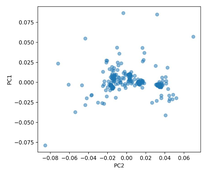
The dimensionality of protein sequence space

Michael Belsky, Gerard Grau, Andrey Stashko
Olga Bochkareva, Lada Isakova

Introduction



The evolutionary distance between biological sequences can be displayed on the multidimensional map called **sequence space** where each sequence is represented by an individual dot. The shape and the dimension of the protein sequence space may contain information about the fundamental constraints on the evolution of individual proteins.

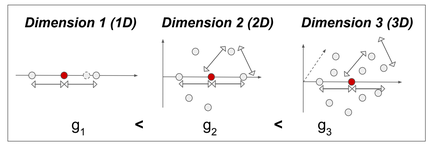
We use the correlation dimension to approximate the dimensionality of a protein space. For instance, it may provide a way to track the epistatic effects in the evolution of individual proteins or to estimate the rate of evolution within different phylogenetic groups.

How do we measure the correlation dimension of protein sequence space?

The number of pairs of points (g_ε) at a given distance Ɛ or closer to each other are proportional to the dimensionality of the space D.

$$g_\varepsilon \sim \varepsilon^D$$

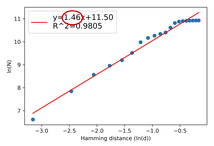
We can use this proportionality to calculate the dimensionality of any set of points. In the case of protein sequence data we calculate the pairwise distances from sequence alignments.



Multiple alignment of orthologous proteins (from 391 vertebrates)

Pairwise distance matrix

Dimensionality plot, k-value



To calculate the correlation dimension, we build a log-log plot of the number of pairs of sequences (ln(N)) at a given distance Ɛ or less against the distance Ɛ (blue dots, ln(d)) and do a linear regression on it. The slope of the obtained cumulative curve is the dimensionality value (k-value) for the sequence space of a given set of orthologous proteins. The dimensionality is not necessarily an integer.

Why did we choose correlation dimension?

Among all existing measures of dimensionality we are considering the correlation and Minkowski dimensions. Both of them are types of Fractal dimension and can be considered as separate cases of Renyi's fractal dimensionality formula.

Meanwhile most researchers use Minkowski dimension and method called Box-counting method (further BCM)

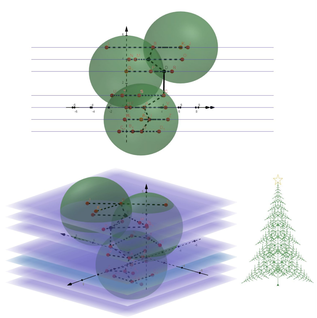
Why didn't we also use it?

1. Our data exists in multidimensional space (much more than 3D) for which the calculation methods are not developed yet
2. BCM requires a lot of computing memory and is based on graph structures that's not easy to work with in an N-dimensional space
3. BCM describes the dimensionality of a graph, that could be constructed from our points, but not necessarily the space in which these points exist

Nevertheless it would be interesting to calculate the Minkowski dimension of our data and compare it with the correlation dimension. So we want to suggest a theoretically effective method for constructing and measuring multidimensional fractal objects. We called this method "Multidimensional Christmas tree method". Here is the outline:

1. Checking the Triangle rule for every possible combination of sequences
2. Putting them in sets of 2D slices
3. Constructing multidimensional graph using this data and simplifying it by eliminating all edges of the length >= ε
4. Building a fractal graph from this graph
5. Measuring the minimum amount of minimum-size multidimensional spheres covering the obtained fractal using the Minkowski dimension formula (1) in its simplified form (2) (right panel).

