**Classical Multidimensional Scaling for Genome and Protein Comparisons**

Vyacheslav Brover

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

# Structured Abstract

The method of classical multidimensional scaling (MDS) is described from the point of view of practical application. It is shown that an MDS application is successful if the input distances or similarities meet certain requirements. The classical MDS method is compared to principal components analysis (PCA) and non-classical MDS methods. The classical MDS is illustrated with real data examples of genome and protein comparisons.

# Introduction

Multidimensional scaling (MDS) is one of the basic explorative methods of multivariate data analysis, together with principal component analysis (PCA) and clustering. MDS is similar to PCA in that a set of new numerical attributes is constructed on which the data points, or objects, are projected, but is different in that the input for MDS is an object-object table of pairwise distances or similarities between objects, whereas the input for PCA is an object-attribute table.

However, MDS is not as popular in bioinformatics, as PCA, though multiple genome and protein comparisons are abundant, because the MDS requires that the object-object similarity matrix should be approximately positive semidefinite.

The MDS method uses a double-centered symmetric matrix as input, and constructs centered and non-correlated numerical attributes y1, y2, y3 … yq as output, like in PCA, such that y1 has the largest variance, y2 is orthogonal to y1 and has the largest variance, and so on.

If objects make up large clusters located distantly from each other, then the MDS plot will clearly show them. Therefore, MDS is usually followed by a clustering in the space of the MDS attributes.

If the object-object table of pairwise distances can be represented by a tree, then this is the best representation of the objects, but only if the tree is constructed by a maximum likelihood estimation (MLE) method. However, building an MLE tree is an NP-hard problem. The MDS plot with clusters is better than a tree in an explorative analysis, because it is faster and easier to use. It also turns out that MDS can reveal some properties of the MLE tree.

# Method

The dataflow is shown in Fig. [Dataflow].

## Matrix of double-centered generalized similarities

### Generalized distances and similarities

Let ℝ be the set of real numbers and I be the set of imaginary numbers.

Let a matrix Xraw be an object-attribute table, the n rows of which represent objects and the p columns of which represent numerical attributes. Each attribute is either in real or imaginary numbers. This space of attributes will be referred to as a generalized Euclidean space.

The matrix of generalized squared distances D2 is defined as

D2[i,j] = (Xraw[i] – Xraw[j])T (Xraw[i] – Xraw[j]),

where Xraw[k] is a vector of attributes of object k.

The superscript 2 in D2 is a part of one whole, and does not mean D × D.

The matrix of generalized distances D is related to the matrix of generalized squared distances by the equation:

(D[i,j])2 = .

Any symmetric matrix in real numbers with a zero diagonal can be a matrix of generalized squared distances or a matrix of generalized distances.

If Xraw then the generalized distances are Euclidean distances and satisfy the distance axioms.

The matrix of generalized similarities Sraw is defined as

Sraw = Xraw XrawT.

If Xraw then the generalized similarities are Euclidean similarities and Sraw is positive-semidefinite (p.s.d.).

Any symmetric matrix in real numbers can be a matrix of generalized similarities.

The matrix of generalized squared distances is related to the matrix of generalized similarities by the equation:

D2[i,j] = Sraw[i,i] + Sraw[j,j] – 2 Sraw[i,j].

### Transformation into a matrix of double-centered generalized similarities

A centered attribute is an attribute whose average is zero. An attribute is made centered by subtracting its average from it. Let X be the object-attribute table obtained from Xraw by centering all attributes. Then the matrix of double-centered generalized similarities S is defined as

S = X XT.

The matrix S is called double-centered because **1**TS = **0** and S**1** = **0.**

A matrix of generalized similarities Sraw is transformed into a matrix of double-centered generalized similarities by the equation:

S = Cn Sraw Cn,

where Cn = I - An is a centering matrix, An = 1/n **1** **1**T is an averaging matrix and **1** is an n-dimensional vector consisting of ones.

A matrix of generalized squared distances D2 is transformed into a matrix of double-centered generalized similarities by the equation:

S = -1/2 Cn D2 Cn.

### Combining several double-centered generalized similarity matrices into one

Several double-centered generalized similarity matrices can be combined into one double-centered generalized similarity matrix by normalizing each one and then averaging them.

The normalization of a double-centered generalized similarity matrix *S* is done by dividing all matrix elements by , which is equivalent to making . This normalization makes each original matrix to have equal contribution in the combined matrix, which is analogous to PCA applied to Xraw where each attribute should have the same contribution.

If all the original matrices are p.s.d., then the resulting matrix is also p.s.d.

## Classical MDS

The goal of the classical MDS is to find an object-attribute table Yq for a double-centered generalized similarity matrix S , such that

.

The procedure has one parameter: the criterion of the selection of the number of attributes q.

The result of MDS consists of:

* object-attribute table Yq;
* quality table.

The procedure is equivalent to these two consecutive steps:

1. construction of the complete MDS space Y, such that S = YYT;
2. PCA of Y with the criterion of the selection of q.

Step 1 is the classical MDS procedure per se, it has no parameters and achieves the goal precisely. Its drawback is that generally the result is not in real numbers.

### Procedure

The procedure of the classical MDS was introduced by [Young, Householder].

For a symmetric matrix S , the first q eigenvectors and their corresponding eigenvalues are generated in the order of |λ1| ≥ |λ2| ≥ |λ3| ≥ … ≥ |λq|, where q is selected by an external criterion, see below. Let the columns of matrix Bq be the eigenvectors.

