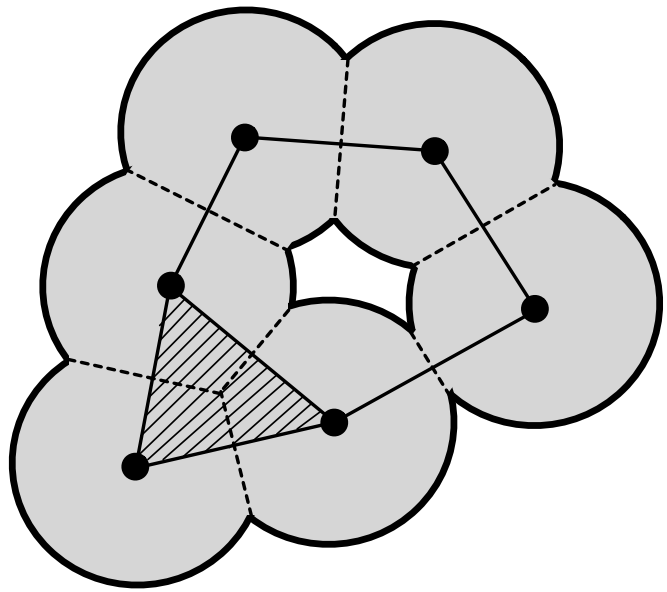


- Need to build a cell complex!

Alpha Shapes

- $X = \{x_i \in \mathbb{R}^n \mid i = 1, \ldots, n\}$
- $\mathbb{R}^n = \bigcup_{i=1}^{n} V_i$: Voronoi decomposition
- $\bigcup_{i=1}^{n} B_i(r) = \bigcup_{i=1}^{n} (B_i(r) \cap V_i)$

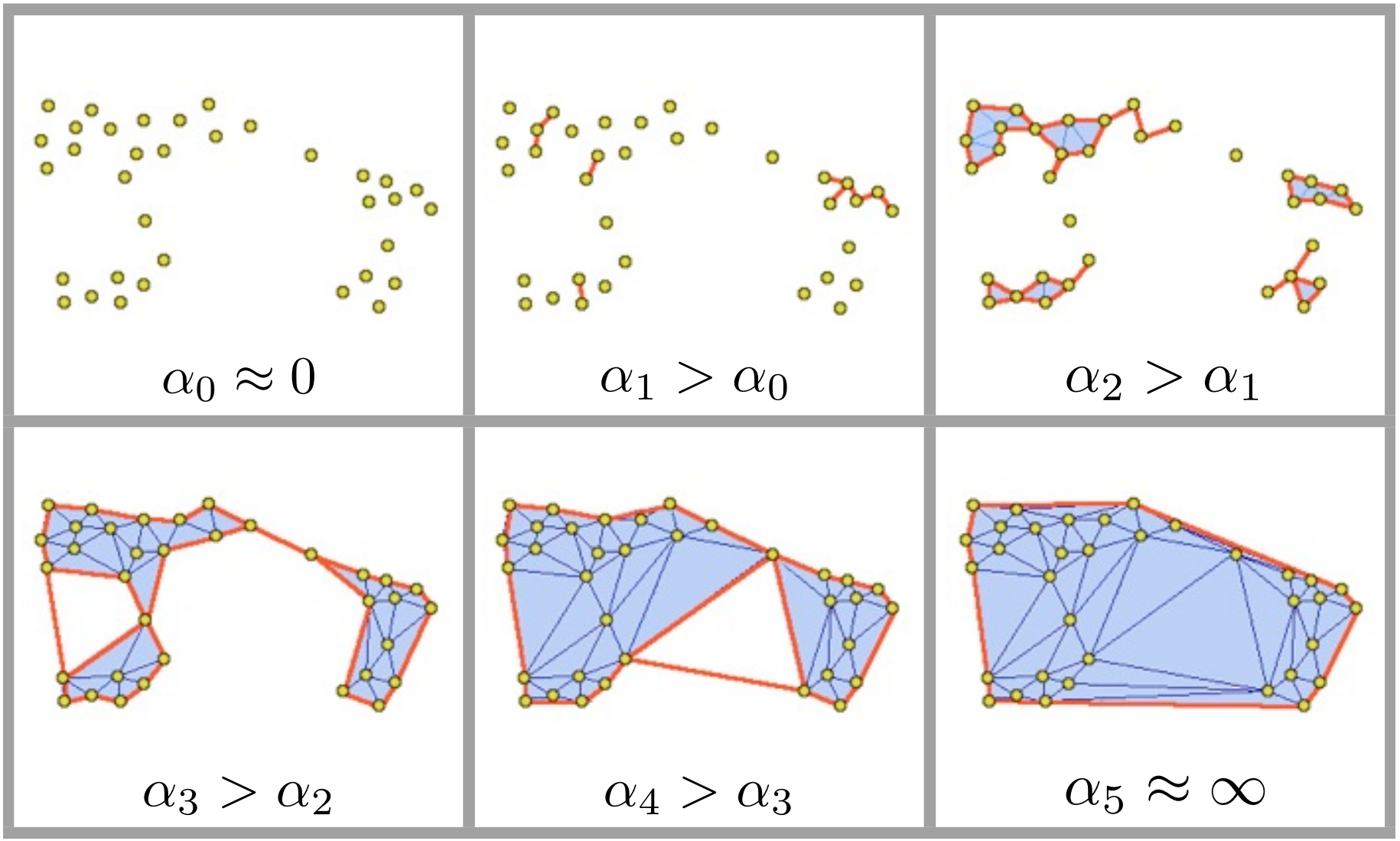
  $B_i(r) = \{x \in \mathbb{R}^n \mid \|x - x_i\| \leq r\}$