Below you can see some visualisations of the "Multidimensional Christmas tree" method and explanations of formules below, which describe measuring the Minkowski dimension:



$$D = -\lim_{\epsilon \to 0} \frac{ln(p_\epsilon)}{ln(\epsilon)} \quad (1) \qquad p_\epsilon \propto \xi^{-D} \quad (2)$$

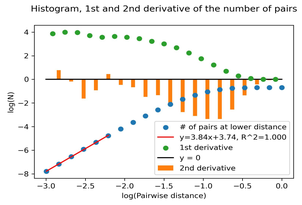
ϵ- radius of the n-dimensional sphere; grid cell size
D - dimensionality
p_ϵ - minimum number of n-dimensional spheres
ξ - scale (the ratio of the size of the subgraph to the original graph)

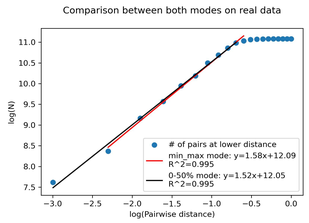
Results

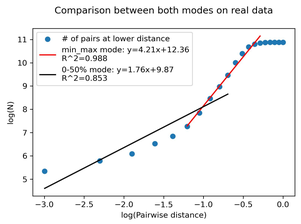
1. Improving the function for dimension estimation

1.1. What part of a curve should we use for the dimension estimation?

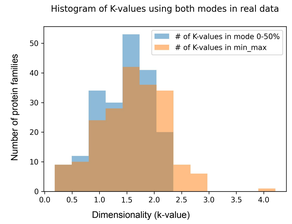
To estimate correlation dimension in case of not completely linear graphs, we have to choose an interval to apply linear regression in.
The most linear part of the graph is the most appropriate for regression. Calculating the first and second derivative on simulated data we see that on the first 6-8 points of the plot the linear regression would be the most accurate (right figure). Though, this is not always the case for real protein families (bottom figures).

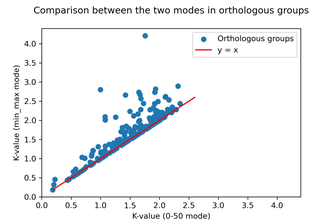






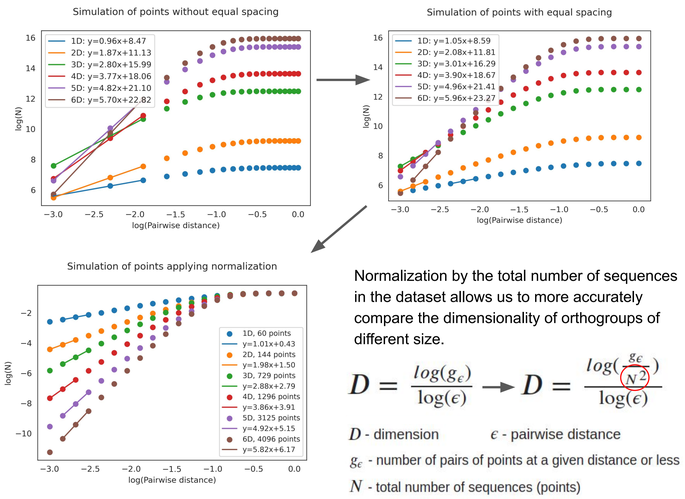
Thus, the part of the curve with the highest slope gives the best fit overall. It was also previously reported as a gold standard way to estimate the dimensionality (Boon et al. 2008). Yet for most protein families the difference between the value obtained using the first ten points of the curve and the value obtained for the region with the highest slope is not large:





1.2. Scale change and normalization

The dimension estimation will be more precise if the points are equally spaced on the log scale, instead of the linear scale.



Normalization by the total number of sequences in the dataset allows us to more accurately compare the dimensionality of orthogroups of different size.

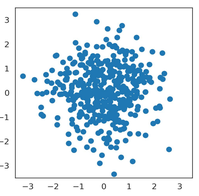
$$D = \frac{log(g_\epsilon)}{\log(\epsilon)} \rightarrow D = \frac{log(\frac{g_\epsilon}{N^2})}{\log(\epsilon)}$$

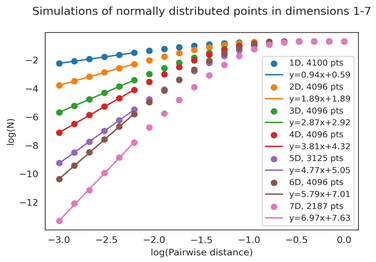
D - dimension ϵ - pairwise distance
g_ϵ - number of pairs of points at a given distance or less
N - total number of sequences (points)