For all i: λi

Bq .

Then the q-dimensional MDS space of S is the matrix Yq computed as

Yq = Bq Λq1/2,

where Λq is a square matrix with all elements equal 0 except for the diagonal elements equal to λ1, λ2, λ3, …, λq.

If λi ≥ 0 then the *i*th column of Yq contains only real numbers. If λi < 0 then the *i*th column contains only imaginary numbers, and if it is projected onto real numbers, the *i*th column will contain only zeros. The projection of Yq onto real numbers can be used for other data analysis procedures.

An MDS space with a larger q contains an MDS space with a smaller q.

The PCA of Yq is Yq.

### Optimality properties

The optimality properties of the classical MDS follow from [Eckart, Young].

The matrix Yq is optimal for a given q in several aspects:

Yq = arg ,

Yq = arg

and

Yq = arg = arg ,

where D2(M) is the matrix of generalized squared distances between the rows of the matrix M.

### Quality table

The quality criteria are the optimization criteria of the optimality properties.

The value

is the explained squared similarity, which is between 0 and 1.

The value

= (1 – sq) tr S2

is referred to as the MDS strain [Cox].

The quality of MDS can be represented by the MDS quality table which contains the percentages of , , …, relative to with the indication of the negativity of λi.

The value

is the explained data variance. This criterion is used in PCA, where it is between 0 and 1. But in MDS it may be not between 0 and 1 because of negative eigenvalues which do not contribute to the explanation of variance.

(1 – rq) tr S = =

An MDS application can be called successful if all eigenvalues in the quality table are positive.

### Criterion of the selection of the number of attributes

Let q be the dimensionality of MDS space, i.e., the number of the MDS attributes.

q ≤ rank(S) ≤ n – 1.

For the maximum q the quality criterion are best and

However, the MDS attributes for small |λi| contain noise and are not accurate due to numerical reasons.

In this paper q is maximum such that /tr S2 > 0.5%.

The fact that q is too large and the MDS attributes with large indices contain noise does not harm the subsequent data analyses of the MDS space, because usually data analysis procedures are tolerant to noise.

### Relation between MDS and PCA of the original object-attribute table

The PCA of X and the MDS of X X’ are the same given the same q. Therefore, the attributes of Yq are also the principal components of the MDS space.

For a matrix Xraw , the matrix D2 or Sraw can be computed, transformed into S and used in MDS. This is an alternative way to do the PCA of X, see Fig. [Dataflow].

MDS is faster than PCA and takes less memory if n < p.

## Generalized distances

### Requirements to generalized distances

If the MDS dimensionality q is large, then the exact value of q is not important. Therefore, the MDS result depends only on the choice of the generalized distance or similarity function which produces the input object-object table.

A generalized distance matrix D represents distances in a Euclidean space if and only if the matrix of double-centered generalized similarities S obtained from D is p.s.d. In this case ≥ 0 for all i, Yq and rq > 0.

In practice, MDS has no negative eigenvalues for large q if necessary conditions of the positive-semidefiniteness hold.

A principal submatrix of S is a matrix containing some rows and columns of S, such that the indices of the rows and of the columns are the same. A symmetric matrix S is p.s.d. if and only if for all its principal submatrices P, det P ≥ 0. A relaxation of this condition is a necessary condition of the positive-semidefiniteness:

for all m ≤ k, P : if P is a principal submatrix of S then det P ≥ 0,

where 1 ≤ k ≤ n.

If k = n then there is no relaxation and S is p.s.d.

If k = 1 then S[i,i] ≥ 0 for all i.

If k = 2 then S2[i,j] ≤ S[i,i] S[j,j] for all i and j, which is the Hölder’s inequality.

If a matrix D satisfies the distance axioms then the induced double-centered matrix of similarities satisfies the Hölder’s inequality.

If a matrix D contains Euclidean distances obtained from a Euclidean space then the induced double-centered matrix of similarities is p.s.d.

If a distance matrix can be represented by a tree, then there is a Euclidean space representing the square roots of the distances of this distance matrix, see below.

### Tree representation of a distance matrix and a Euclidean space representing the tree

If the n objects are the leaves of an undirected tree whose arcs have non-negative lengths, then each pair of objects has a unique path connecting them, and the sum of the lengths of the arcs on this path satisfies the distance axioms. It will be referred to as the tree distance.

Let m be the number of the arcs in the tree.

Let *root* be any node of the tree, and *path(x,root)* be the set of arcs on the path from a leaf x to *root*. Then each object x can be placed in an *m*-dimensional Euclidean space by assigning the coordinates

xi = , if i *path(a,root)*, otherwise 0,

where i is an arc, is the length of the arc, and 1 ≤ i m.

There is a 1-1 relation between the tree arcs and the dimensions of this Euclidean space. However, generally there is no 1-1 relation between the tree arcs and the attributes of the MDS space, because the number of arcs is between n and 2n-1, but the number of MDS attributes is less than n.

For each pair of objects, the tree distance between them is the squared distance between them in the Euclidean space.

Therefore, if for a distance d(x,y) there theoretically exists a tree representing this distance, and if a new dissimilarity measure d1/2(x,y) is defined as d1/2 (x,y) = and transformed into a double-centered similarity matrix, then this matrix is p.s.d. The dissimilarity measure d1/2 (x,y) satisfies the distance axioms.

Each internal node of the tree has a coordinate. The root of the tree has coordinate 0.

