

## The choice of optimal distance measure in genome-wide data sets

Galina Glazko<sup>1,4</sup>, Alexander Gordon<sup>2</sup>, and Arcady Mushegian<sup>1,3</sup>

<sup>1</sup>*Stowers Institute for Medical Research, 1000 E 50<sup>th</sup> St., Kansas City MO 64110,*

<sup>2</sup>*Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642 and* <sup>3</sup>*Department of Microbiology, Molecular Genetics, and Immunology, University of Kansas Medical Center, Kansas City, KS 66160, USA*

<sup>4</sup>Corresponding author.

E-MAIL [gvv@stowers-institute.org](mailto:gvv@stowers-institute.org)

## ABSTRACT

**Motivation:** Many types of genomic data are naturally represented as binary vectors. Numerous tasks in computational biology can be cast as analysis of relationships between these vectors, and the first step is frequently to compute their pairwise distance matrix. Many distance measures have been proposed in the literature, but there is no theory justifying the choice of distance measure.

**Results:** We examine the approaches to measuring distances between binary vectors and study the characteristic properties of various distance measures and their performance in several tasks of genome analysis. Most distance measures between binary vectors turn out to belong to a single parametric family, namely generalized average-based distance with different exponents. We show that descriptive statistics of distance distribution, such as skewness and kurtosis, can guide the appropriate choice of the exponent. On the contrary, the more familiar distance properties, such as metric and additivity, appear to have much less effect on the performance of distances.

**Availability:** R code GADIST is available from the corresponding author upon request.

## INTRODUCTION

Vigorous genome sequencing and analysis in recent years has produced a lot of genome-wide data, substantial portion of which takes the form of binary vectors. Some examples of binary vectors in genomics are: presences and absences of genes in complete genomes (Tatusov et al. 1997); recoded gene expression data, for example where 1 stands for a change in gene expression, and 0 stands for no change, or when different tissues are interrogated for detectable levels of expression of each gene (Shmulevich and Zhang 2002); gene/protein interaction matrices, where 1 stands for registered interaction, (in a systems-biology network, this is presented as an edge), and 0 stands for lack of interaction (Bader and Hogue 2002; Krause et al. 2003; Lesage et al. 2005; Wong et al. 2004).

The frequent purpose of genome-scale data analysis is to uncover the subsets in the data that are related by a similarity of some sort. One way to do it is by computing the distances between vectors. The major question here is: how to choose the distance measure, when several of them are available?

The choice of the distance measure is more straightforward when there is a hypothetical model, or several competing hypotheses, about the process that generates data. In that case, the likelihood of the models given the data can be compared, and the best model can be selected using, for example, likelihood ratio test (Posada and Crandall 2001). This approach is applied, for example, when genome-wide data are used to study evolutionary relationships: one compares several models of character evolution, determines the likelihood of each model given the phylogenetic tree, and selects the best model, in this way obtaining the most plausible distance measure between characters. For many kinds of genome-wide data, however, there is no causative process model at all.

The choice of the distance measure can be also facilitated by a training sample, which contains information about desired biological properties of the solution. For example, assuming that the objective of clustering of gene expression vectors is to find groups of functionally linked genes, one may choose the distance measure empirically, by answering the question “Which distance measure yields the clusters with the highest proportion of genes known to be functionally linked?” (Gibbons and Roth 2002).

In this study we examine distance measures from yet another angle. We survey the properties of the distances between binary vectors representing different types of genome-wide data, and ask whether there are any commonalities in distance measures that prove to be the most successful in revealing the relationships between genes and genomes. We compare the performance of various distances on three specific problems, namely phylogenetic inference from gene content (Problem 1), protein-protein interaction inference from the co-purification data (Problem 2) and inference of tight clusters of periodically expressed genes from yeast cell-cycle expression data (Problem 3). We find that most distances between binary vectors proposed in the literature belong to a single parametric family, namely generalized average-based distance with different exponents. Virtually every available distance measure is neither metric nor additive, yet, unexpectedly, the extent of deviation from those properties does not correlate with the ability of the distance measure to arrive at a biologically correct solution. On the other hand, skewness and kurtosis, two statistics describing the shape of the distance distribution, show good correlation with the recovery of correct solution: namely, the best-performing distance measure tends to have the most extreme values of these statistics. We have written a program that examines a range of distance measures for a given data set, and selects a distance with the best fit to these criteria.

## RESULTS

### Definitions

Let  $X$  be a set of  $n$  elements and let  $d_{ij}$  be an arbitrary measure of the proximity of two elements  $i, j \in X$ . If  $d_{ij} = d(i, j)$  is a non-negative real function  $d: X \times X \rightarrow \mathbb{R}^+$  and satisfies (1)  $d_{ij} > 0$  for  $i \neq j$ ; (2)  $d_{ij} = 0$  for  $i=j$ ; (3)  $d_{ij} = d_{ji}$  for all  $i, j$ , then  $d$  is a distance measure (DM) and  $D = \{d_{ij}\}$  is a distance matrix. If, in addition,  $d_{ij}$  satisfies triangle inequality (4)  $d_{ij} \leq d_{ik} + d_{kj}$ , then  $d$  is a metric. An unrooted tree  $T$  is a connected graph of  $n$  vertices of degree 1 or 3 with no cycles and with  $2n-3$  edges (branches); a nonnegative real number (branch length) is assigned to every branch. If for a distance matrix  $D = \{d_{ij}\}$  there exists a tree  $T$  such that sum of branch lengths along the shortest path between terminal vertices  $i, j$  equals to  $d_{ij}$  for all  $i, j$ ,  $D$  is said to be additive. (Note that we use a tree for an operational definition of additivity, but the problems that we study do not always require a tree-like clustering solution). A necessary and sufficient condition for additivity of  $D$  (or, in other words, for  $d$  to be a tree metric) is the four-points condition found by Zaretsky (Zaretsky 1965) and Buneman (Buneman 1974): for all sets of four elements there exists some labeling  $i, j, k, l \in X$  such that (5)  $d_{ij} + d_{kl} = d_{jl} + d_{ki} \geq d_{il} + d_{jk}$ . Further,  $D$  is said to be ultrametric if the three-points condition holds: for any three elements,  $i, j, k \in X$  the two closest elements  $i, j$  are at the same distance from the third element, that is  $d_{ij} \leq d_{ik} = d_{kj}$ . Ultrametric condition is stronger than additivity, and additivity is stronger than metric property.