- Alpha Shape $\mathcal{A}(X, r)$: dual of $\{B_i(r) \cap V_i \mid i = 1, \ldots, n\}$ (simplicial complex)
- Homotopy Equivalence: $\bigcup_{i=1}^{n} B_i(r) \simeq \mathcal{A}(X, r)$ (same Homology)

Alpha Filtration

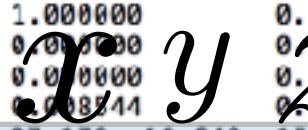
- $\mathcal{A}(X, \alpha_1) \subset \mathcal{A}(X, \alpha_2)$ for $\alpha_1 \leq \alpha_2$



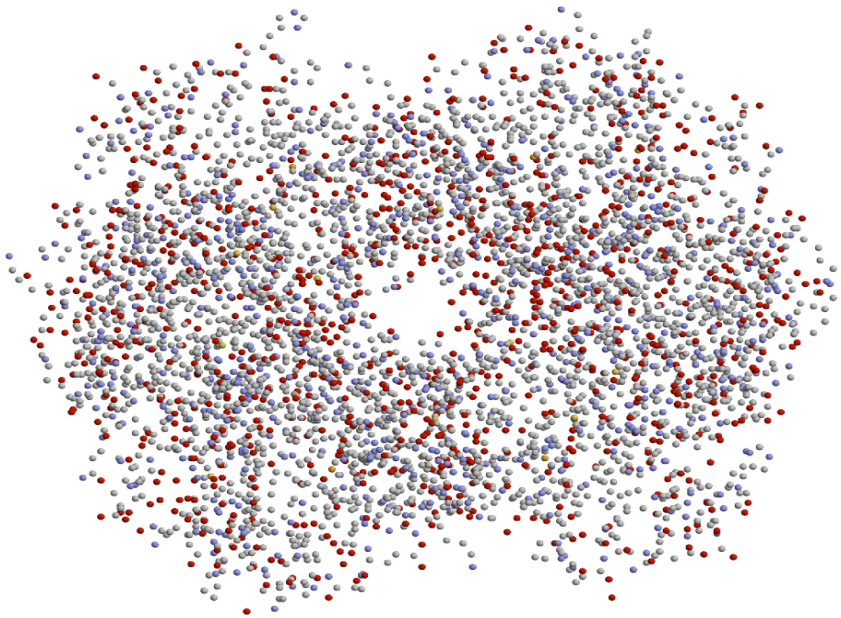
CGAL provides fast software to compute alpha shape