If for two internal nodes the paths from the root to them do not intersect, then the vectors from 0 to their coordinates are orthogonal.

Since the construction of the Euclidean space depends on the arbitrarily chosen root, this space is not unique, but its PCA is unique, which is the same as the MDS of the distances which are the square roots of tree distances.

### Some simple generalized distances

#### Generalized distances based on whole genome evolution models

If objects evolve from one ancestral object then they are leaves of a directed tree. Each interior node of the tree is an extinct object. An evolution model specifies evolution events of some kind which transform parent objects into child objects or the other way around.

The evolution events act on parts of objects. It is assumed that each part evolves independently, each arc (a,b) of the tree has length λab ≥ 0 , and the number of evolution events on the arc per part has a Poisson distribution with parameter λab. Then for any pair of nodes (x,y) the number of evolution events on the path from x to y per part has a Poisson distribution with parameter λxy = λxu + λuv + … λwy, where (x, u, v, w, …, y) is the path from x to y.

The values λxy are tree distances between objects x and y. The generalized distances between objects x and y are estimations of λxy.

The choice of symbol λ is traditional for a Poisson distribution and is not related to eigenvalues.

##### ANI-based Jukes-Cantor model

The evolution event is a mutation of one nucleotide.

The DNA of two genomes x and y is aligned by NCBI MegaBLAST [MegaBLAST] with the parameters: -max\_target\_seqs 100000; -xdrop\_gap 150; -xdrop\_gap\_final 150; -penalty -1; -gapopen 3; -gapextend 1; -dbsize 10000000; -searchsp 10000000.

The average nucleotide identity (ANI) is computed for HSPs longer than 10000 bp as the sum of the identical nucleotides in an alignment divided by the sum of aligned nucleotides in the alignment, i.e., ignoring gaps. The minimum HSP length is chosen to so that HSPs should be conserved enough to include intergenic regions.

The Jukes-Cantor model [Jukes-Cantor 1969] assumes that nucleotides mutate independently, genomes mutate independently, all nucleotides have equal probability and mutate into any nucleotide with equal probability.

Then the estimated number of mutations per nucleotide is

= - log

where pchance = ¼ is the probability for two nucleotides to coincide by chance.

And the ANI distance is defined as

dANI = 100 ,

where the coefficient 100 is selected to make the average of dANI to be around 1 in intra-species comparisons.

Since nucleotide distribution and mutation depends on whether the nucleotide is in CDS and in which codon position in CDS, the assumptions of this model are not accurate, but show a good fit to data in practice.

Since long enough homologous segments are needed to compute ANI, and horizontally transferred genes should be a small fraction of all homologous matches, in this paper this distance is used only at taxonomic level of genus or below in prokaryotes.

The value of ANI = 94%, or dANI = 8.3, is suggested as the threshold to identify most of prokaryotic species [Richter].

##### Conservation model

The evolution event is a mutation of a segment of genome into any element of an infinite set of genome segments with equal probability. All segments have a fixed small size.

For a pair of genomes, x and y, aligned homologous segments separated by < 400 bp are merged to form conserved segments. The conservation fractions Cx and Cy are computed as the total length of the conserved segments divided by the total length of DNA in genomes x and y respectfully.

A genome is partitioned into many small segments of equal length, which in reverse evolution are either retained or not. Then two genomes share a homologous segment if this segment is retained in the both genomes, which has the probability

Cx Cy = ,

where λxa and λya are the tree distances from the least common ancestor *a* of x and y to x and y respectfully.

Then the estimated number of mutations per segment is

= + = - (log Cx + log Cy)

and the conservation distance is defined as

dcons = 10 ,

where the coefficient 10 is selected to make the average of dcons to be around 1 in practice.

The conservation model is equivalent to the Jukes-Cantor model if pchance = 0, which is possible if the alphabet of mutating states is infinite.

The alignment of two genomes of different prokaryotic phyla by MegaBLAST consists only of 16S and 23S ribosomal RNA, which do not follow the statistical assumptions of dcons. The same happens even within the same phylum, e.g., in comparing the genomes of the orders *Spirochaetales* and *Leptospirales* of the phylum *Spirochaetes*. Therefore, in this paper dcons is computed only for taxonomic levels of family or below in prokaryotes. Since 5-15% of genome consists of accessory genes [Tettelin] which are exchanged horizontally at a high rate, dcons is less precise than the ANI-based Jukes-Cantor distance at a species level or below.

#### Protein transformation distance

This generalized distance is a distance. It is described in [Setubal, Meidanis].

Let sequences be transformed one into another, and each non-identity transformation have a positive cost. Then the minimum cost of a transformation of one sequence into another satisfies the distance axioms.

The minimum cost of a transformation can be computed via an optimal global alignment.

The score of global alignment of two protein sequences is based on a substitution matrix M and scores gopen for opening gaps and *g* for all gaps. A global alignment defines a transformation of one sequence into the other consisting of a series of substitutions, insertions and deletions. And a transformation defines a global alignment.

Let the cost of a substitution of amino acid *a* by *b* be

M(a,a) + M(b,b) – 2M(a,b),

the cost of an insertion or deletion of amino acid *a* be

M(a,a) – 2g

and the cost of starting an insertion or deletion be

-2gopen.

Let s(x,y) be the score of a global alignment of sequences x and y, and d(x,y) be the cost of the corresponding transformation of x into y, then

2 s(x,y) + d(x,y) = s(x,x) + s(y,y).

Let a transformation be simple if in the process of the transformation any amino acid is affected at most once. A global alignment defines a simple transformation.