Consider the random sample of  $n$  observed distances  $d_1, \dots, d_n$  from  $D$ . Mean  $\bar{d}$ , variance  $s^2$ , skeweness  $\sqrt{b_1}$  and kurtosis  $b_2$  of the distribution  $f_D$  are four statistics characterizing  $f_D$ , computed as follows:

$$\bar{d} = \frac{\sum_i d_i}{n}, \quad s^2 = \frac{\sum_i (d_i - \bar{d})^2}{n-1}, \quad \sqrt{b_1} = \frac{\sqrt{n} \sum_i (d_i - \bar{d})^3}{[\sum_i (d_i - \bar{d})^2]^{3/2}}, \quad b_2 = \frac{n \sum_i (d_i - \bar{d})^4}{[\sum_i (d_i - \bar{d})^2]^2} - 3 \quad (\text{Rencher}$$

2002). When the population is normally distributed,  $\sqrt{b_1}$  and  $b_2$  both equal 0. If  $\sqrt{b_1} <(>)0$ , the distribution  $f_D$  is skewed to the left (right) and has long left (right) tail. When  $b_2 <(>)0$ , we say that the distribution has negative (positive) kurtosis. A distribution with negative kurtosis is flatter than normal, is less peaked, and with heavier flanks and thinner tails. A distribution with positive kurtosis has a higher peak than the normal, with an excess of values near the mean and in the tails, but with thinner flanks (Rencher 2002) (e.g. Fig. 1, distribution with  $exp = -2$ ). We will use these statistics to characterize the distance distributions for different distance measures.

### **Data sets and methods**

Vectors of presences and absences of orthologous genes in completely sequenced genomes (phyletic vectors, also known as “phyletic patterns” or “phylogenetic profiles”), have several uses. One use involves studying the distances between genes in the species space - gene vectors separated by short distance indicate that these genes tend to co-occur in the same sets of genomes, which may suggest functional interactions between these genes (Date and Marcotte 2003; Pellegrini et al. 1999; Strong et al. 2003; von Mering et al. 2003). Another use is to examine the distances between species in the gene space. In that case, small distance between species’ vectors indicates that the species are phylogenetically close (Dutilh et al. 2004; Korbelt et al. 2002; Wolf et al. 2002). Problem 1 below is an example of phylogenetic inference from these species’ vectors.

Presences and absences of protein in biochemical purifications is another binary genome-wide data set. Methods of large-scale protein-protein interaction (PPI) screening include yeast two-hybrid technology (Uetz et al. 2000), which registers only pairwise PPIs, and various affinity purification schemes (Gavin et al. 2002), which record the content of protein complexes, but do not specify which protein pairs interact directly. Independent biochemical purifications that contain the same protein typically share additional proteins. Comparison of multiple purifications can aid in separating spurious co-purifications from stable, functionally relevant protein complexes (Gavin et al. 2002). Unsupervised clustering of binary vectors has been used recently to extract non-redundant protein complexes from noisy purification data (Krause et al. 2003). A related strategy is to construct interaction graph from binary protein-protein interaction vectors and select dense clusters of high connectivity in this graph (Bader and Hogue 2003; Brun et al. 2004; King et al. 2004). Our Problem 2 addresses the most appropriate way of measuring distances between binary vectors of PPIs, which is a prerequisite of any clustering.

Our third case study deals with gene expression vectors recoded into a binary form. While binary recoding leads to the loss of quantitative information about expression, it facilitates the comparison of data from different expression platforms, simplifies the data representation, and in addition, is more compatible with many machine-learning algorithms than the interval-valued attributes (Pfahringer 1995). Recently it has been shown that different tumor types can be perfectly separated using binary recoded expression vectors (Shmulevich and Zhang 2002), confirming that this transformation preserves information content of expression vectors. Our Problem 3 addresses a frequently asked question in gene

expression analysis, i.e., which parameter setting gives clusters with the highest content of relevant genes.

*Problem 1: gene content-based phylogeny of Proteobacteria*

In this Problem, the set of binary vectors of the form  $\mathbf{X}_i=(x_{i1},x_{i2},\dots,x_{iN})$  represents the gene content of  $N$  genomes, where  $x_{ik}$  is 1 if  $k$ th gene is present in  $i$ th genome and 0 otherwise. The phylogeny of Proteobacteria has been extensively studied using a combination of many molecular traits, including several types of gene content-based trees, protein sequence phylogenies, and 16S rRNA tree (Fig. 2; (Korbel et al. 2002; Lerat et al. 2003; Wolf et al. 2001). Our goal is to study the performance of different distance measures in recovering this phylogeny.

Information on gene presences and absences was taken from Clusters of Orthologous Group Database (COG DB, <http://www.ncbi.nlm.nih.gov/COG/new/>), containing 4589 prokaryotic COGs. The number of COGs found in more than one proteobacterial genome (the  $k$  value) is 625, and the number of genomes  $N$  is 24. We used neighbor-joining algorithm (Saitou and Nei 1987) to build the tree using different distances between gene content vectors. To estimate performance of different distances, we compared the trees to a consensus tree, modeled by phylogeny inferred from 16S rRNA using the Kimura distance (Kimura 1980) (the latter tree is almost the same as the tree based on the concatenation of 205 protein families – see Fig. 2 in (Lerat et al. 2003) – with necessary correction for taxon sampling).

*Problem 2: protein complex assembly from TAP data*

In this Problem, the binary vectors  $\mathbf{X}_i=(x_{i1},x_{i2},\dots,x_{iN})$ , represent the set of protein purifications for  $N$  proteins, where  $x_{ik}$  is 1 if  $k$ th protein is present in  $i$ th purification and 0

otherwise. We used TAP data set from Gavin et al. (Gavin et al. 2002) and removed purifications that retrieved nothing but the bait, retaining 455 purifications that contain 1361 proteins ( $i=455, N=1361$ ). We wish to distinguish signal (true protein complexes) from noise (spurious co-purifications). Our Problem 2 is to find a distance which gives best coverage of already known MIPS protein complexes. To estimate performance of different distance measures, we compared purifications vectors, clustered using the UPGMA method with different similarity cutoffs, with the set of 252 predicted TAP complexes (Krause et al. 2003); this set is the best coverage of MIPS collection of protein complexes (Mewes et al. 2002), obtained for TAP data (Krause et al.).

*Problem 3: partition clustering of gene expression vectors.*

In Problem 3,  $\mathbf{X}_i=(x_{i1}, x_{i2}, \dots, x_{iN})$ , represents the well-studied Cdc28 data set, monitoring gene expression throughout the yeast cell cycle (Cho et al. 1998), with  $i$  gene expression values over  $N$  time points ( $i=6214, N=17$ ). Several hundreds of genes are expressed at the particular phases of yeast mitotic cell cycle, in four distinct waves (Rustici et al. 2004). The first wave is expressed at the start of S phase (G1/S), the second wave during S phase, the third during G2 and the fourth wave at the end of M-phase and the start of the next cell cycle. It is known that many genes are specifically expressed in one of these waves, and we want to recover groups of these wave-specific genes as different clusters defined by the similarity of their expression profiles. Our Problem 3 is to find a distance resulting in the tight clusters with the highest number of known periodically expressed genes. As a benchmark set of known periodically expressed genes we used combined benchmark set ( $B_{123}$ ) suggested in Lichtenberg et al. (de Lichtenberg et al. 2005) (see ref. for the detail of benchmark construction). Best-performing distance would be the one that maximizes the sum

of ratios (score): size of in-cluster benchmark ( $NB_i$ ) (i.e. the number of genes from  $B_{123}$  found in cluster  $i$ ), to the cluster size ( $NCL_i$ ) over all clusters  $d_{best} = \arg \max_d (\sum_i NB_i / NCL_i)$ . For consistency, we fixed the number of clusters ( $K=4$ ) and clustering algorithm (partitioning around medoids, PAM (Kaufman and Rousseeuw 1990)): in this way only differences in distances performance would influence the score. In order to find the best-performing distance measure, we compared score values for clustering solutions provided by different distances.