Protein Data (PDB)

```
SITE     1 AC1  6 ASP A  43  VAL A  46  GLY A  48  ASP A  50
SITE     2 AC1  6 SER A  52  HOH A2091
SITE     1 AC2  5 THR A 171  ASP A 222  THR A 225  ILE A 228
SITE     2 AC2  5 ASP A 230
SITE     1 AC3 25 ARG A  31  ALA A  34  SER A  35  ARG A  38
SITE     2 AC3 25 PHE A  41  SER A  73  PRO A 139  ALA A 140
SITE     3 AC3 25 PRO A 141  LEU A 148  PHE A 152  LEU A 156
SITE     4 AC3 25 GLY A 169  HIS A 170  PHE A 172  GLY A 173
SITE     5 AC3 25 LYS A 174  ASN A 175  GLN A 176  PHE A 179
SITE     6 AC3 25 PHE A 221  SER A 246  FMT A1310  HOH A2353
SITE     7 AC3 25 HOH A2354
SITE     1 AC4  5 ARG A  38  PHE A  41  HIS A  42  HEM A1307
SITE     2 AC4  5 HOH A2189
CRYST1   40.330   68.302  117.048  90.00  90.00  90.00 P 21 21 21    4
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.024795  0.000000  0.000000        0.00000
SCALE2      0.000000  0.014641  0.000000        0.00000
SCALE3      0.000000  0.000000  0.008544        0.00000
ATOM      1  N   GLN A   1      27.178  16.649  15.783  1.00 26.78           N
ATOM      2  CA  GLN A   1      26.756  15.666  15.822  1.00 24.87           C
ATOM      3  C   GLN A   1      25.529  16.146  17.589  1.00 23.01           C
ATOM      4  O   GLN A   1      25.581  17.171  18.267  1.00 23.79           O
ATOM      5  CB  GLN A   1      27.896  15.425  17.018  1.00 26.64           C
ATOM      6  CG  GLN A   1      29.093  14.680  17.252  1.00 28.43           C
ATOM      7  CD  GLN A   1      28.821  13.203  17.036  1.00 29.14           C
ATOM      8  OE1 GLN A   1      29.684  12.465  16.563  0.00 29.08           O
ATOM      9  NE2 GLN A   1      27.617  12.764  17.388  0.00 29.08           N
ATOM     10  N   LEU A   2      24.429  15.405  17.484  1.00 19.34           N
ATOM     11  CA  LEU A   2      23.215  15.765  18.204  1.00 17.22           C
ATOM     12  C   LEU A   2      23.460  15.514  19.688  1.00 17.26           C
ATOM     13  O   LEU A   2      24.194  14.597  20.050  1.00 19.22           O
ATOM     14  CB  LEU A   2      22.033  14.917  17.722  1.00 14.98           C
ATOM     15  CG  LEU A   2      21.687  15.012  16.232  1.00 14.35           C
ATOM     16  CD1 LEU A   2      20.440  14.184  15.950  1.00 12.62           C
ATOM     17  CD2 LEU A   2      21.454  16.464  15.839  1.00 14.43           C
ATOM     18  N   THR A   3      22.849  16.330  20.541  1.00 15.96           N
ATOM     19  CA  THR A   3      23.013  16.190  21.986  1.00 16.22           C
ATOM     20  C   THR A   3      21.703  16.503  22.707  1.00 15.16           C
ATOM     21  O   THR A   3      20.991  17.435  22.340  1.00 15.46           O
ATOM     22  CB  THR A   3      24.118  17.133  22.511  1.00 17.01           C
ATOM     23  OG1 THR A   3      24.193  17.036  23.938  1.00 20.44           O
ATOM     24  CG2 THR A   3      23.826  18.571  22.117  1.00 17.32           C
ATOM     25  N   PRO A   4      21.370  15.725  23.748  1.00 16.50           N
ATOM     26  CA  PRO A   4      20.135  15.926  24.515  1.00 16.92           C
```



