Significantly distinct branches of hierarchical trees: A framework for statistical analysis and applications to biological data

Guoli Sun[1,2] and Alexander Krasnitz[1]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

[2]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, USA.

Abstract

We formulate a method termed Tree Branches Evaluated Statistically for Tightness (TBEST) for identifying significantly distinct tree branches in hierarchical clusters. For each branch of the tree a measure of tightness is defined as a rational function of heights, both of the branch and of its parent. A statistical procedure is then developed to determine the significance of the observed values of tightness. We test TBEST as a tool for tree-based data partitioning by applying it to four benchmark datasets, each from a different area of biology and each with a well-defined partition of the data into classes. In all cases TBEST performs on par with or better than the existing techniques.

1 INTRODUCTION

Hierarchical clustering (HC) is widely used as a method of partitioning data and of identifying meaningful data subsets. Most commonly an application consists of visual examination of the dendrogram and intuitive identification of sub-trees that appear clearly distinct from the rest of the tree. Obviously, results of such qualitative analysis and conclusions from it may be observer-dependent. Quantifying the interpretation of hierarchical trees and introducing mathematically and statistically well-defined criteria for distinctness of sub-trees would therefore be highly beneficial and is the focus of this work.

The need for such quantification was recognized some time ago, and methods have been designed for (a) identifying distinct data subsets while (b) making use of hierarchical tree organization of the data. These methods fall into two categories, depending on whether or not they employ statistical analysis. The simplest approach that does not rely on statistical analysis is a static tree cut, wherein the tree is cut into branches at a given height. This procedure is guaranteed to produce a partition of the data, but provides no way to choose the height at which to cut. Dynamic Tree Cut, or DTC in the following (Langfelder, Zhang et al. 2008), is a more sophisticated recipe wherein the tree is generally partitioned into branches of unequal heights,

but here again the partition depends on a parameter (the minimal number of leaves in a branch) that cannot be determined by the method.

In addition, there are methods for choosing a tree partition from considerations of branch distinctness and its statistical significance. Sigclust, or SC in the following (Liu, Hayes et al. 2008), is a parametric approach wherein a two-way split of the data is deemed significant if the null hypothesis that the data are drawn from a single multivariate normal distribution is rejected. The method is designed to work in the asymptotic regime, where the dimensionality of the objects being clustered far exceeds the number of the objects. In application to trees SC works in a top-down fashion, by first examining the split at the root node and proceeding from a parent node to its daughter nodes only if the split at the parent node has been found significant. Unlike SC, the sum of the branch lengths method, or SLB in the following (Munneke, Schlauch et al. 2005) is designed specifically for hierarchical trees and utilizes a measure of distinction between two nodes joined at a parent node that is linearly related to the heights of the two daughter nodes and that of the parent. Similarly to SC, SLB adopts a top-down scheme.

A method introduced here is termed Tree Branches Evaluated Statistically for Tightness (TBEST) and shares features with the existing approaches. Like SC and SLB, TBEST employs statistical analysis to identify significantly distinct branches of a hierarchical tree. Similarly to DTC and SLB, it uses tree node heights to assess the distinctness of a tree branch. At the same time, TBEST differs from the existing designs in several aspects, two of which are critical. First, unlike both SC and SLB, it examines all the tree nodes simultaneously for distinctness. Secondly, unlike SLB, it combines node heights non-linearly to construct a statistic for distinctness that is better able to handle a tree in which distinct branches of approximately equal numbers of leaves occur at different heights. The key properties of all four methods are summarized in Table 1.

In the remainder of this work we formulate TBEST and systematically compare its performance to that of DTC, SC and SLB on a number of benchmark datasets originating from a variety of biological sources. In all cases we find that TBEST performs as well as or better than the three published methods. We conclude by discussing generalizations of TBEST and its relation to other aspects of cluster analysis.

Table 1. Properties of TBEST, and other three published methods.

Method	Order of examining the tree	Significance estimated
TBEST	All internal nodes in parallel	Yes
DTC	top down and bottom up	No
SC	top down	Yes
SLB	top down	Yes

2 METHODS

Consider a set of objects with pair-wise relations given by a dissimilarity matrix. Given a linkage rule, a hierarchical tree can be grown for the set. This tree is specified, in addition to its branching structure, by the heights of its nodes. The height of the node quantifies the dissimilarity within the data subset defined by the node. In the special case of the objects being points in a Euclidean space, and the dissimilarities defined as distances between the points, the node height is a measure of the linear extent of the subset. Accordingly, the difference in heights between a parent $P(n)$ of node n and that of n itself quantifies how distinct the data represented by n are from those represented by the other child of $P(n)$. We wish to construct, for each node of the tree, a measure of how distinct the data subset corresponding to the node is from the data set. An example of a one-dimensional dataset, tabulated in Supplementary Data and shown in Figure 1, clarifies considerations involved in constructing such a measure. Both the subsets shown in blue and in green are clearly distinct from the rest of the data, but the blue node is not as different in height from its parent as the green node is from its parent. Based on the parent to child difference in heights, one would conclude, counter-intuitively, that the blue subset is not nearly as distinct as the green subset.

