THE BIOMOLECULAR UNIVERSE: INCEPTION

Peter

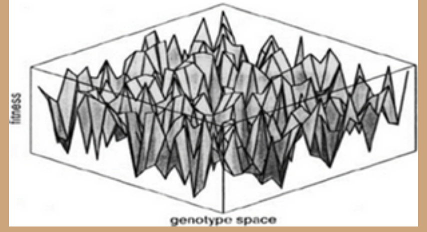
Ilya

Participants

Liliya
Bohdan

Alexey
Matyash

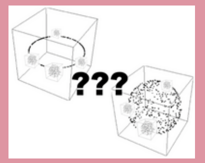
Irina
Malysheva



1

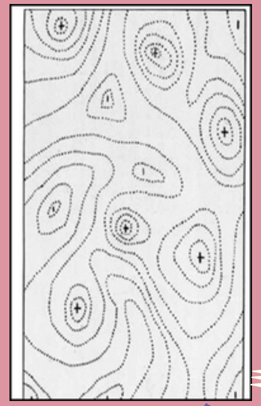
THE PROJECT DESCRIPTION

The "fitness landscape" is a key concept in evolutionary biology that describe the relationship between genotype and reproductive success (fitness). In this representation, genotypes that are similar (by sequences) are said to be "close" to each other, and genotypes that are very different are "far" from each other. This is a convenient theoretical concept - however, despite years of theory and experiments, we still know little about the real structure of this landscape, especially on a very big (evolutionary) scale.



Also, considering the space of sequences and the points "inhabiting" it, it is quite intuitive to move on to observing some areas of increased and reduced density of points that actually occur in nature (with many functional, "suitable" sequences). Then one can operate with geometric and topological categories to study this space. Until recently, the number of available (sequenced) biological molecules (DNA, RNA, proteins) was not large enough and simply did not allow us to analyze the sequence space... but now the number of annotated sequences is quite enough for such an analysis.

In our project, we are using approaches from geometry and topology to study the distribution of proteins in their "sequence space" and across the fitness landscape at very different distances. For the protein families, we evaluated the "individual" properties of their sequence space - and visualized it. Additionally, it was interesting to test the hypothesis that due to the fundamental features of the evolutionary process the distribution of sequences (in the global "space" of all their possible variants) remains constant for protein families over billions of years of evolution.

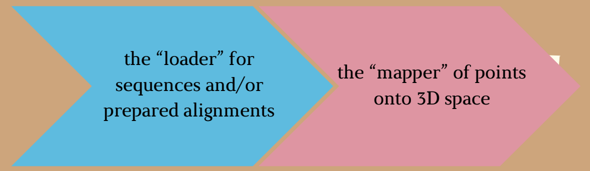


THE ENGINE description for the Sequence Space visualization

2

- Data analysis takes place in the Python 3 environment in the Jupyter Notebook instance, which affilated with the freely-available Kaggle server:
- The laptop reads the alignment files uploaded to the FASTA server and extracts the aligned sequences, to build their sequence space.
- The next step: the calculation of the matrix of relative pairwise Hamming distances between all sequences extracted from the alignment.
- Based on this matrix, the average pairwise distance is calculated, and the correlation dimension of the space of these sequences is also calculated.
- After, using multidimensional scaling, the original multidimensional sequence space of these sequences is projected onto a three-dimensional one and then visualized using interactive scatter plots.

The key modules:



the "loader" for sequences and/or prepared alignments

the "mapper" of points onto 3D space

THE IDEAS for exploring the Sequence Space - implemented already or for the future development of the project:

3

- comparison of the topologies of a "natural" set of points (taken from nature) and a "stochastic" set (generated by a random distribution of mutations)
- phylogeny restoration (for a specific family) and subsequent visualization of Sequence Space, taking into account the distribution of points (sequences) on evolutionary branches
- comparing the distribution of points for working and non-working copies of the SAME gene (because then we can observe / depict them in one "sector" of Sequence Space) - and, thus, assess the influence of selection (and its absence) on topology formation
- compiling lists of "extreme" species (bacteria) with unusual properties - thermophiles, fast-evolving, etc. - and writing an AUTOMATIC PASER that could match the names of species from such lists with names in a large dataset in order to "mark special points"
- the SequenceSpace visualization and analysis of genes that "represent" some important interspecies transition... for example: genes that were under strong driving selection during the separation of humans and primates (it is assumed that these genes are somehow connected with the formation of intelligence)

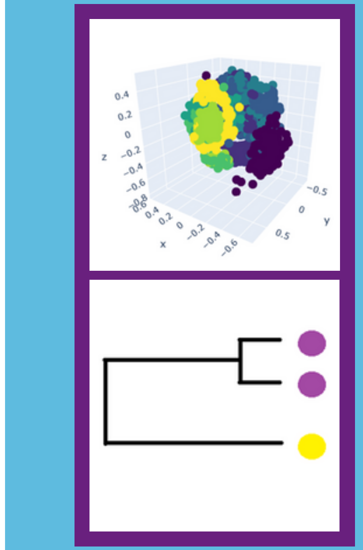
ZIMIN FOUNDATION

SMTB

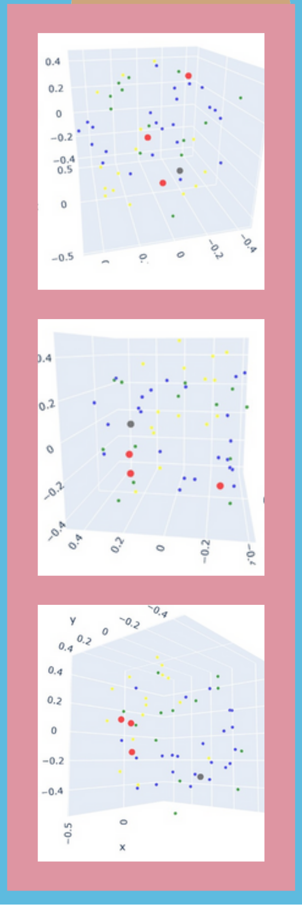
hhmi Howard Hughes Medical Institute

4

This diagram was created by projecting the alignment of proteins sorted by the time of their "last modification". The oldest and newest protein versions are marked yellow and purple respectively.



palette viridis



$$\sqrt{\frac{\sum(d_{ij}-\delta_{ij})^2}{\sum d_{ij}^2}}$$

The evaluation of projection quality is called **stress**.

Stress is defined as *
where dij are the original pairwise distances between points and 6ij are the corresponding pairwise distances in the new dimension space.

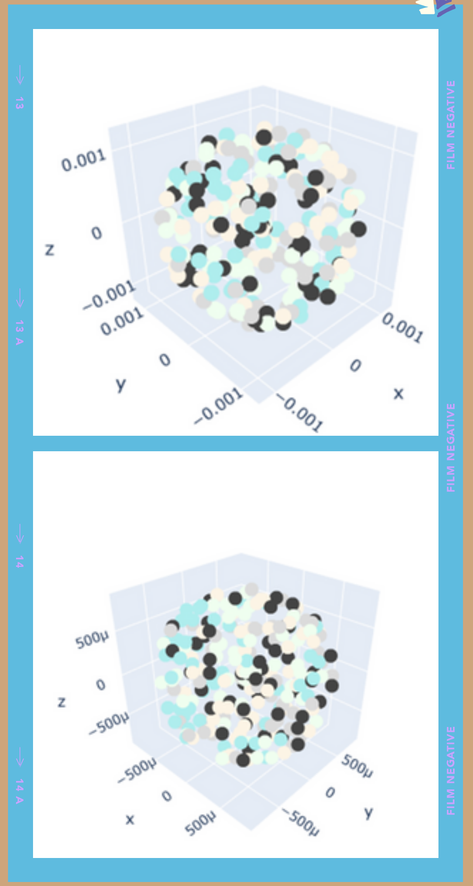
With the high "stress" different patterns are obtained each time. Pay attention to the red dots.