Renormalized Cdc28 data set (obtained from <http://www.cbs.dtu.dk/cellcycle>, (de Lichtenberg et al. 2005)) was processed by replacing the outlying expression values (deviating more than 3 s.d. from the mean) by the mean. After that, all genes with the variance smaller than the variance in the upper quartile (75<sup>th</sup> percentile) of the variance distribution for the entire gene pool were removed. To convert the coordinates into a binary form, average expression value  $\bar{x}_i$  was calculated for every gene  $i$ , and every expression value  $x_i$  was set to 1 if it was more than  $\bar{x}_i$  and 0 otherwise. By this binary recoding around the mean, we transform the hypothetical sine wave, which may be expected of many genes in the cyclic process, into rectangular wave.

### **Types of distance measures**

Sneath and Sokal (Sneath and Sokal 1973) divide all distance measures into several types: distance coefficients (e.g., Euclidean distance), association coefficients (e.g., Jaccard coefficient), correlation coefficients, and probabilistic coefficients (e.g., an information-theoretical measure). Sometimes, a measure of one type can be easily expressed in terms of a measure of another type; for example, the matching coefficient (association coefficient) is a

complement to the Hamming distance (distance coefficient). In what follows we discuss several popular similarity coefficients and distance measures.

The Jaccard coefficient for a pair of vectors,  $\mathbf{X}_m$  and  $\mathbf{X}_n$  is

$$J_{mn} = \frac{X_{mn}}{X_{mm} + X_{nn} - X_{mn}} \quad (1)$$

where  $X_{ij} = \mathbf{X}_i \cdot \mathbf{X}_j$  (the dot product of two vectors). The value of  $J_{mn}$  ranges from 0 to 1 and is equal to the number of bits “on” in both vectors, divided by the number of bits “on” in either vector. The distance measure based on  $J_{mn}$ ,  $d_{J_{mn}} = 1 - J_{mn}$ , is called the Soergel distance and satisfies the triangle inequality (Lipkus 1999), i.e.

$$d_{J_{mn}} \leq d_{J_{mk}} + d_{J_{kn}}.$$

Mirkin and Koonin (Mirkin and Koonin 2003) have noted that, when phyletic vectors of gene content in completely sequenced genomes are compared, Jaccard coefficient systematically underestimates the similarity between genomes. If two genomes have approximately the same size and half of the genes in each genome is also found in the other genome, one would expect the similarity to be about 1/2, whereas the Jaccard coefficient gives counterintuitive 1/3 (Mirkin and Koonin 2003). Mirkin and Koonin (Mirkin and Koonin 2003) suggested what they called the Maryland Bridge coefficient, representing the average proportion of the overlap in the genomes,

$$MB_{mn} = \frac{1}{2} \left( \frac{X_{mn}}{X_{mm}} + \frac{X_{mn}}{X_{nn}} \right).$$

$MB_{mn}$  is free of the aforementioned artifact. Another similarity coefficient has been suggested by Korbel et al. (Korbel et al. 2002):

$$WA_{mn} = \frac{\sqrt{X_{mm}^2 + X_{nn}^2}}{\sqrt{2} X_{mm} X_{nn}} X_{mn}.$$

Dutilh et al. (Dutilh et al. 2004) observed that  $WA_{mn}$  worked better in phylogeny inference than several other coefficients, based on different kinds of normalization by genome sizes.

We note that  $MB_{mn}$  and  $WA_{mn}$  are instances of the following parametric family:

$$A_{I,mn} = \frac{X_{mn}}{B_I}, \quad -\infty < I < \infty, \quad (I \neq 0) \quad (2),$$

where

$$B_I = \left( \frac{X_{mm}^I + X_{nn}^I}{2} \right)^{\frac{1}{I}}$$

is the generalized average cardinality of two sets, of exponent  $I$ .  $MB_{mn}$  and  $WA_{mn}$  indices can be obtained from (2), if  $I=-1$  and  $I=-2$ , respectively. Other notable similarity coefficients have the following limiting values of  $I$ :

$$\begin{aligned} I \rightarrow \infty, \quad B_I &\rightarrow \max(X_{mm}, X_{nn}), \quad A_{\infty,mn} = \frac{X_{mn}}{\max(X_{mm}, X_{nn})}; \\ I \rightarrow -\infty, \quad B_I &\rightarrow \min(X_{mm}, X_{nn}), \quad A_{-\infty,mn} = \frac{X_{mn}}{\min(X_{mm}, X_{nn})}; \\ I \rightarrow 0, \quad B_I &\rightarrow \sqrt{X_{mm} X_{nn}}, \quad A_{0,mn} = \frac{X_{mn}}{\sqrt{X_{mm} X_{nn}}}. \end{aligned} \quad (3)$$

The similarity coefficient with  $I=-\infty$  is available for phylogenetic inference from gene content in SHOT software (Korbel et al. 2002). This coefficient is in effect the number of shared orthologs divided by the size of the smallest of the two genomes (maximum possible number of shared orthologs). In fact,  $A_{-\infty,mn}$  is the Simpson similarity index, frequently used in different areas of biology (Pielou 1975; Sneath and Sokal 1973).

For  $I = 0$ ,  $I = -1$  and  $I = 1$ ,  $B_I$  represents, respectively, the geometric, harmonic and arithmetic average cardinalities of the two sets.  $B_I$  increases with the increase of  $I$ , therefore,  $MB_{mn} \leq WA_{mn}$ . We call the distance measures based on  $A_{I,mn}$ , namely  $d_{A_{I,mn}} = 1 - A_{I,mn}$ ,

“the  $A_I$ - distance family” (or generalized average-based distance measures). It can be shown that  $d_{A_I}$  for  $I=-1$  and  $I=-2$  ( $1-MB_{mn}$  and  $1-WA_{mn}$ , respectively) are not metric (triangle inequality does not hold).

In the past, various similarity indices, corresponding to different values of  $I$ , have been introduced in connection with clustering tasks. As we already mentioned,  $A_{-\infty, mn}$  is the Simpson similarity index; for  $I=1$ ,  $A_{1, mn}$  is the arithmetic average cardinality of two sets, and is known as Dice similarity index. Dice and Jaccard similarity indices are related by

$$A_{1, mn} = \frac{2J_{mn}}{1 + J_{mn}}, \text{ or } d_{A_1} = \frac{d_{J_{mn}}}{2 - d_{J_{mn}}} \text{ (Kosman and Leonard 2005).}$$

Therefore, for the case of binary vectors, the  $A_I$ -distance family covers many popular measures, arising from distance and association coefficients, mentioned by Sneath and Sokal (Sneath and Sokal 1973).

The standard correlation coefficient is frequently used to measure the similarity between vectors (e.g. (Eisen et al. 1998)). For the pair of binary vectors  $\mathbf{X}_m$  and  $\mathbf{X}_n$ , the correlation coefficient can be written as

$$r(X_m, X_n) = \frac{(X_{mn} - n\bar{X}_m\bar{X}_n)}{\sqrt{(X_{mm} - n\bar{X}_m^2)(X_{nn} - n\bar{X}_n^2)}} \quad (4)$$

and is not covered by the  $A_I$ -distance family. Probabilistic coefficients, such as the mutual information, which tends to give results close to those obtained using the correlation coefficient (Glazko and Mushegian 2004), are also not covered by the  $A_I$ -distance family.

