Space oddity. Hyperbolic Learning on Omics Data

Abstract

Many objects in biology are connected by hierarchical relationships. To clarify how snippets of data are associated, we apply embeddings, that is, mapping of multidimensional objects into space so that similar objects are positioned at close points. Analyzing such data with tools operating in Euclidean spaces is problematic as the tools may not account for the underlying data hierarchy. We applied several dimensionality reduction methods based on hyperbolic geometry to diverse datasets including gene expression, gene interaction, microbiome composition, and gene phyletic patterns. We compare hyperbolic embeddings to the more conventional ones (PCA, UMAP) and discuss the difference.

Datasets

C. elegans

- Single cell gene expression [1]: table of read counts with rows corresponding to cells and columns, to genes

Human

- BioGRID gene interaction network [2]: pairs of genes involved in confirmed interactions either as RNA or proteins

Bacterial

- Earth microbiome [3]: table of read counts with rows corresponding to samples and columns, to bacterial species
- EggNOG gene patterns [4]: table of zeros and ones with rows for each bacteria species and columns for gene orthology groups

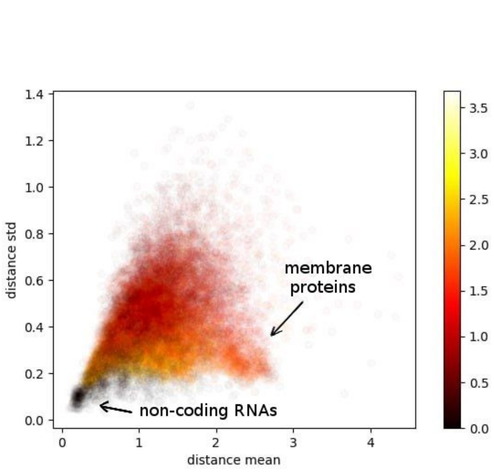
Methods

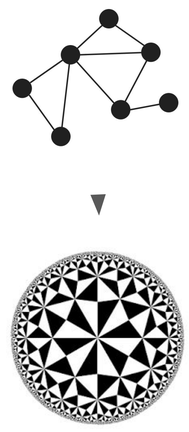
Hyperbolic:

- PoincareMap [5]
- Gensim PoincareModel [6]
- Hyperbolic autoencoders [7]

Euclidean:

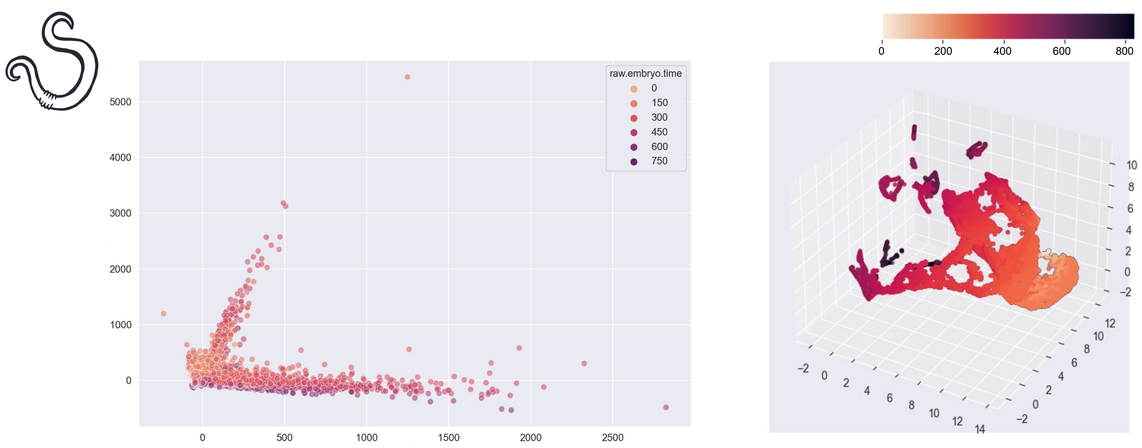
- PCA
- UMAP





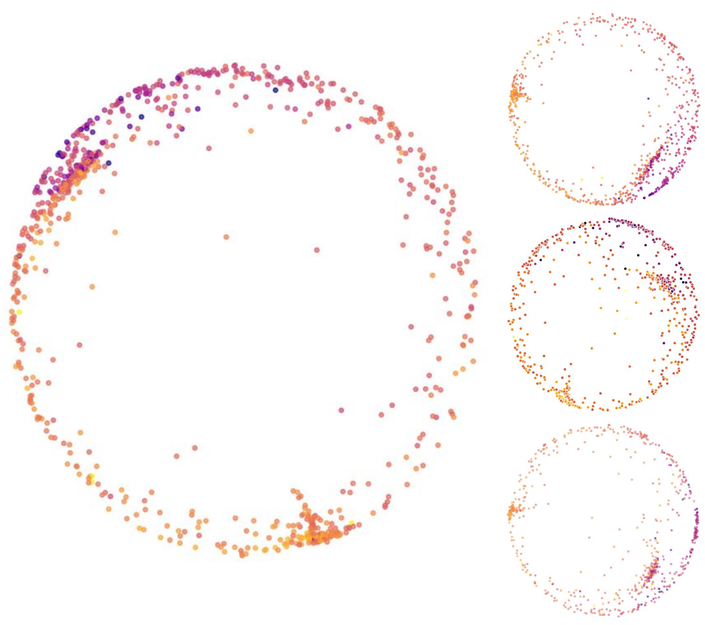
Gene embeddings are stable for genes presumed to be high in the interaction hierarchy, but also for some of low-hierarchy genes

Each dot is a gene. Gene-gene interaction network was subsampled 150 times and embedded to the Poincare disk. For each gene in each sample we measured its distance to the disk center. The closer teh dot is to the center, the higher it is supposed to be in the hierarchy. The Y-axis shows the variance of the distance, the node color represents the node degree (in the logarithm scale, right) of each gene in the gene interaction network.



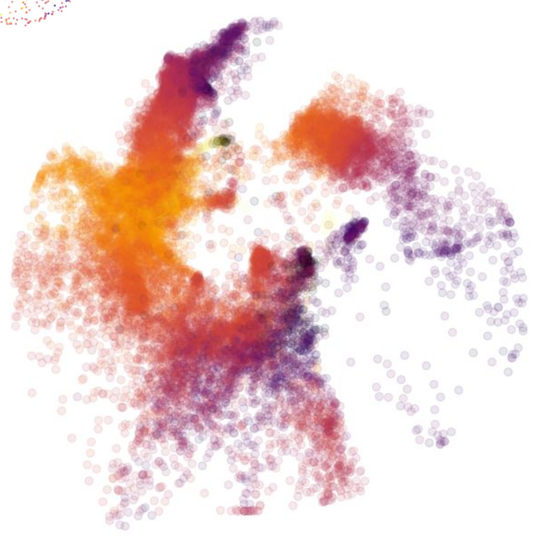
Cells at similar embryo stages are placed together by PCA and UMAP

Each dot is a cell from *C.elegans* single-cell RNA dataset. Dots are colored by the stages of embryo development.



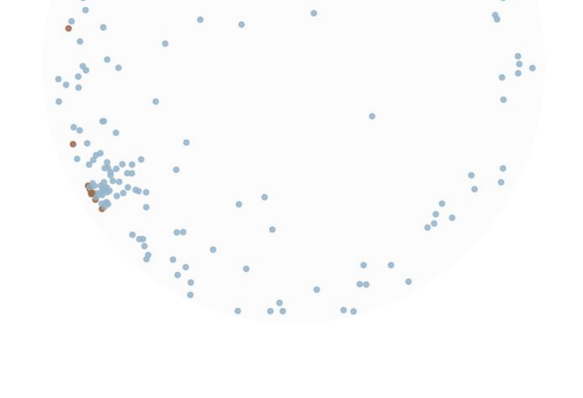
Hyperbolic embedding shows cell hierarchy consistent with the embryo stage

It can be seen by the Poincare analysis of cells correlations in *C. elegans*. Each dot corresponds to a single cell, the darker the color, the older is the cell. Most maps feature conserved structures, clusters of early-stage cells (a yellow triangle below) and gradients of later-stage cells (a darker structure above). As expected, darker dots are all placed at the edge of the map, consistent with the older cells being more differentiated.