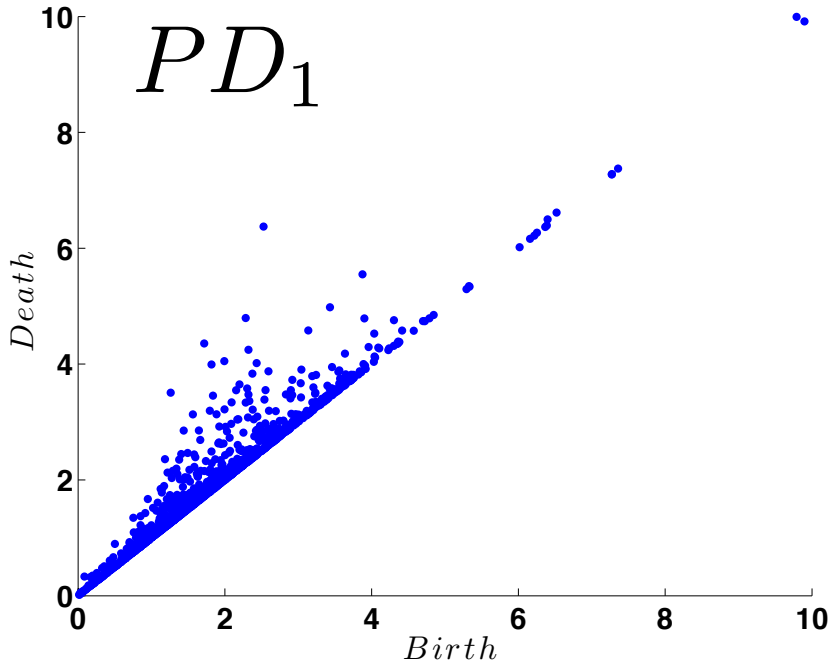
Persistence Diagrams of Protein Data

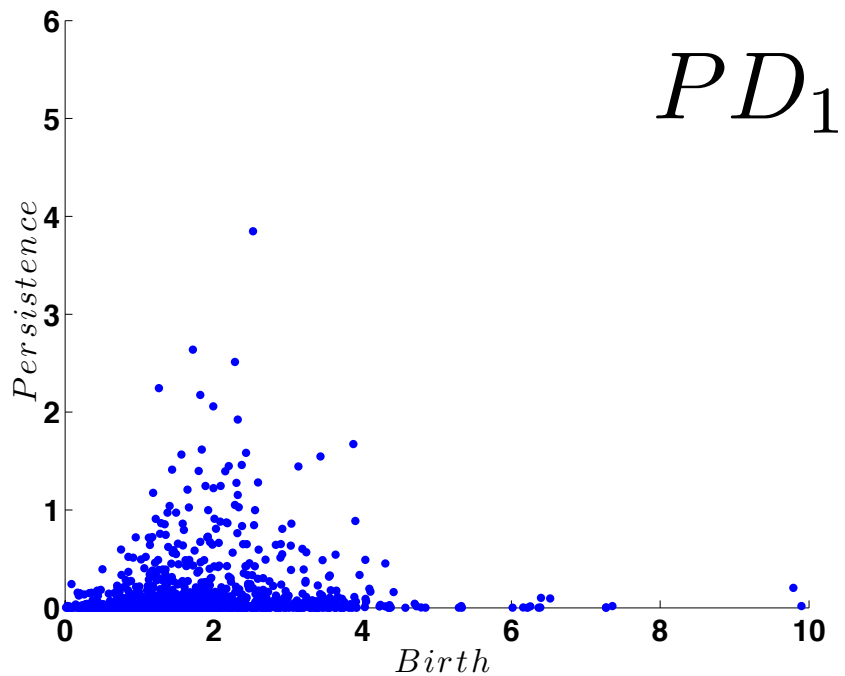


Hemoglobin

PDB id: 1A3N

Persistence = death - birth





Vectorized Persistence Diagrams

Vectorize using persistence image (Adams et al., 2017)

Persistence diagram (birth, persistence) pairs

$$PD_\ell = \{(b_i, p_i) \mid i = 1, \ldots, N\}$$

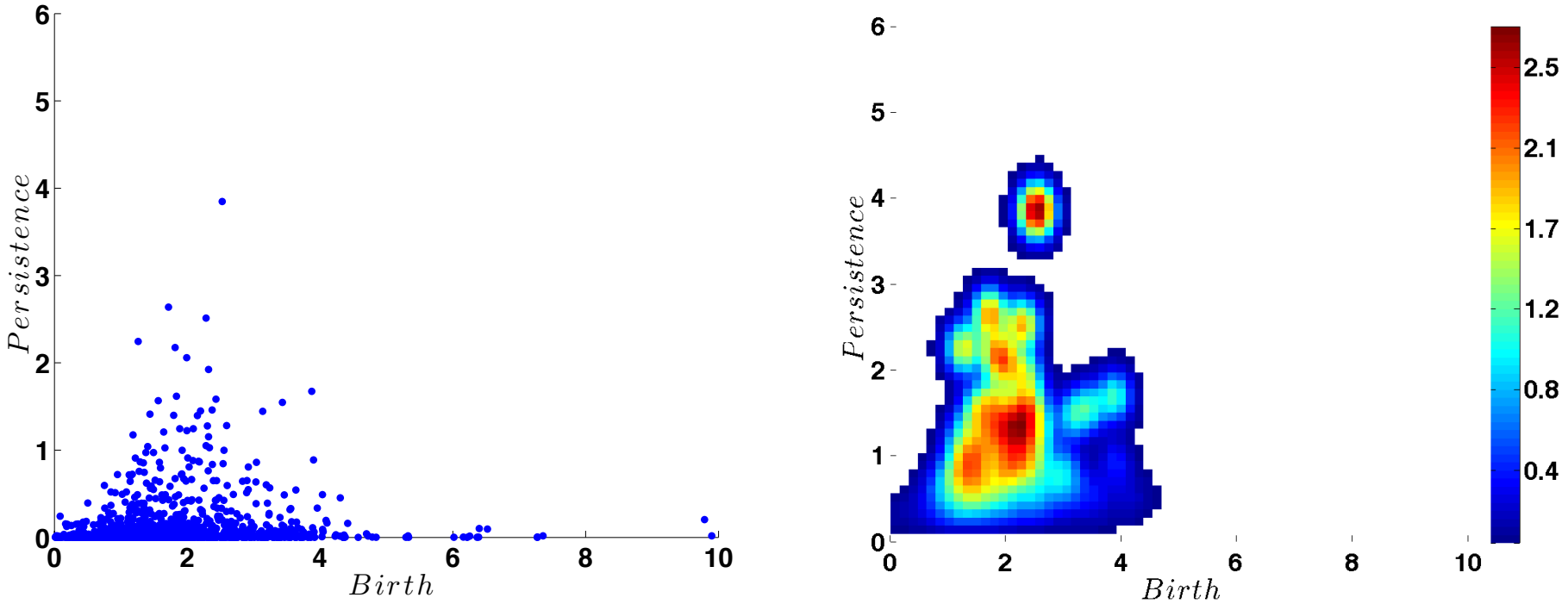
Consider

$$\psi(x, y) = \sum_{i=1}^{N} |p_i|^p \exp\left(-\frac{(b_i - x)^2 + (p_i - y)^2}{2\sigma^2}\right)$$