In pattern recognition tasks, the distance measures that satisfy the triangle inequality (i.e, they are metrics) are often recommended (Arkin et al. 1991), and distance additivity is considered advantageous in phylogenetic inference (Gusfield 1999; Waterman et al. 1977). One reason for such preference is that if the distance is additive,

there exists a unique tree corresponding to the distance matrix, and this tree can be constructed from the distance matrix in a  $O(n^2)$  time (Gusfield 1999; Waterman et al. 1977); in contrast, the problem of reconstructing tree from a non-additive distance matrix is *NP*-hard (Gusfield 1999).

Though the relationship between additivity/metric properties and the ability of certain algorithms to arrive at a unique solution in polynomial time is well-studied, distances between real-life data points typically do not satisfy either of these conditions; hence the interest in employing distances with improved properties, which would make them closer to metric or additive distances (Atchley et al. 2005; Xu and Miranker 2004). A relevant question in various clustering tasks, however, is whether the extent of deviation from additivity is a good predictor of this distance's performance, or whether there are better predictors.

### **Performance of different distance measures**

In this section we assess performance of different distance measures (DM) in phylogenetic inference from gene content vectors, in protein complexes assembly and in detecting tight clusters of periodically expressed genes. In all three cases, the best DM is defined as the distance measure giving the result closest to the known correct (or, at least, consensus) answer.

#### *Problem 1: Proteobacterial Phylogeny*

We compared the gene content-based trees to the consensus Proteobacterial tree (Fig. 2). The gene-content trees were inferred using  $d_{A^l}$  with different exponents ( $l = -\infty, -3, -2, -$

1, 0, 1,  $+\infty$ ), as well as correlation-based DM and  $d_{J_{mn}} \cdot (1 - J_{mn})$ . The Robinson-Foulds criterion (Robinson and Foulds 1981), as implemented in PAUP\*, was used to compute the differences between the trees. All trees can be found in the Supplementary Information. The smallest tree difference was observed for  $d_{AI}$  with  $I = -2$  ( $1 - WA_{mn}$ ), and the next smallest was for  $I = -1$  ( $1 - MB_{mn}$ ). We also computed correlation  $r$  between the distance matrices (Fig. 3, light blue bars). The highest correlations were again observed for distance measures with  $I = -2$ ,  $I = -1$ . For correlation-based DM, the  $r$  value was also high (0.77).  $d_{AI}$  with  $I \geq 0$  all had low correlation with  $d_K$ . Distance measures with  $I \geq 0$  attach more importance to shared presence of 1s, weighting shared genes by a factor of  $2^I$ . Generally this would make two larger genomes more similar to each other, than any large and small genome. Probably for that reason, DM with  $I \geq 0$  frequently give rise to clustering artifacts, grouping together genomes of similar sizes (Supplementary Information) (Korbel et al. 2002). Thus, even within the same family of distance measures, certain values of the  $I$  parameter appear to induce more artifacts than the others.

We then quantified the deviation from additivity for different distance measures between gene content vectors of proteobacteria, using the  $d$ -plots approach (Holland et al. 2002). In this approach, for every quartet of taxa (external nodes)  $q$ , a quantity  $d_q$  ( $0 \leq d_q \leq 1$ ) is computed, indicating how much  $q$  violates four-point condition, which is necessary condition for distance to be additive (Buneman 1974; Zaretsky 1965). The  $d$ -plot for all distances is shown in Fig. 3 (light-gray bars). Among generalized average-based distances, the smallest deviation from additivity is observed for the distance with  $I = \infty$  and the largest

for the complement to Simpson similarity index ( $d_{A_I}$  with  $I = -\infty$ ).  $d_q$  for Kimura distance and for  $A_I$ -distances are a statistical tie.

Interestingly, we did not observe any significant negative correlation between distance additivity and correct topology (the correlation between  $d_q(d_{A_I})$  and  $\text{corr}(d_{A_I}, d_K)$  is -0.145). Thus, distance additivity is weakly correlated to distance performance and can not drive the choice of distance. It should be noted that correlation reflects linear relationship and the relationship between distance performance and additivity may be more complex.

Different distance measures emphasize different underlying properties of the data. As we have already noted,  $d_{A_I}$  with positive  $I$ 's increase the weight of shared entities. Negative  $I$  values tend to balance the shared 1's and the total number of genes. The distance measures distributions clearly expose differences between distances; for example, when  $I = 2$  or  $I = 1$ , there is a tendency of averaging all distances, whereas at  $I = -2$  or  $-1$ , distributions become more skewed (Fig. 1). Examination of the moments for these distributions (Table 1) indicates that there is a marked difference between their behavior; in particular, skewness and kurtosis of distribution appear to distinguish quite well between distances with different performance.

Perhaps unexpectedly, more familiar descriptive statistics, such as mean, median and variance, are not that different for different distance distributions (Table 1). The kurtosis of the best performing distance is five times larger than that of the worse performing one, whereas their mean ratio is 1.5. The absolute values of skewness and kurtosis gradually increase with the growth of the exponent from  $I = -\infty$  to  $I = -2$  and then gradually decrease from  $I = -2$  to  $I = \infty$  (Fig. 4).

The correlation-based DM has the highest absolute values of skewness and kurtosis. The best performing  $d_{AI}$  with  $I = -2$  (Fig. 4) has the second highest values of these statistics. Thus, for phylogenetic analysis, the best distance measure is the one that effectively polarizes pairwise distance measures, with only few DMs close to the average.

*Problem 2: Distinguishing true protein-protein interactions from the spurious ones*

Clustering of affinity purification vectors (Krause et al. 2003) may help us to define a native multiprotein complex, as illustrated by the following example. Let  $x_i$  stand for  $i^{th}$  protein and  $X_j$  – for the  $j^{th}$  purification. Suppose that we see the following evidence:  $X_1 = (x_1, x_2)$ ,  $X_2 = (x_1, x_2, x_3, x_4)$  and  $X_3 = (x_1, x_3, x_5, x_6)$ . Intuitively, we may expect  $d_M(X_1, X_2) < d_M(X_2, X_3)$ : if  $x_3, x_4$  are abundant proteins, then  $X_2$  is more likely to be a noisy realization of  $X_1$ , and  $X_3$  is more likely to be a separate complex. Therefore a good distance measure should be able to cluster  $X_1$  and  $X_2$  to the exclusion of  $X_3$ .

The best performing distance should also reveal the largest number of known protein complexes. We explore which parameter settings maximize this number, as follows. First, we use UPGMA to cluster purification vectors with different maximum allowed pairwise distance between in-cluster purifications (parameter1). Second, we explore several distance measures (parameter2). Third, we find the combination of parameter1 and parameter2 giving the highest coverage of known annotated protein complexes. As shown in Figure 5, the best-performing distance was derived from Simpson similarity index with 0.5 distance threshold.