If all minimum cost transformations contain a simple transformation and s(x,y) is maximum over all alignments of x and y then d(x,y) is minimum over all transformations of x into y.

And the minimum cost of a transformation is

dmin(x,y) = smax(x,x) + smax(y,y) – 2 smax(x,y),

where smax(x,y) is the score of an optimal global alignment of x and y.

The function dmin(x,y) satisfies the distance axioms and will be referred to as a protein transformation distance.

A minimum cost transformation is not simple if, for example, two consecutive substitutions at the same position cost less than one resulting substitution. For the following substitution matrices available for the NCBI BLASTP, and default gopen = -11 and g = -1, all minimum cost transformations are simple: BLOSUM45, BLOSUM45.50, BLOSUM50, BLOSUM50.50, BLOSUM60, BLOSUM60.50, BLOSUM62, BLOSUM62.50, BLOSUM65, BLOSUM65.50, BLOSUM85, BLOSUM85.50, BLOSUM90, BLOSUM90.50, BLOSUMN, BLOSUMN.50 and MATCH. In this paper the default BLASTP matrix BLSUM62 is used.

Though this distance is not based on an evolution model, it fits a tree distance. Therefore, the matrix of dmin(x,y) should be submitted to MDS as a squared distance matrix, which is equivalent to submitting the matrix of smax(x,y) as a similarity matrix to MDS.

## Clustering

The clustering of objects in a Euclidean space is made by an EM algorithm of decomposition of multivariate data into a mixture of multivariate normal distributions. The clustering method is not important, but this method should be tolerant to noise.

If objects have a tree structure then there is no natural number of clusters. In most applications of this paper the clustering has been done into 6 clusters.

The MDS can be applied recursively to each cluster of objects. This way an MDS tree of objects can be produced. This tree can be used to build an MLE tree as an initial tree to which local optimizations are applied.

## Interpretation of MDS result in terms of the tree

If objects can be represented by a tree, then MDS of the tree Euclidean space can reveal some properties of the tree. This does not require building the tree, which is an NP-hard problem.

The statements in this section are rather empirical than mathematical.

### Interpretation of an MDS plot in terms of the tree

The clusters of objects in the MDS plot are large distant tree clades. If only distant clades and/or outliers are the goal of an explorative analysis, then the MDS plot is better than a tree.

The most distant clusters on the plot are the most distant tree clades. The arcs on the tree paths from the common center of these clades to each clade are parallel to the plane, and the other arcs incident to these paths are orthogonal to the plane. Therefore, the clades appear as rays going out of one center. Since the rays are orthogonal to each other in the Euclidean space, three rays have the angles of 120° between them on the plane, and the other rays appear as small clouds in the neighborhood of the center. This allows to guess major paths in the tree topology.

The gap between a distant cluster and the other objects represents the arc going into the least common ancestor of the clade of the cluster, which allows not only to identify a tree arc, but also the lower bound of its length.

### Interpretation of an MDS quality table in terms of the tree

If an MDS has a negative eigenvalue for a large sq, then there is no tree representing this distance matrix with a large likelihood.

Assume a tree represents distances with a high accuracy.

An eigenvalue with a larger index reflects the arcs of the tree with a larger depth. A large means long arcs going out of the tree root relative to the other arcs of the tree. A large q means a large tree depth, where arc lengths are not noise.

If the objects have been obtained by a random sampling, then the percentages of follow the Zipf’s law

by the principle of maximum entropy [***reference***]. The Zipf law is visualized as a straight line if are plotted vs. *i* in a log-log plot. A deviation from a random sampling, for example, stratified sampling, appears as a broken straight line in the log-log plot.

# Materials and Applications

For each MDS application there is:

* an object-object distance or similarity matrix in the Supplementary Materials;
* an explanation why a specific distance is chosen;
* an MDS plot showing objects in the first 3 MDS attributes, where the third attribute is indicated by the diameter of circles;
* a clustering of the objects in the MDS space as a decomposition of a mixture of multivariate normal distributions by an EM algorithm; each cluster has a different color on the plot;
* an MDS quality table - squared eigenvalues;
* a log-log plot of the squared eigenvalues.

In each application the explained squared similarity is above 99.5%.

## Cities

Nine USA cities are in a 2-dimensional Euclidean space, assuming the globe is flat, and the generalized distances between them are approximate Euclidean distances measured in miles. This data is borrowed from <http://www.analytictech.com/networks/mds.htm>. The MDS result is in Fig. [Cities without noise]. The clustering is not done. The MDS contains only 2 attributes as expected, explaining 99.92% of squared similarity, and the cities are placed according to the USA map.

As an experiment, to the X and Y axes of the flat Earth, 10000 new artificial Earth dimensions have been added, distributed as random normal variables with mean = 0 and SD = 50. The MDS is in Fig. [Cities with noise in data]. The MDS space has 8 = 9 - 1 attributes. The distortion in the placement of cities is little, though the first two MDS attributes explain only 37.90% of squared similarity. The range of the first MDS attribute is from -2500 to 2000 miles, whereas this range used to be from -1500 to 1500 without adding noise. The log-log line of the squared eigenvalues is broken into two pieces, which shows a mixed source of data: real distance and noise. MDS is resilient to the noise in the object-attribute table.