Evaluate this function on a grid on the (x, y) plane

Organize values as a vector (vectorized persistence diagram)

Vectorized Persistence Diagrams



Retains the continuity of persistence diagrams

Resolution and regions of diagrams to use are adjustable

Linear Regression

Linear regression model

$$y = wx + b$$

Vectorized persistence diagram

Error measure (RMSD)

The vector w is called the dual persistence diagram (Obayashi and Hiraoka 2017)

Solved by minimizing the mean squared error function

$$E = \sum_{n=1}^{N} (wx_n + b - y_n)^2 + \lambda \|w\|_2^2$$

Regularization term

using a training dataset of pairs (x_n, y_n)

Linear Regression

Linear regression with $\lambda\|w\|_2^2$ regularization is called Ridge

Linear regression with $\lambda\|w\|_1$ regularization is called Lasso

The Ridge regularization term is differentiable

Lasso regularization is know to produce sparse learned vectors

With Lasso we get a sparse dual persistence diagram

Useful to identify the main regions on the proteins responsible for the error

Predicted Structures Dataset

Data for 48 different proteins types (48 PDB ids)

Small molecules (around 100 amino acids each)

For each PDB id we have data for predicted structures and for the true native molecule

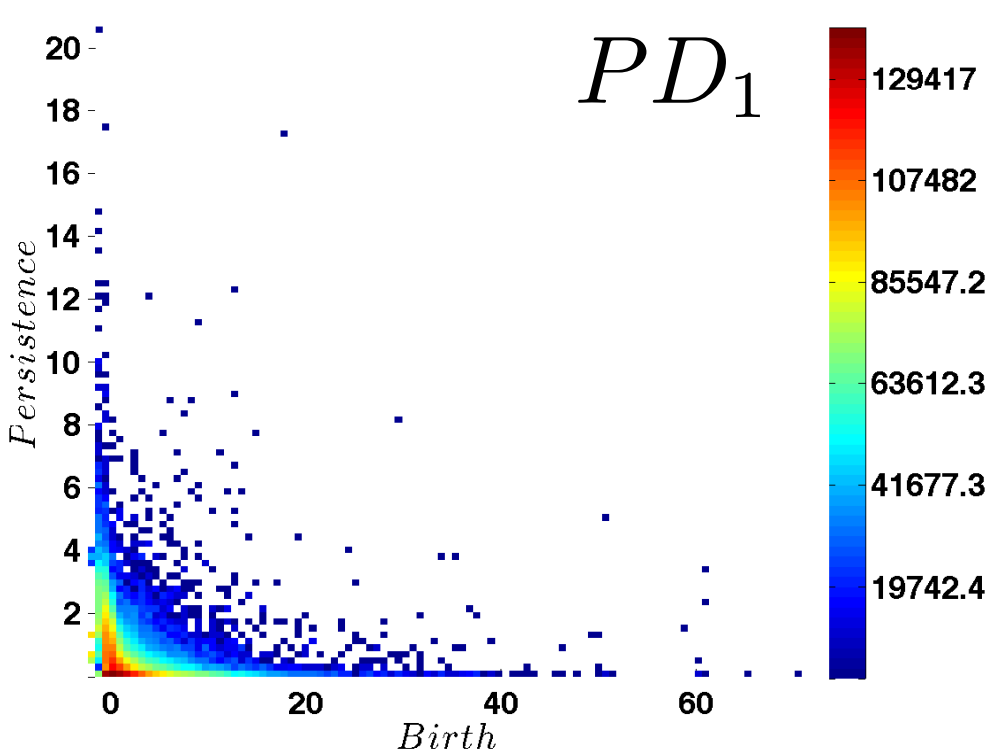
All predicted structure have low energy values (minima)