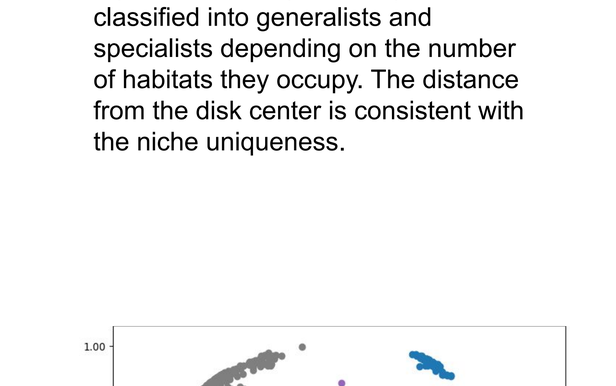
Hyperbolic embedding co-localizes cells with similar embryonic time

It can be seen on an autoencoder analysis of *C.elegans* cell expressions.
Each dot corresponds to a single cell. The dot color depends on the embryonic time of the cell - the darker is the dot hue the older is the cell.



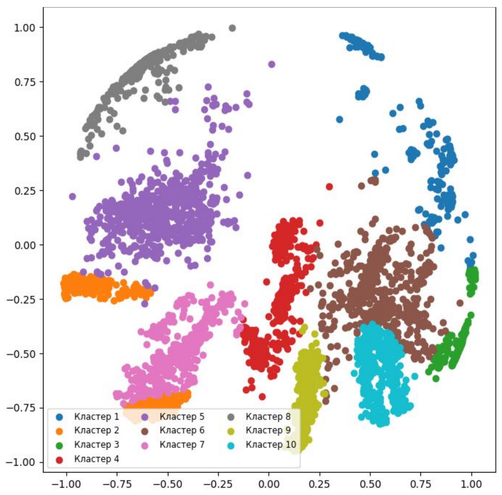
Microbiome samples taken from same ecological niches are clustered by PCA and UMAP

Each dot is a bacterial microbiome of single sample from a certain niche. The niches are colored (see the legend).



Specialist bacterial species are located further from the Poincare disk center than generalists

The distribution is shown using the Poincare analysis (gensim) of bacterial metagenomic data. One dot represents one species. Species are classified into generalists and specialists depending on the number of habitats they occupy. The distance from the disk center is consistent with the niche uniqueness.



Bacterial species with similar phyletic patterns co-localize in the autoencoder embedding of the EggNOG database

Dots on this embedding represent bacterial species. Clustering is consistent with the phyla. Clusters 1, 3, 6, and 10 consist of Gammaproteobacteria, cluster 9 formed by Alphaproteobacteria, clusters 2, 5, and 8 contain Bacillota, and cluster 4 is the FCB (Bacteroidota/Fibrobacterota) group.

Conclusions

We applied hyperbolic dimensionality reduction methods to several biological datasets. To our knowledge, these methods were previously applied only to gene expression data, and we succeed to reproduce the published results on the hierarchy of cell hyperbolic embedding of *C. elegans* single cells. However, we failed to obtain stable results with PoincareMap method used by the authors [1], as its results depend on multiple hyperparameters.

For the first time we predicted hierarchy of human genes based on hyperbolic embeddings from gene-gene interactions and showed it to be stable against data permutations. We also embedded bacterial species based on their presence in various environmental samples and showed the generalist species to be higher than specialist species in the reconstructed hierarchy.

INSTITUTE OF MOLECULAR BIOLOGY

References:
http://bitly.ws/SMQR