We then quantified the deviation from additivity for different distance measures between protein purification vectors, using the  $d$ -plots approach (Holland et al. 2002). Among the generalized average-based distances, the smallest deviation from additivity is

observed for generalized-average based distance measure with  $I = -2$  and the largest for correlation-based distance. Similar to Problem 1, there is no correlation between distance performance and deviation from additivity.

Examination of distance distribution for Problem 2 indicates that skewness and kurtosis gradually increase with the growth of the exponent from  $I = -\infty$  to  $I = +\infty$ . They reach maxima for the distance derived from Jaccard similarity index. For this problem, the best-performing distance (Simpson similarity index,  $I = -\infty$ ) has the smallest absolute values of skewness and kurtosis (Table 1). Descriptive statistics such as mean, median and variance were again more similar for different distance distributions.

### *Problem 3: Tight clusters of periodically expressed genes*

Experimental evidence suggests that during yeast mitotic cell cycle hundreds of genes are expressed only at a particular phase (Cho et al. 1998; Rustici et al. 2004). One way to simultaneously discover several groups of genes with distinct phase specificity is to cluster the space of expression vectors. The choice of similarity measure is important for the successful clustering. We applied PAM clustering to binary recoded *cdc28* data set with four clusters, using 7 generalized-average based distances ( $\lambda = -\infty, -2, -1, 0, 1, 2, +\infty$ ), as well as correlation-based distance and distance derived from Jaccard similarity index. According to score criterion, the distance measure derived from Jaccard similarity index performed better than all the others (Table 2).

As before, we did not observe the correlation between distance additivity and its performance. However, and also as before, the best performing distance measure ( $d_{J_{mn}}$ ) has the largest absolute values of kurtosis and skewness (Table 1). Similar to previous examples,

first and second moments of distance distributions were more similar among different distance distributions, while the values of skewness and kurtosis were more widely spread.

Thus, the best performing distance in Problems 2 had the smallest values of skewness and kurtosis, while the best performing distance in Problem 3 and 1 had the highest and the second-highest absolute skewness and kurtosis. It appears that the extreme values of these statistics correlate with distance performance in contrast to the values of first two moments, or metric and additivity properties.

## **DISCUSSION**

Many distances between multidimensional vectors have been introduced, and it has been noticed that performance of a measure varies with the dataset (see for review ref. (Wilson and Martinez 1997)). The choice of the measure should depend on the goal of the analysis and should be informed by the properties of the dataset, but “there is no theory how to choose the best distance measure” (Brazma and Vilo 2000). In this study, we examined the characteristic properties of several types of distances between binary vectors, and found, that many popular distances between binary vectors can be represented via generalized average with different exponents. We studied the performance of different distance measures for three problems: phylogenetic inference by gene content, protein complex assembly and tight clustering of gene expression vectors.

We tested the behavior of distance function under different values of the exponent ( $I$  parameter), and noticed the trend that may be useful for choosing the distance measure, namely, that the best performing distance measures tend to highly polarize distances between vectors, with only few distances close to the average. This property is reflected in the

absolute values of the skewness and kurtosis of distance distribution, which tend to lie close to the extremes. Notably, this trend was observed in all three Problems that we examined, though the optimal distance measures were all different, but all can be obtained from the vast distance measure family defined here.

In conclusion, we suggest that the analysis of distributions for different distance measures may be a useful preliminary step of any task that involves clustering vectors with high dimensionality. We would also expect that selection of a few measures with the extreme values of skewness and kurtosis tend to produce results close to optimal in many other cases. The reasons why extreme values of skewness and kurtosis appear to be good predictors of the suitability of a distance measure remain to be studied.

## FIGURE LEGENDS

Figure 1. Distance distributions with different exponents (red lines) for Problem 1. Abbreviation 'exp' stands for the exponent in  $A_I$ -distance family (see main text);  $exp=-inf$  and  $exp=inf$  correspond to  $I = -\infty$  and  $I = \infty$ . Distribution of correlation-based distance is shown in the intersection of the first row and last column. Blue lines show the distribution of 10,000 random numbers generated from normal distributions with mean and standard deviation of the corresponding distance sample.

Figure 2. Phylogenetic tree of Proteobacteria built from aligned 16 sRNA sequences, using Kimura distance and NJ algorithm (see text). Bootstrap support percentages are shown next to each branch. Three letter species abbreviations are as follows: **a-Proteobacteria:** *Agrobacterium tumefaciens* strain C58 (Atu), *Sinorhizobium meliloti* (Sme), *Brucella melitensis* (Bme), *Mesorhizobium loti* (Mlo), *Caulobacter crescentus* CB15 (Ccr), *Rickettsia prowazekii* (Rpr), *Rickettsia conorii* (Rco); **b-Proteobacteria:** *Neisseria meningitidis* MC58 (Nme), *Neisseria meningitidis* Z2491 (NmA), *Ralstonia solanacearum* (Rso), **e-Proteobacteria:** *Helicobacter pylori* 26695 (Hpy), *Helicobacter pylori* J99 (jHp), *Campylobacter jejuni* (Cje); **g-Proteobacteria:** *Escherichia coli* K12 (Eco), *Escherichia coli* O157:H7EDL933 (EcZ), *Escherichia coli* O157:H7 (Ecs), *Yersinia pestis* (Ype), *Salmonella typhimurium* LT2 (Sty), *Buchnera* sp. APS (Buc), *Vibrio cholerae* (Vch), *Pseudomonas aeruginosa* (Pae), *Haemophilus influenzae* (Hin), *Pasteurella multocida* (Pmu), *Xylella fastidiosa* 9a5c (Xfa).

Figure 3. Correlation with Kimura distance and additivity for several selected distance measures in Problem 1. Distance abbreviations are as follows:  $dK$ , Kimura distance;  $dJC$ , distance based on Jaccard similarity index;  $d_{cor}$ , correlation-based distance;  $d_{-1}$ ,  $d_{-2}$ ,  $d_{-3}$ ,  $d_{-4}$ , and  $d_{-5}$  are  $A_I$ -distances with negative exponents ( $I = -1, -2, -3, -4, -5$ );  $d_0$ ,  $d_1$ ,  $d_2$  are  $A_I$ -distances with  $I = 0, 1, 2, 3$  respectively;  $d_{inf}$  and  $d_{-inf}$  are  $A_I$ -distances with  $I = \infty$  and  $I = -\infty$ .

Figure 4. Statistics of distribution of several distance measures in Problem 1. Distance abbreviations are as for Figure 3.

Figure 5. Best performing distance has the highest ratio of perfectly matched known protein complexes to the total number of predicted complexes (distance derived from Simpson similarity index with 0.5 distance cutoff for clustering).

Table 1. Descriptive statistics of distance distributions for Problem 1, 2 and 3. Best performing distances are shown in bold.