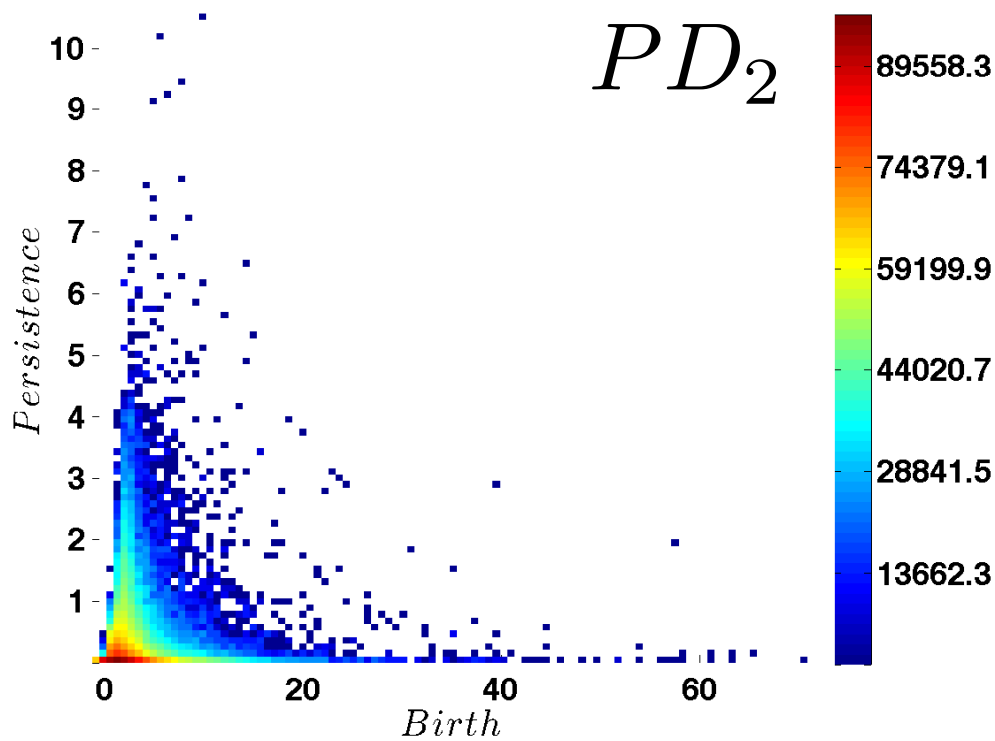
269 are false minima (predicted structures with low energy, but large error)

Protein Folding Analysis

Dataset: 400 persistence diagrams of predictions with small and large errors.