(the stress ~ 0.4)

sets of stochastic (random) points

5

It was decided to conduct an **experiment** (which severely tested our belief in the applicability of the methods used in the project). A random amino acid sequences set was generated based on the consensus and the average distance between the sequences. Surprisingly, they were not four concentric (consensus-centered) spheres as expected, but the figures shown above.



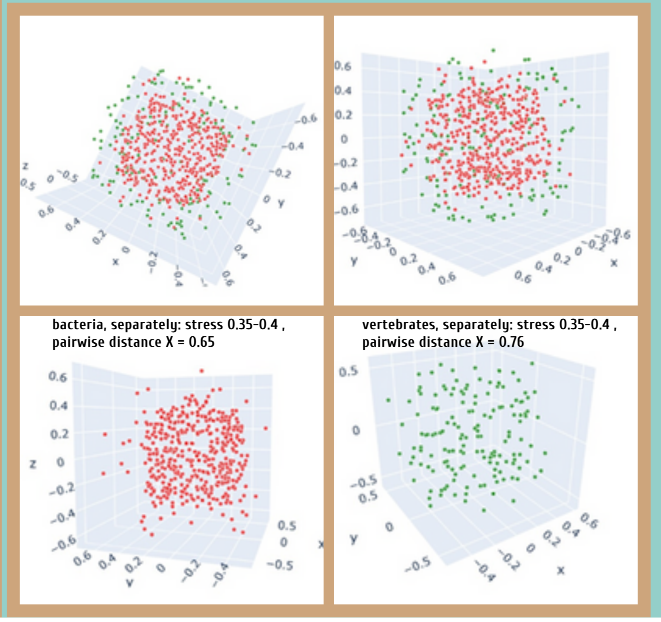
The final visualization for two very different evolutionary groups' sequences **vertebrates and bacteria**

green dots = vertebrates,
red dots = bacteria

6

1

stress = 0.35-0.4, average pairwise distance X = 0.8
bacteria fit inside the sphere while vertebrates on its shell, mostly



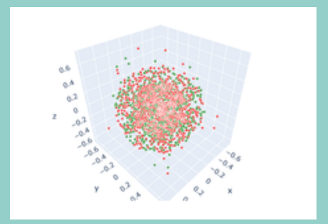
bacteria, separately: stress 0.35-0.4 , pairwise distance X = 0.65

vertebrates, separately: stress 0.35-0.4 , pairwise distance X = 0.76

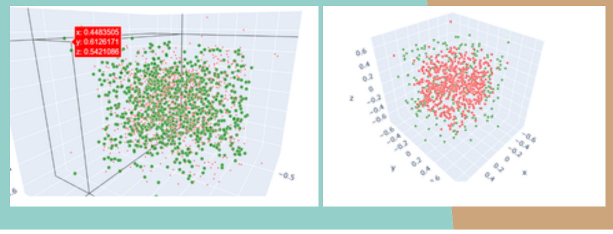
We chose proteins whose sequences are very similar in bacteria AND vertebrates. So we wanted to look at the Sequence Space for families with points (sequences) went through very different evolutionary selection, that (probably!) formed very different patterns in the sequence space (but, these proteins retained a noticeable similarity for their **sequences = structures = functions**).

2

The same pattern



3



RESULTS

7

- A convenient practical tool for analyzing and visualizing the space of sequences has been created
- The sequence spaces are visualized for protein families whose sequences in bacteria and vertebrates are very similar (therefore - have a common evolutionary history). This allows you to "look" at SequenceSpace in a situation where points (genes) have undergone very different evolutionary selection (such a difference is obvious for bacteria VS vertebrates)
- Unexpected result - in terms of some intuitive expectations: the visual patterns of the distribution of sets of points that are separated from a given center by various discrete "steps". One could assume that such sets from spheres of various diameters nested into each other, like in the "Russian Doll" - BUT, in fact, the points of these sets are mixed with each other, forming some single spherical layer