	Problem1					Problem2					Problem3				
	Mean	Var	Median	Kurtosis	Skewness	Mean	Var	Median	Kurtosis	Skewness	Mean	Var	Median	Kurtosis	Skewness
d_inf	0.201	0.008	0.187	-0.567	-0.057	<b>0.989</b>	<b>0.005</b>	<b>1.000</b>	<b>84.957</b>	<b>-8.487</b>	0.462	0.040	0.444	-0.119	0.151
d-10	0.245	0.007	0.239	-0.054	-0.327	0.989	0.004	1.000	85.266	-8.501	0.487	0.036	0.466	-0.130	0.183
d-9	0.249	0.007	0.243	0.044	-0.370	0.989	0.004	1.000	85.346	-8.505	0.488	0.036	0.469	-0.132	0.186
d-8	0.253	0.006	0.248	0.176	-0.427	0.989	0.004	1.000	85.460	-8.510	0.490	0.035	0.472	-0.134	0.190
d-7	0.259	0.006	0.256	0.359	-0.503	0.989	0.004	1.000	85.630	-8.517	0.492	0.035	0.477	-0.136	0.194
d-6	0.266	0.006	0.267	0.622	-0.609	0.990	0.004	1.000	85.892	-8.528	0.495	0.035	0.481	-0.139	0.199
d-5	0.274	0.006	0.279	1.007	-0.760	0.990	0.004	1.000	86.322	-8.546	0.498	0.034	0.487	-0.142	0.204
d-4	0.285	0.006	0.290	1.566	-0.976	0.990	0.004	1.000	87.078	-8.577	0.501	0.034	0.493	-0.144	0.209
d-3	0.299	0.006	0.314	2.276	-1.263	0.990	0.004	1.000	88.550	-8.635	0.504	0.033	0.500	-0.144	0.212
<b>d-2</b>	<b>0.317</b>	<b>0.006</b>	<b>0.338</b>	<b>2.720</b>	<b>-1.516</b>	0.991	0.003	1.000	91.826	-8.761	0.508	0.033	0.500	-0.143	0.212
d-1	0.340	0.008	0.361	2.072	-1.417	0.991	0.003	1.000	100.104	-9.073	0.513	0.032	0.500	-0.140	0.208
d0	0.365	0.012	0.377	0.898	-0.910	0.992	0.003	1.000	116.371	-9.707	0.517	0.032	0.500	-0.139	0.198
d1	0.387	0.016	0.386	0.280	-0.479	0.993	0.002	1.000	130.798	-10.280	0.521	0.032	0.500	-0.141	0.186
d2	0.404	0.020	0.400	0.015	-0.285	0.993	0.002	1.000	139.807	-10.625	0.525	0.032	0.524	-0.145	0.172
d3	0.416	0.022	0.416	-0.128	-0.221	0.993	0.002	1.000	145.599	-10.837	0.529	0.031	0.530	-0.150	0.160
d4	0.426	0.024	0.428	-0.214	-0.209	0.993	0.002	1.000	149.530	-10.975	0.532	0.031	0.532	-0.154	0.149
d5	0.432	0.025	0.440	-0.266	-0.216	0.994	0.002	1.000	152.311	-11.070	0.534	0.031	0.533	-0.158	0.139
d6	0.438	0.026	0.448	-0.298	-0.230	0.994	0.002	1.000	154.347	-11.138	0.537	0.031	0.533	-0.160	0.130
d7	0.442	0.026	0.452	-0.318	-0.244	0.994	0.002	1.000	155.878	-11.188	0.539	0.031	0.534	-0.162	0.123
d8	0.445	0.027	0.456	-0.330	-0.258	0.994	0.002	1.000	157.058	-11.226	0.541	0.031	0.535	-0.163	0.116
d9	0.448	0.027	0.461	-0.336	-0.271	0.994	0.002	1.000	157.985	-11.256	0.542	0.031	0.536	-0.164	0.111
d10	0.451	0.027	0.464	-0.339	-0.282	0.994	0.002	1.000	158.728	-11.279	0.543	0.031	0.536	-0.164	0.106
dinf	0.478	0.028	0.500	-0.229	-0.402	0.994	0.002	1.000	164.454	-11.444	0.563	0.029	0.556	-0.146	0.057
dJac	0.545	0.021	0.557	1.753	-1.121	0.996	0.001	1.000	231.798	-13.472	<b>0.667</b>	<b>0.025</b>	<b>0.667</b>	<b>0.255</b>	<b>-0.475</b>
<b>dcor</b>	<b>0.531</b>	<b>0.016</b>	<b>0.552</b>	<b>3.457</b>	<b>-1.723</b>	0.997	0.002	1.004	117.194	-9.735	0.998	0.102	1.015	-0.219	-0.031

NOTE: see Figure 3 legend for distance abbreviations.

Table 2. The values of scores (sum of ratios: size of in-cluster benchmark to the cluster size over all clusters) for different distance measures computed in Problem 3.

	d_inf	d_2	d_1	d0	d1	d2	dinf	dcor	dJC
score	0.561	0.576	0.580	0.586	0.587	0.575	0.546	0.622	0.870

NOTE: see Figure 3 legend for distance abbreviations.

## REFERENCES

- Arkin, E., *et al.* (1991) An efficiently computable metric for comparing polygonal shapes. In *IEEE Trans. Patt. Anal. Mach. Intell.*, pp. 209-216.
- Atchley, W.R., *et al.* (2005) Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A* **102**: 6395-6400.
- Bader, G.D. and C.W. Hogue. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol* **20**: 991-997.
- Bader, G.D. and C.W. Hogue. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**: 2.
- Brazma, A. and J. Vilo. (2000) Gene expression data analysis. *FEBS Lett* **480**: 17-24.
- Brun, C., *et al.* (2004) Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics* **5**: 95.
- Buneman, P. (1974) A note on the metric properties of trees. *Journal of Combinatorial Theory (B)* **17**: 48-50.
- Cho, R.J., *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**: 65-73.
- Date, S.V. and E.M. Marcotte. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **21**: 055-1062.
- de Lichtenberg, *et al.* (2005) Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* **21**: 1164-1171.
- Dutilh, B.E., *et al.* (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* **58**: 527-539.
- Eisen, M.B., *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863-14868.
- Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141-147.
- Gibbons, F.D. and F.P. Roth. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res* **12**: 1574-1578.
- Glazko, G.V. and A.R. Mushegian. (2004) Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol* **5**: R32.
- Gusfield, D. (1999) *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press.
- Holland, B.R., *et al.* (2002) Delta plots: a tool for analyzing phylogenetic distance data. *Mol Biol Evol* **19**: 2051-2059.
- Kaufman, L. and P.J. Rousseeuw. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.
- Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111-120.
- King, A.D., *et al.* (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*.
- Korbel, J.O., *et al.* (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* **18**: 159-162.