“Cumulative” density histogram of the dataset

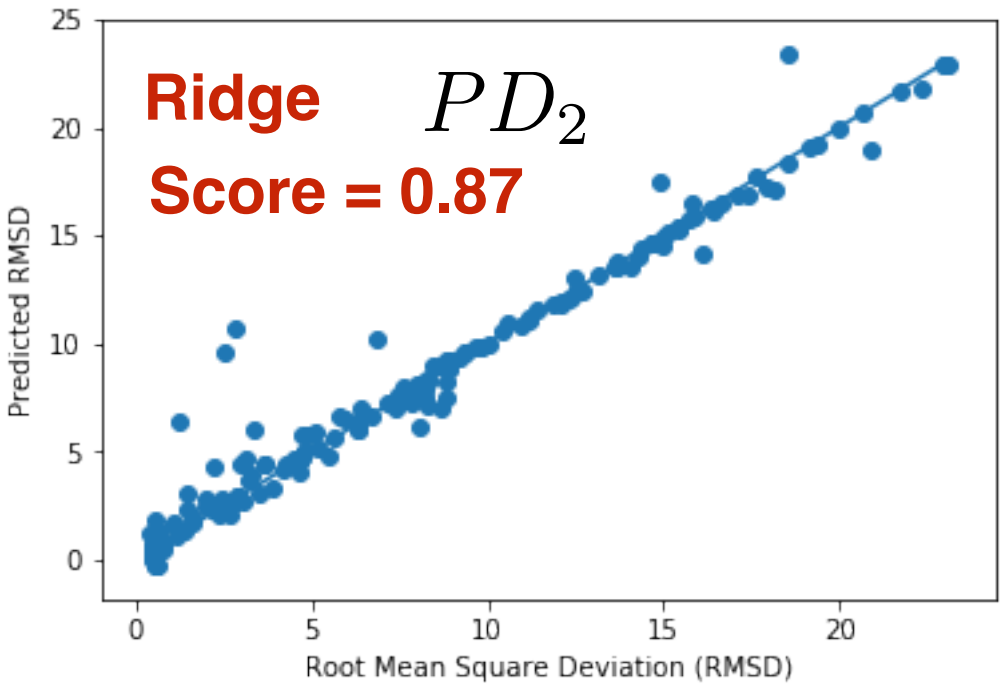


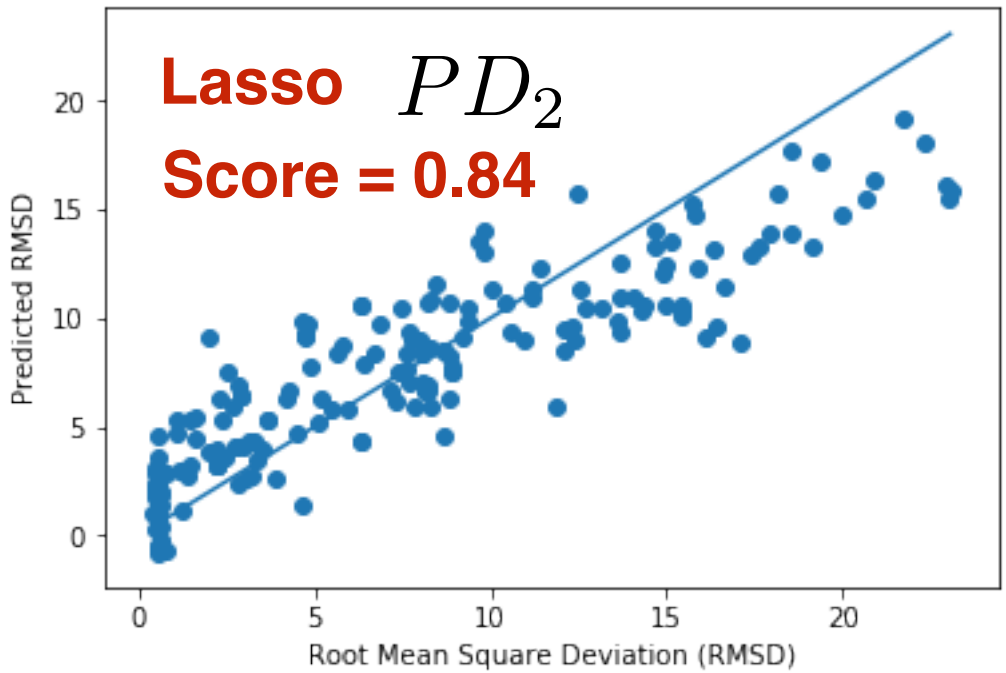


Protein Folding Analysis

Linear regression on dataset with 400 persistence diagrams

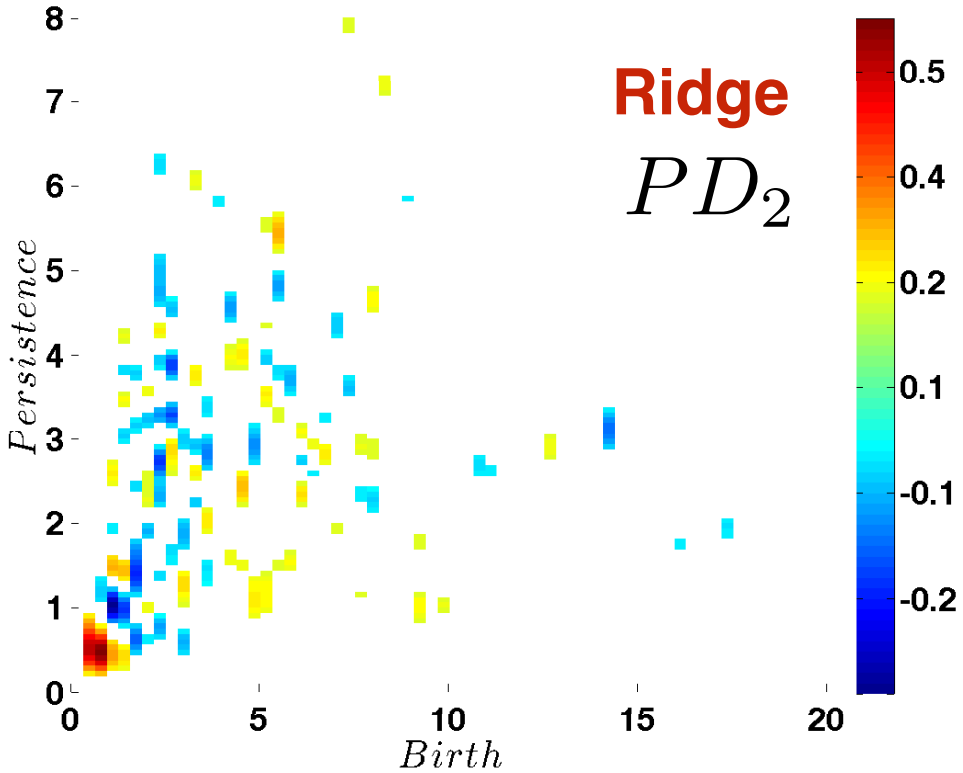
Grid size for vectorization of diagrams: 50x50