As another experiment, a noise has been added to the double-centered similarity matrix as a random normal variable with mean = 0 and SD = 0.5 × average trace element of S. The MDS is in Fig. [Cities with noise in measurement]. The configuration of cities still resembles the map, but there are distortions, e.g., Seattle is between Los Angeles and San Francisco and Chicago is to the South of Denver. The MDS space has 5 attributes, the first two explain only 88.80% of squared similarity, the 3rd and 5th eigenvalues are negative. MDS is harmed by the noise in the distance or similarity object-object table.

## Genome comparisons

The bacterial genomes are identified by GenBank assembly ids and can be found at <https://www.ncbi.nlm.nih.gov/assembly>. The taxonomy classes are identified by NCBI taxonomy identifiers, which can be found at <https://www.ncbi.nlm.nih.gov/taxonomy>.

Whole genome evolution models are used for generalized distances.

Each MDS plot shows well separated clusters.

The clusters form rays going out of the center of the plot with approximately 120° between them.

### *Enterobacteriaceae*

For the family *Enterobacteriaceae* deposited in GenBank as of December 2016 the species with at least 10 genomes have been selected, and for each species a representative genome is used [O’Leary], resulting in 26 genomes. Since all comparisons are made above the species level, the conservation model distance is used. The MDS result is in Fig. [Enterobacteriaceae].

Three rays are visible going out of point 0 with approximately angles of 120 between them:

1. to *Citrobacter* to S*almonella* to *Escherichia coli/Shigella*;
2. to *Enterobacter cloacae complex*;
3. to *Pluralibacter* to *Enterobacter* to *Raoultella/Klebsiella*.

The 4th ray is orthogonal to the plane and goes to *Cronobacter*.

These rays show, for example, that the tree path from *Klebsiella/Raoultella* to *Escherichia coli/Shigella* will pass through *Pluralibacter , Cronobacter/Enterobacter cloacae complex*, *Citrobacter* and *Salmonella* in this order.

### Groups of close bacterial species

In each application a stratified sampling was used: from each species or biovar random 50 genomes have been sampled. If the number of available genomes was less than 50 then all of them were sampled.

Since the genomes are at species level or below, the ANI-based Jukes-Cantor model distance is used.

The log-log plots of the squared eigenvalues are broken lines due to stratified sampling.

#### Escherichia coli/Shigella

The species of *E. coli* and *S. boydii*, *S. dysenteriae*, *S. flexneri* and *S. sonnei* have been sampled from GenBank on 4/3/17, resulting in 207 genomes. The *E. coli* genomes were required to contain the strain identifier. ANI-based Jukes-Cantor distance was used. The MDS result is in Fig. [Escherichia coli/Shigella].

The light blue cluster in the top part of the plot is the serovar O157:H7 of *E. coli*. The magenta cluster in the center is mostly serovar O104:H4 of *E. coli*. The red cluster in the center is commensal *E. coli*. The cyan cluster in the left part is *S. flexneri*. The green cluster in the bottom-right part is *S. sonnei*. The blue cluster in the right part of the plot is mostly *S. boydii*. The species *S. dysenteriae* had only 11 genomes, split by the blue *S. boydii* and central yellow clusters.

#### Bacillus cereus group

The species group *Bacillus cereus group* consists of species *B. anthracis*, *B. cereus*, *B. mycoides*, *B. thuringiensis*, *B. weihenstephanensis*, *B. toyonensis*, *B. gaemokensis*, *B. manliponensis*, *B. cytotoxicus* and *B. wiedmannii*, deposited in GenBank. Random 50 genomes from *B. anthracis*, *B. cereus* and *B. thuringiensis* and all genomes from the other species as of 3/15/17 are used, totaling 207 genomes. The MDS result is in Fig. [Bacillus cereus group].

*B. anthracis* has 47 genomes in one cluster, red, and the other 3 genomes in the neighboring blue cluster in the left part of the plot. *M. thuringeinsis* is in the yellow and cyan clusters in the upper part of the plot, and *M. cereus* is dispersed all over the plot.

#### Brucella melitensis biovars

The genus *Brucella* has several species grouped as “Brucella melitensis biovars”: *B. abortus*, *B. ovis*, *B. melitensis*, *B. neotomae*, *B. suis*, *B. canis*, *B. pinnipedialis*, *B. ceti*, *B. microti*, *B. vulpis* and *B. inopinata*.

Random 50 genomes from *B. abortus*, *B. melitensis* and *B. suis* and all genomes from the other species as of 3/15/17 are used, totaling 200 genomes. The MDS result is in Fig. [Brucella melitensis biovars].

All genomes of *B. abortus* are in the green cluster in the bottom-left part of the plot. *B. melitensis* has 48 genomes in the red cluster in the left-upper part of the plot. *B. suis* has 47 genomes in the cyan and magenta clusters in the right part of the plot.

#### Mycobacterium tuberculosis complex

The species group *Mycobacterium tuberculosis complex* consists of species *M. tuberculosis*, *M. bovis*, *M. africanum*, *M. caprae*, *M. orygis* and *M. mungi*, deposited in GenBank. Random 50 genomes from the first two species and all genomes from the other species as of 3/15/17 are used, totaling 133 genomes.

The ANI-based Jukes-Cantor distance dANI turned out to be very variable in different parts of the alignment of two genomes. Therefore, dANI was computed not over all alignments, but only over the most typical part of alignments by this procedure:

* the alignment is cut into 10000 bp long pieces;
* the pieces are sorted by their dANI;
* the shortest range of dANI is found containing ≥ 75% of all alignments;
* the dANI of the two genomes is the average dANI in this range.