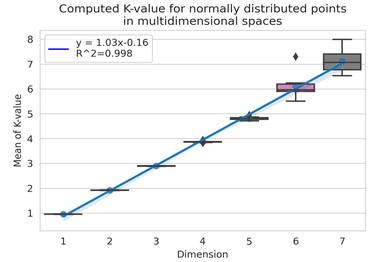
2. Dimensionality on a simulated set of random points in the space of a known dimension

An N-dimensional protein sequence space can be simulated by randomly placing some points in an N-dimensional space following a normal distribution.

Random normally distributed points in a 2D space





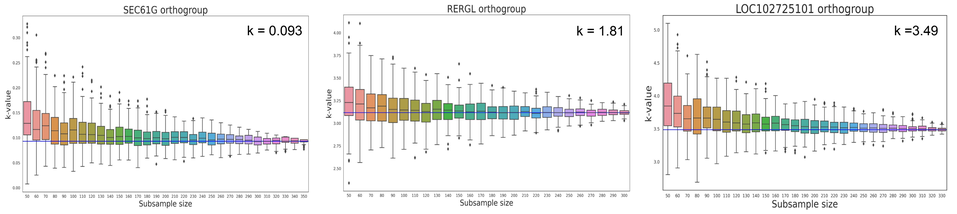


We performed data simulations to proof the validity of our method of dimensionality calculation (left). Indeed, the estimated dimensionality occurred to be close to the dimension the points were simulated in. The box-plot graph shows that the computed dimensionality in simulated spaces is very close to the given dimensionality of the points (right). For higher given dimensionality, the calculated dimensionality values are more disperse as we have a smaller number of points per dimension to play with.

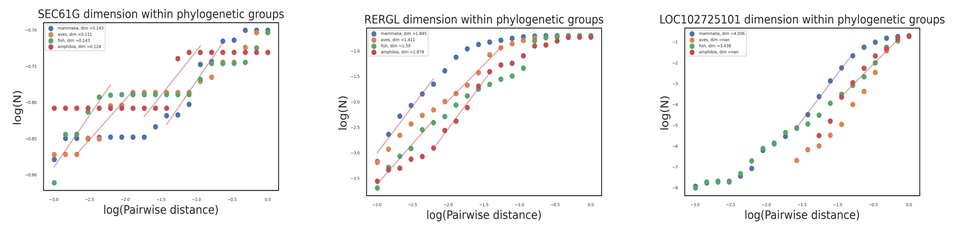
3. Robustness of dimensionality estimation for vertebrates

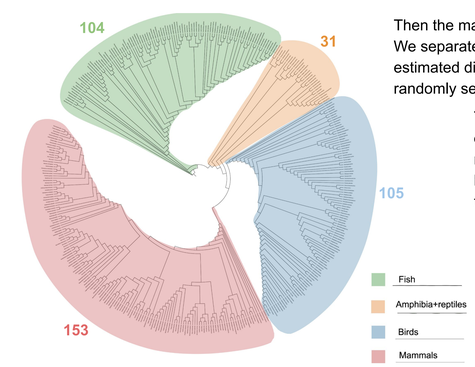
3.1. Random subsampling

A sampling of multiple random subsets of sequences from alignments to perform calculations is a way to check robustness of the calculations. Thus, box plots represent the variance in k-value between random subsamples with different sizes. The horizontal line represents the dimensionality for the initial matrix. We see that the variance of the estimation decreases with the increasing subsample size.



3.2. Taxonomy subsampling





Then the matrices were subsampled according to the phylogenetic tree structure. We separated proteins from fish, mammalia, amphibia, and birds and compared the estimated dimensionalities. For this analysis we used the sequences from 199 randomly selected orthogroups.

The dimensionality estimates of the same protein in different classes of organisms differ. This might be explained by biological or computational reasons. In particular, amphibia group is the smallest one that may explain higher errors in dimensionality estimation in this subsample in comparison to other classes.