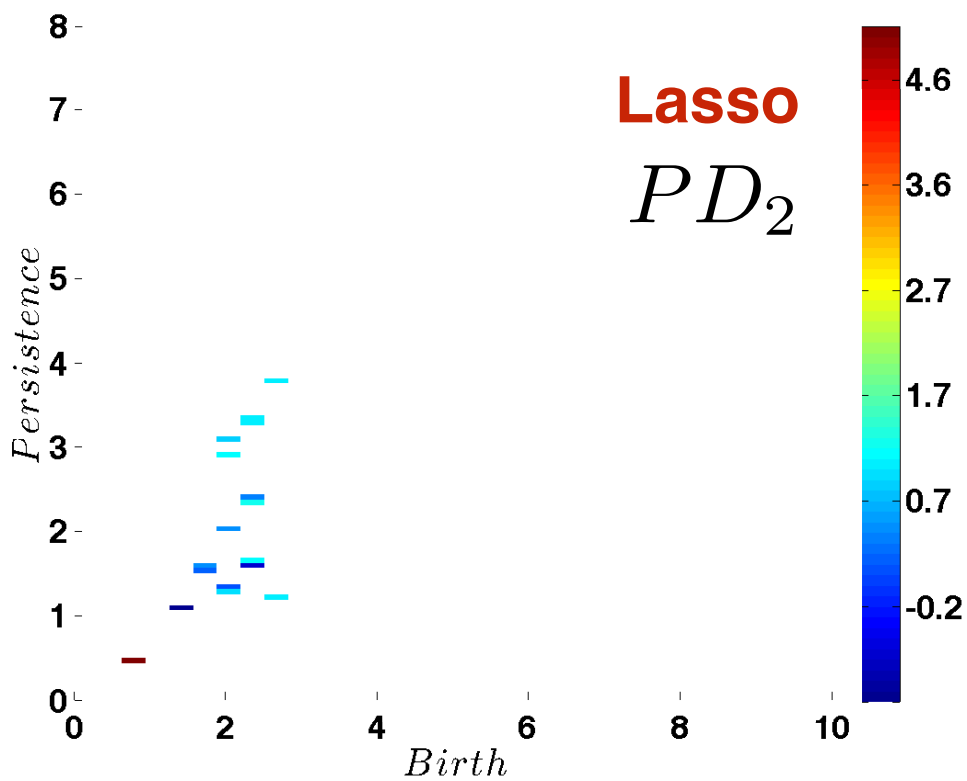




Protein Folding Analysis

Plot of thresholded dual persistence diagrams

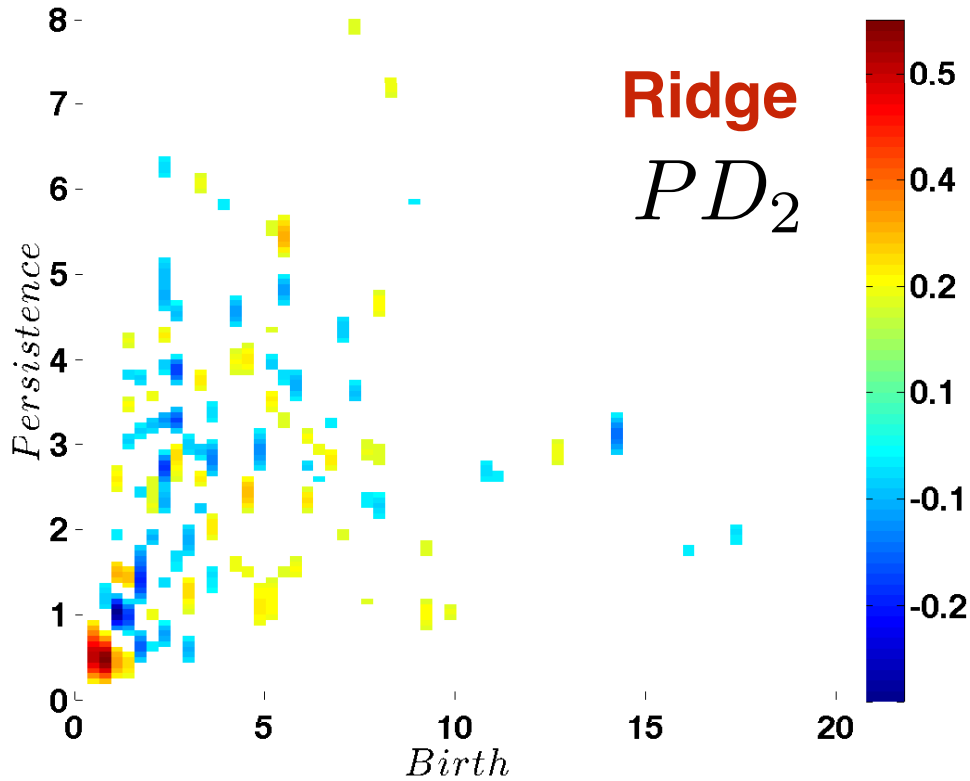




Protein Folding Analysis

High density in dual persistence provide more relevant regions

Need to identify these regions on the proteins (generators)



Compare the regions on the proteins with small and large errors

This could provide information about regions where energy functional is more problematic

Thank you for your attention!