The MDS result is in Fig. [Mycobacterium tuberculosis complex]. The 3 rays are *M. tuberculosis*, *M. bovis* and *M. africanum*, all other species are in the center.

Using dANI without cutting alignment into pieces resulted in λ3 < 0 and diffused clusters.

## Protein comparisons

The BLOSUM62-based protein transformation distance is used.

### Protein family of class A beta-lactamases (bla-A)

There are 1200 proteins in the family of class A beta-lactamases (NCBI Anti-Microbial Resistance database, January 2017), see NCBI BioProject PRJNA313047.

The MDS result is in Fig [Class A beta-lactamases]. The upper-left red cluster is blaTEM, the bottom-left cyan cluster is blaSHV, the magenta cluster at the right side is blaCTX-M.

### ABC transporters

There are about 3 million proteins in RefSeq whose name contains “ABC transporter”, out of them the 717 proteins whose GI number is divisible by 4000 have been used for analysis.

The canonical projection of the MDS result in in Fig. [ABC transporters].

### Comparison of bacterial genomes by universal protein families

There are 360 bacterial species in GenBank containing ≥ 10 genomes (as of 4/14/2017). In each species a representative genome is selected [O’Leary]. The list of these representative genomes is in [supplemental materials].

There are 106 universal protein families, which exist in each bacterial genome usually as one locus. For these universal protein families there are 108 Pfam [Pfam] or TIGRFAMs protein HMMs [Haft], listed in [supplemental materials]. The protein families are identified by gene symbols.

The representative genomes have been matched against the HMMs of the universal protein families by the NCBI Prokaryotic Genome Annotation Pipeline [O’Leary] by running hmmsearch with parameters –cut\_tc [HMMer]. For each genome-gene symbol pair the protein of the genome hitting the HMM with the highest score was selected resulting in a table of 360 genomes × 106 gene symbols, see in [supplemental materials]. In 201 genomes all 106 universal protein families have been found, and the minimum number of universal protein families found in a genome is 90, so that this table has missing values.

For each universal protein family the pairwise protein transformation distances have been computed, which were transformed into a double-centered similarity matrix. These 106 similarity matrices have been combined into one double-centered similarity matrix. This distance matrix obtained from this similarity matrix can be represented by a tree.

The MDS result of the combined similarity matrix is in Fig. [Bacteria]. The three rays are *Proteobacteria*, *Firmicutes* and *Actinobacteria*.

# Discussion

A close, but different MDS method is metric MDS [Cox]. This method constructs a Euclidean space in real numbers given a distance object-object table, but requires its dimensionality q as a parameter and the optimization is local. The optimization criterion is the weighted sum of residuals which is a function of input and constructed distances. This criterion is referred to as stress. An MDS space with a larger q does not contain an MDS space with a smaller q.

For multiple protein comparisons the distance can be computed from multiple alignment, see, for example, [Pele 2011], but an optimal multiple alignment is an NP-hard problem, usually requires expert knowledge in protein domains and active sites.

[Not finished]

# Funding

This work was supported by …

# Acknowledgements

# References

[Cox] Cox, T.F.; Cox, M.A.A. (2001). Multidimensional Scaling. Chapman and Hall.

[Eckart, Young] C.Eckart, G.Young, The Approximation of One Matrix by Another of Lower Rank. Psychometrika – Vol. 1, No. 3, September, 1936.

[Haft] Daniel H. Haft, Brendan J. Loftus, Delwood L. Richardson, Fan Yang, Jonathan A. Eisen, Ian T. Paulsen, Owen White; TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 2001; 29 (1): 41-43. doi: 10.1093/nar/29.1.41

[HMMer] http://hmmer.org/.

[Jukes-Cantor] Jukes TH & Cantor CR (1969). *Evolution of Protein Molecules*. New York: Academic Press. pp. 21–132. In Mammalian Protein Metabolism, Vol. III, ed. M.N.Munro, Academic Press, New York.

[MegaBlast] Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

[Pele 2011] Pele´ J, Abdi H, Moreau M, Thybert D, Chabbert M (2011) Multidimensional Scaling Reveals the Main Evolutionary Pathways of Class A G-Protein-Coupled Receptors. PLoS ONE 6(4): e19094. doi:10.1371/journal.pone.0019094.

[O’Leary] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional** annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-45

[Pfam] [The Pfam protein families database: towards a more sustainable future](https://nar.oxfordjournals.org/content/44/D1/D279.long): R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman, Nucleic Acids Research (2016)  Database Issue 44:D279-D285

[Richter] [Richter M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Richter%20M%5BAuthor%5D&cauthor=true&cauthor_uid=19855009), [Rosselló-Móra R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rossell%C3%B3-M%C3%B3ra%20R%5BAuthor%5D&cauthor=true&cauthor_uid=19855009). Shifting the genomic gold standard for the prokaryotic species definition. [Proc Natl Acad Sci U S A.](https://www.ncbi.nlm.nih.gov/pubmed/19855009) 2009 Nov 10;106(45):19126-31. doi: 10.1073/pnas.0906412106. Epub 2009 Oct 23.