- Kosman, E. and K.J. Leonard. (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Mol Ecol* **14**: 415-424.
- Krause, R., *et al.* (2003) A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens. *Bioinformatics* **19**: 1901-1908.
- Lerat, E., *et al.* (2003) From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the gamma-Proteobacteria. *PLoS Biol.* **1**: E19.
- Lesage, G., *et al.* (2005) An interactional network of genes involved in chitin synthesis in *Saccharomyces cerevisiae*. *BMC Genet* **6**: 8.
- Lipkus, A.H. (1999) A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry* **26**: 263-265.
- Mewes, H.W., *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**: 31-34.
- Mirkin, B. and E.V. Koonin. (2003) A top-down method for building genome classification trees with linear binary hierarchies. In *Bioconsensus*. (ed. J.-F.L. M. Janowitz, F. McMorris, B. Mirkin, and F. Roberts.), pp. 97-112. American Mathematical Society, Providence.
- Pellegrini, M., *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci USA* **96**: 4285-4288.
- Pfahringer, B. (1995) Compression-based discretization of continuous attributes. In *Proceedings of the 12th International Conference on Machine Learning.*, pp. 456-463.
- Pielou, E.C. (1975) *Ecological diversity*. Wiley, New York.
- Posada, D. and K. Crandall. (2001) Selecting the Best-Fit Model of Nucleotide Substitution. *Syst. Biol.* **50**: 580-601.
- Rencher, A.C. (2002) *Methods of multivariate analysis*. A John Wiley & Sons, Inc., New York.
- Robinson, D.R. and L.R. Foulds. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences* **53**: 131-147.
- Rustici, G., *et al.* (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat Genet* **36**: 809-817.
- Saitou, N. and M. Nei. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-425.
- Shmulevich, I. and W. Zhang. (2002) Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* **18**: 555-565.
- Sneath, P. and R. Sokal. (1973) *Numerical taxonomy. The principles and practice of numerical classification.*, San Francisco.
- Strong, M., *et al.* (2003) Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. *Genome Biol.* **4**: R59.
- Tatusov, R.L., *et al.* (1997) A genomic perspective on protein families. *Science* **278**: 631-637.
- Uetz, P., *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623-627.
- von Mering, C., *et al.* (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl. Acad. Sci USA* **100**: 15428-15433.

- Waterman, M.S., *et al.* (1977) Additive evolutionary trees. *J Theor Biol* **64**: 199-213.
- Wilson, D. and T. Martinez. (1997) Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research* **6**: 1-34.
- Wolf, Y.I., *et al.* (2002) Genome trees and the tree of life. *Trends Genet.* **18**: 472-479.
- Wolf, Y.I., *et al.* (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**: 8.
- Wong, S.L., *et al.* (2004) Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A* **101**: 15682-15687.
- Xu, W. and D.P. Miranker. (2004) A metric model of amino acid substitution. *Bioinformatics* **20**: 1214-1221.
- Zaretsky, K. (1965) Reconstruction of a tree from the distances between its pendant vertices. *Uspekhi Math. Nauk (Russian Mathematical Survey)* **20**: 90-92.