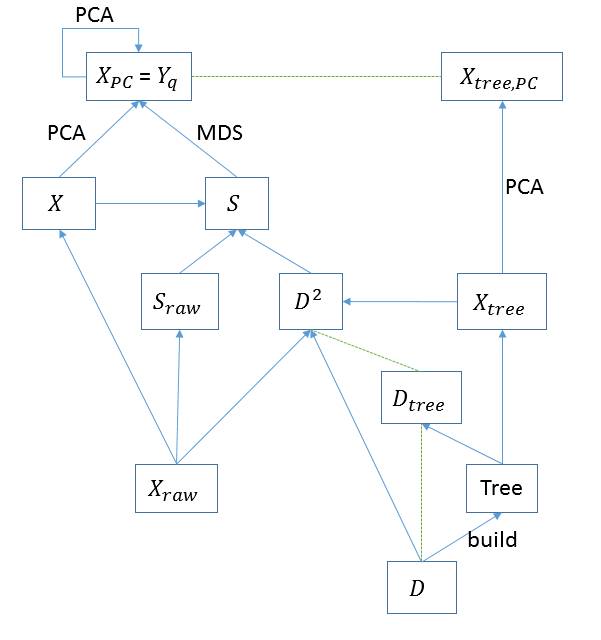
[Setubal, Meidanis] J. Setubal and J. Meidanis, Introduction to Computational Molecular Biology, PWS Publishing Company, 1997.

[Tettelin] [Tettelin H](https://www.ncbi.nlm.nih.gov/pubmed/?term=Tettelin%20H%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Masignani V](https://www.ncbi.nlm.nih.gov/pubmed/?term=Masignani%20V%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Cieslewicz MJ](https://www.ncbi.nlm.nih.gov/pubmed/?term=Cieslewicz%20MJ%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Donati C](https://www.ncbi.nlm.nih.gov/pubmed/?term=Donati%20C%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Medini D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Medini%20D%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Ward NL](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ward%20NL%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Angiuoli SV](https://www.ncbi.nlm.nih.gov/pubmed/?term=Angiuoli%20SV%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Crabtree J](https://www.ncbi.nlm.nih.gov/pubmed/?term=Crabtree%20J%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Jones AL](https://www.ncbi.nlm.nih.gov/pubmed/?term=Jones%20AL%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Durkin AS](https://www.ncbi.nlm.nih.gov/pubmed/?term=Durkin%20AS%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Deboy RT](https://www.ncbi.nlm.nih.gov/pubmed/?term=Deboy%20RT%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Davidsen TM](https://www.ncbi.nlm.nih.gov/pubmed/?term=Davidsen%20TM%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Mora M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Mora%20M%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Scarselli M](https://www.ncbi.nlm.nih.gov/pubmed/?term=Scarselli%20M%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Margarit y Ros I](https://www.ncbi.nlm.nih.gov/pubmed/?term=Margarit%20y%20Ros%20I%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Peterson JD](https://www.ncbi.nlm.nih.gov/pubmed/?term=Peterson%20JD%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Hauser CR](https://www.ncbi.nlm.nih.gov/pubmed/?term=Hauser%20CR%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Sundaram JP](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sundaram%20JP%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Nelson WC](https://www.ncbi.nlm.nih.gov/pubmed/?term=Nelson%20WC%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Madupu R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Madupu%20R%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Brinkac LM](https://www.ncbi.nlm.nih.gov/pubmed/?term=Brinkac%20LM%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Dodson RJ](https://www.ncbi.nlm.nih.gov/pubmed/?term=Dodson%20RJ%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Rosovitz MJ](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rosovitz%20MJ%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Sullivan SA](https://www.ncbi.nlm.nih.gov/pubmed/?term=Sullivan%20SA%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Daugherty SC](https://www.ncbi.nlm.nih.gov/pubmed/?term=Daugherty%20SC%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Haft DH](https://www.ncbi.nlm.nih.gov/pubmed/?term=Haft%20DH%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Selengut J](https://www.ncbi.nlm.nih.gov/pubmed/?term=Selengut%20J%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Gwinn ML](https://www.ncbi.nlm.nih.gov/pubmed/?term=Gwinn%20ML%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Zhou L](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zhou%20L%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Zafar N](https://www.ncbi.nlm.nih.gov/pubmed/?term=Zafar%20N%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Khouri H](https://www.ncbi.nlm.nih.gov/pubmed/?term=Khouri%20H%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Radune D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Radune%20D%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Dimitrov G](https://www.ncbi.nlm.nih.gov/pubmed/?term=Dimitrov%20G%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Watkins K](https://www.ncbi.nlm.nih.gov/pubmed/?term=Watkins%20K%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [O'Connor KJ](https://www.ncbi.nlm.nih.gov/pubmed/?term=O%27Connor%20KJ%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Smith S](https://www.ncbi.nlm.nih.gov/pubmed/?term=Smith%20S%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Utterback TR](https://www.ncbi.nlm.nih.gov/pubmed/?term=Utterback%20TR%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [White O](https://www.ncbi.nlm.nih.gov/pubmed/?term=White%20O%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Rubens CE](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rubens%20CE%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Grandi G](https://www.ncbi.nlm.nih.gov/pubmed/?term=Grandi%20G%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Madoff LC](https://www.ncbi.nlm.nih.gov/pubmed/?term=Madoff%20LC%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Kasper DL](https://www.ncbi.nlm.nih.gov/pubmed/?term=Kasper%20DL%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Telford JL](https://www.ncbi.nlm.nih.gov/pubmed/?term=Telford%20JL%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Wessels MR](https://www.ncbi.nlm.nih.gov/pubmed/?term=Wessels%20MR%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Rappuoli R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Rappuoli%20R%5BAuthor%5D&cauthor=true&cauthor_uid=16172379), [Fraser CM](https://www.ncbi.nlm.nih.gov/pubmed/?term=Fraser%20CM%5BAuthor%5D&cauthor=true&cauthor_uid=16172379). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". [Proc Natl Acad Sci U S A.](https://www.ncbi.nlm.nih.gov/pubmed/16172379) 2005 Sep 27;102(39):13950-5. Epub 2005 Sep 19.