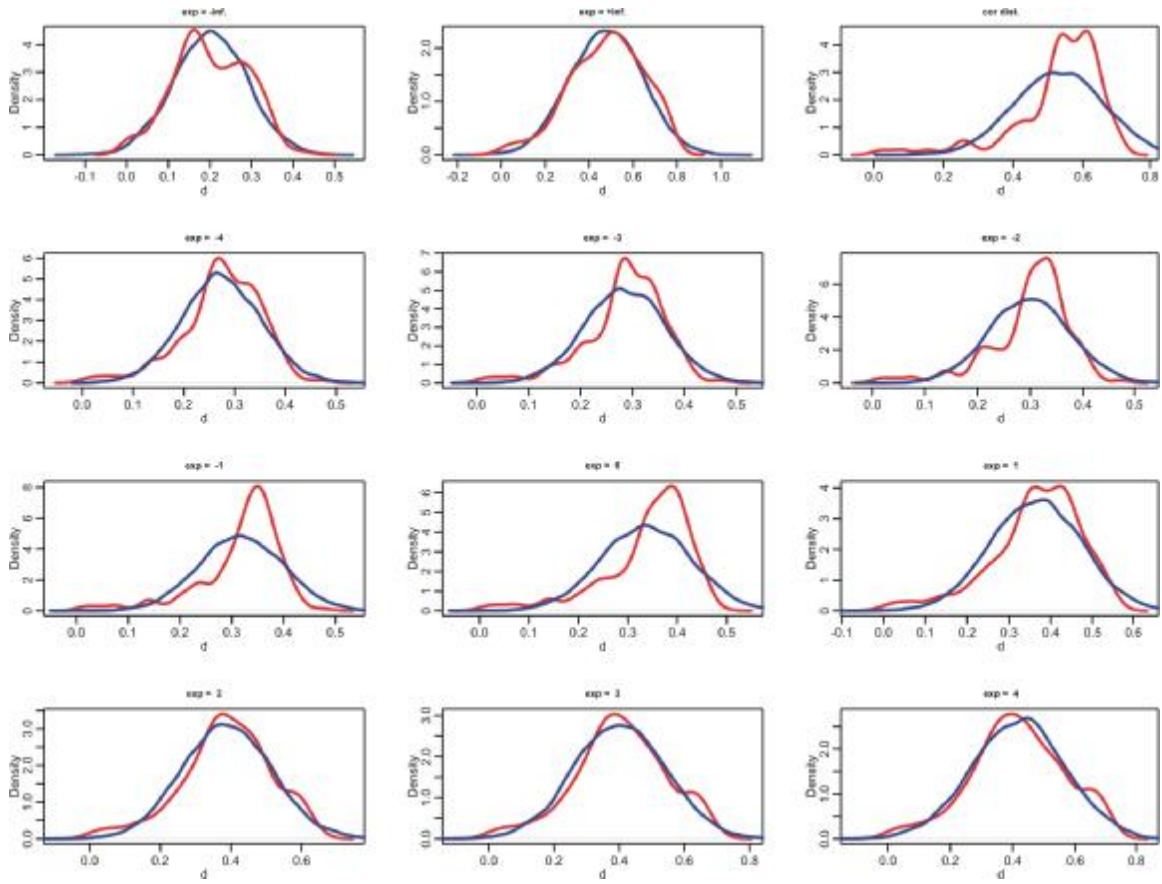


Figure 1

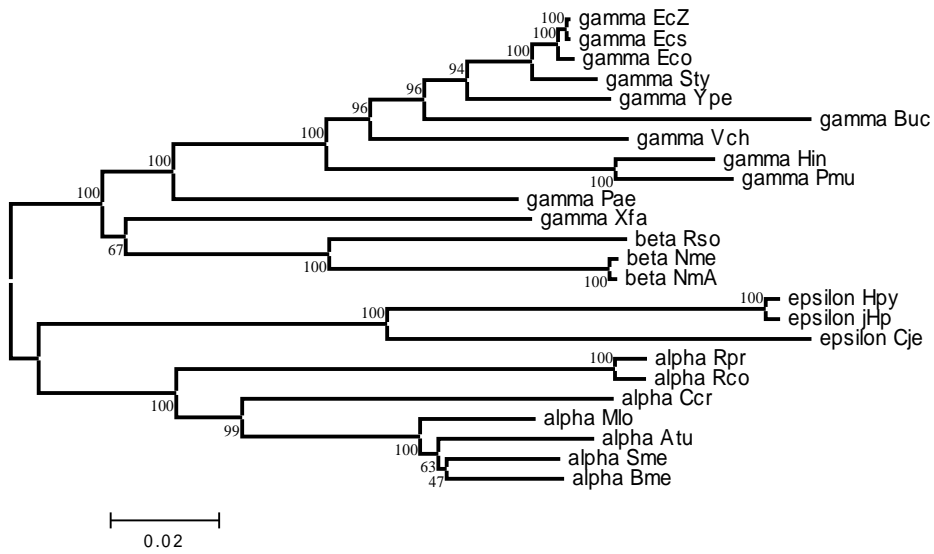


Figure 2

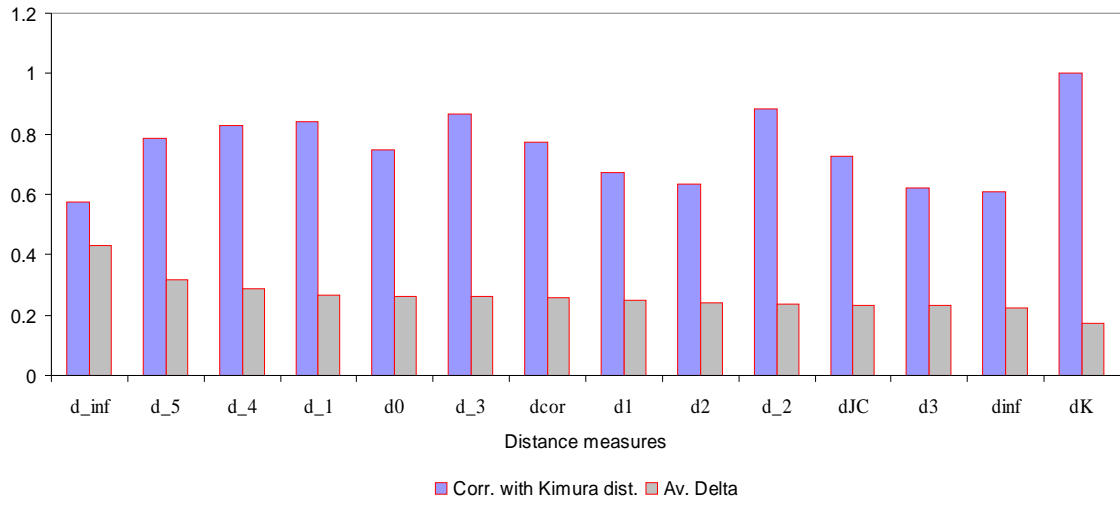


Figure 3

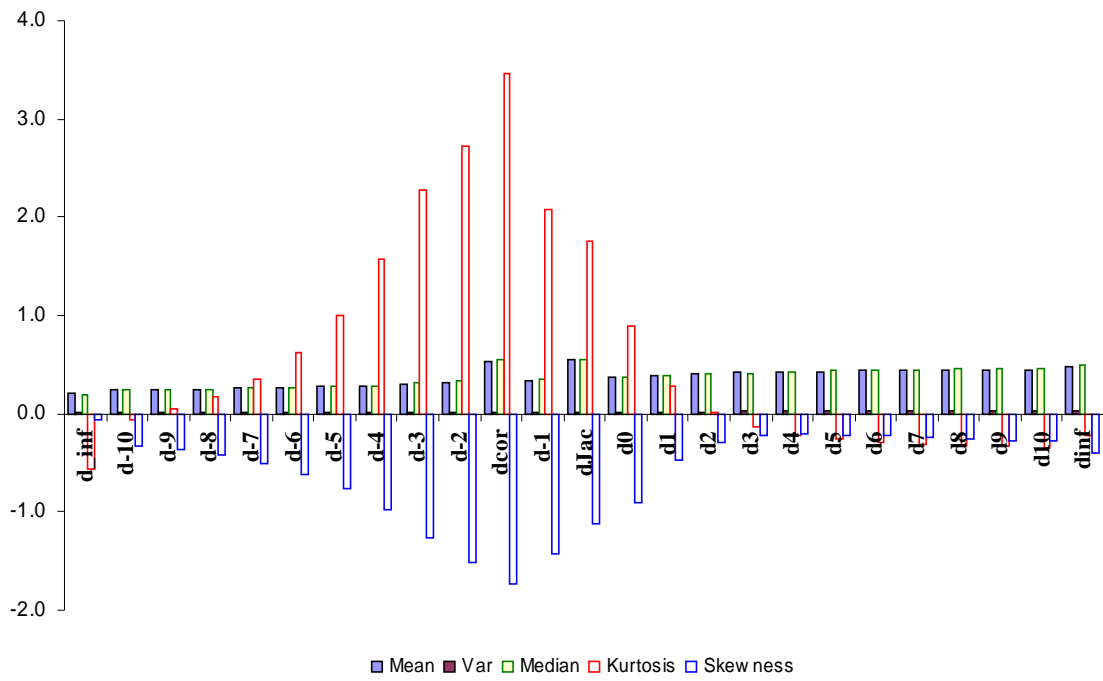


Figure 4

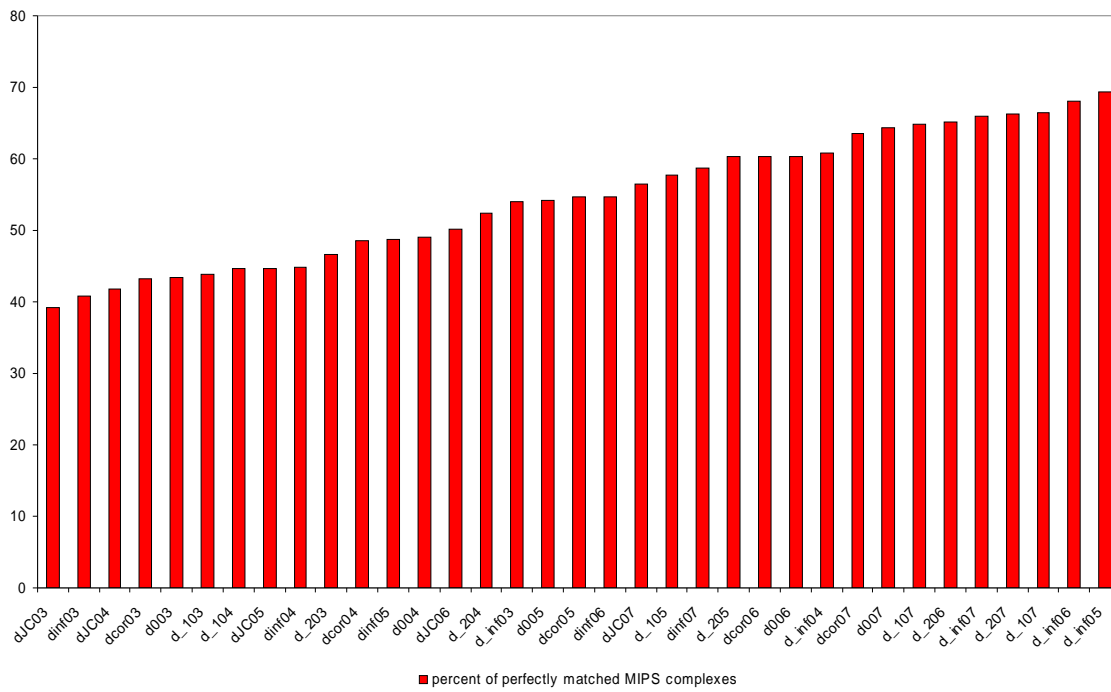


Figure 5