[Young, Householder] G.Young, A.S.Householder, Discussion of a Set of points in Terms of Their Mutual Distances. Psychometrika – Vol. 3, No.1, March, 1938.

# Figures

## Dataflow



Xraw is the original object-attribute table, X is the centered object-attribute table, Sraw is the matrix pf generalized similarities, D2 is the matrix of generalized squared distances, S is the matrix of double-centered generalized similarities, XPC is the PCA space, Yq is the MDS space, D is the matrix of generalized distances, Tree is the tree of objects, Dtree is the matrix of tree distances, Xtree is the Euclidean space defined by the tree, Xtree,PC is the PCA space for Xtree. Blue arcs are computational transformations, green arcs mean identity or approximate identity. This graph shows that Xtree,PC can be achieved from D by the path D = Dtree = D2 → S → Yq = Xtree,PC via MDS, without building the Tree.

## Cities without noise

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\cities.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 97.66 | 97.66 | | 2 | 2.27 | 99.92 | |

## Cities with noise in the data used for distance measurement

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ ,%** | | 1 | 25.30 | 25.30 | | 2 | 12.59 | 37.90 | | 3 | 10.82 | 48.72 | | 4 | 10.71 | 59.43 | | 5 | 10.51 | 69.94 | | 6 | 10.38 | 80.32 | | 7 | 9.99 | 90.31 | | 8 | 9.69 | 100.00 | |

## Cities with noise in the distance measurement

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ ,%** | | 1 | 79.67 | 79.67 | | 2 | 9.14 | 88.80 | | 3 | (-)5.01 | 93.81 | | 4 | 3.93 | 97.74 | | 5 | (-)1.75 | 99.49 | |

## Enterobacteriaceae

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\Enterobacteriaceae.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 53.99 | 53.99 | | 2 | 24.31 | 78.30 | | 3 | 6.63 | 84.93 | | 4 | 3.68 | 88.61 | | 5 | 2.55 | 91.17 | | 6 | 1.75 | 92.92 | | 7 | 1.38 | 94.30 | | 8 | 0.95 | 95.25 | | 9 | 0.89 | 96.13 | | 10 | 0.67 | 96.80 | | 11 | 0.57 | 97.37 | |

## *Escherichia coli/Shigella*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\Ecoli.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 47.36 | 47.36 | | 2 | 26.01 | 73.36 | | 3 | 17.05 | 90.42 | | 4 | 4.35 | 94.77 | | 5 | 2.60 | 97.37 | | 6 | 1.16 | 98.52 | |

## *Bacillus cereus group*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\Bacillus_cereus.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 52.91 | 52.91 | | 2 | 29.88 | 82.79 | | 3 | 8.80 | 91.59 | | 4 | 2.91 | 94.50 | | 5 | 1.52 | 96.02 | | 6 | 0.87 | 96.89 | | 7 | 0.67 | 97.57 | | 8 | 0.61 | 98.18 | |

## *Brucella melitensis* biovars

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\Brucella.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 52.61 | 52.61 | | 2 | 25.82 | 78.42 | | 3 | 11.33 | 89.76 | | 4 | 7.51 | 97.27 | | 5 | 0.82 | 98.09 | | 6 | 0.62 | 98.71 | |

## *Mycobacterium tuberculosis* complex

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\Mtb.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 72.67 | 72.67 | | 2 | 21.61 | 94.28 | | 3 | 1.81 | 96.10 | | 4 | 1.30 | 97.40 | | 5 | 0.54 | 97.94 | |

## Class A beta-lactamases

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\bla-A.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 68.87 | 68.87 | | 2 | 11.80 | 80.67 | | 3 | 10.38 | 91.05 | | 4 | 2.89 | 93.94 | | 5 | 2.03 | 95.97 | | 6 | 1.05 | 97.02 | | 7 | 0.55 | 97.57 | |

## ABC Transporters

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ ,%** | | 1 | 61.05 | 61.05 | | 2 | 8.66 | 69.71 | | 3 | 2.91 | 72.61 | | 4 | 1.80 | 74.41 | | 5 | 1.73 | 76.14 | | 6 | 1.23 | 77.37 | | 7 | 0.87 | 78.23 | | 8 | 0.78 | 79.02 | | 9 | 0.68 | 79.70 | | 10 | 0.61 | 80.30 | |

## All Bacteria compared by universal protein families

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **C:\Users\brovervv\work\MDS\Paper\Bacteria-univ.JPG** | |  |  |  | | --- | --- | --- | | ***i*** | **λ2,%** | **Σ, %** | | 1 | 59.09 | 59.09 | | 2 | 13.83 | 72.92 | | 3 | 6.40 | 79.31 | | 4 | 3.74 | 83.05 | | 5 | 2.72 | 85.77 | | 6 | 2.15 | 87.92 | | 7 | 1.87 | 89.79 | | 8 | 1.08 | 90.88 | | 9 | 1.05 | 91.92 | | 10 | 0.81 | 92.73 | | 11 | 0.74 | 93.48 | | 12 | 0.66 | 94.13 | | 13 | 0.63 | 94.77 | | 14 | 0.51 | 95.28 | |