A measure in better agreement with intuition is the relative difference of heights:

$$S(n) \equiv \frac{h(P(n)) - h(n)}{h(P(n))} \qquad (1)$$

where $h(n)$ is the height of node n. In the following we refer to $S(n)$ as the tightness of node n. In the absence of inversions the tightness of any node is a number between 0 and 1. In particular, $S(n) = 1$ identically if n is a leaf. The two subsets highlighted in Figure 1 are nearly equally tight by this measure, despite the disparity in their heights.

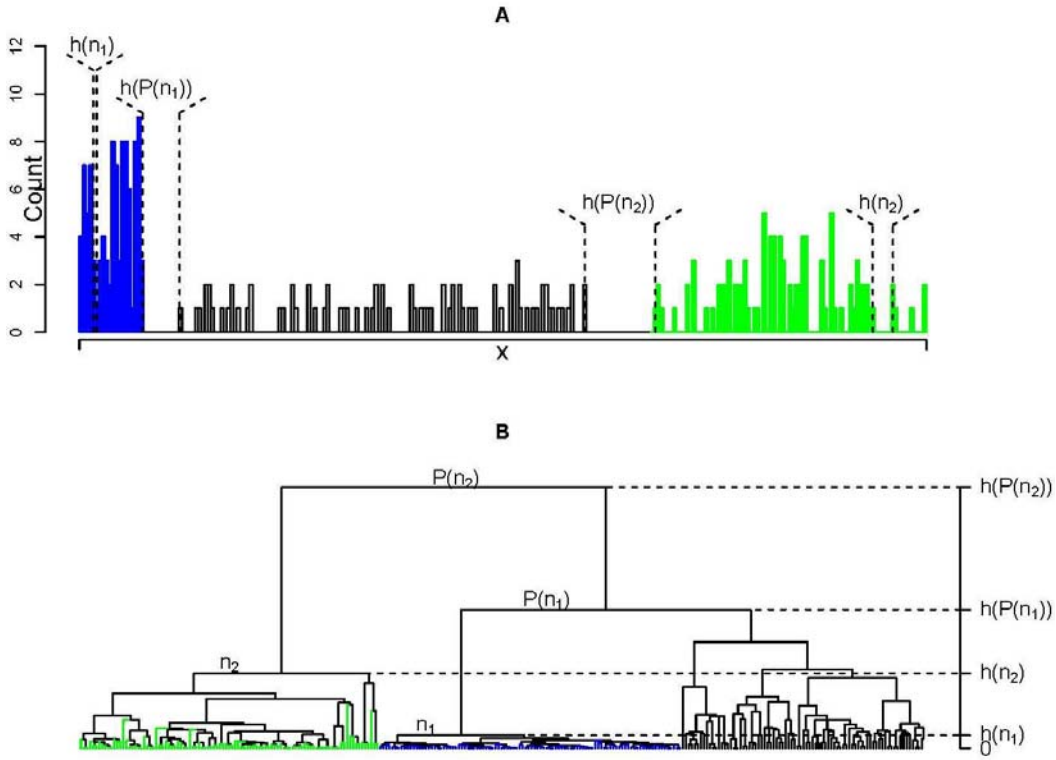

Fig. 1. Tightness analysis for a data set of 280 positive real values sampled from a mixture of three normal components: $N(0.5, 0.4^2)$; (blue), $N(11, 1^2)$; (green) and $N(5, 2^2)$. (black). **A)** A histogram of the data. The inset is an enlarged portion of the histogram enclosed in the red frame **B)** A hierarchical tree of the data, grown using the absolute difference of the data values as the dissimilarity measure, and single linkage. Thus, the node heights shown in (**B**) are equal to the corresponding gaps in the data, as indicated in (**A**).

Next, we consider statistical analysis of tightness. To this end we view $S(n)$ as a statistic and seek to define a suitable null distribution to which the observed $S(n)$ is to be compared. This null distribution of tightness is obtained by randomizing the dataset from which trees are grown. We do not specify at this point how such randomization is to be performed, and different data types may require different randomization prescriptions. Instead, we design a general procedure for constructing the null distribution of tightness for any given data randomization scheme. To guide this design, we generated distributions of tightness in trees grown from randomized data for multiple combinations of datasets, definitions of dissimilarity, linkage rules and randomization methods, as listed in Table 2.

Table 2. Combinations of datasets, dissimilarity, linkage and randomization methods, used for testing TBEST.

Dataset	Dissimilarity	Linkage	Data permutation Method
Leukemia	Euclidean	Ward	Independently for each gene (column)
	(1 - Pearson correlation)	Average	
T10	Euclidean	Ward	Independently for each chromosome; identically for all cores (columns) in a chromosome
	(1 - Pearson correlation)	Average	
Organelles	(1 - Pearson correlation)	Ward	Independently for each protein (column)
	(1 - Pearson correlation)	Average	
Chondrosarcoma	(1 - Spearman correlation)	Ward	Independently for each surface marker (column)
	(1 - Kendall correlation)	Average	
	Manhattan	Ward	

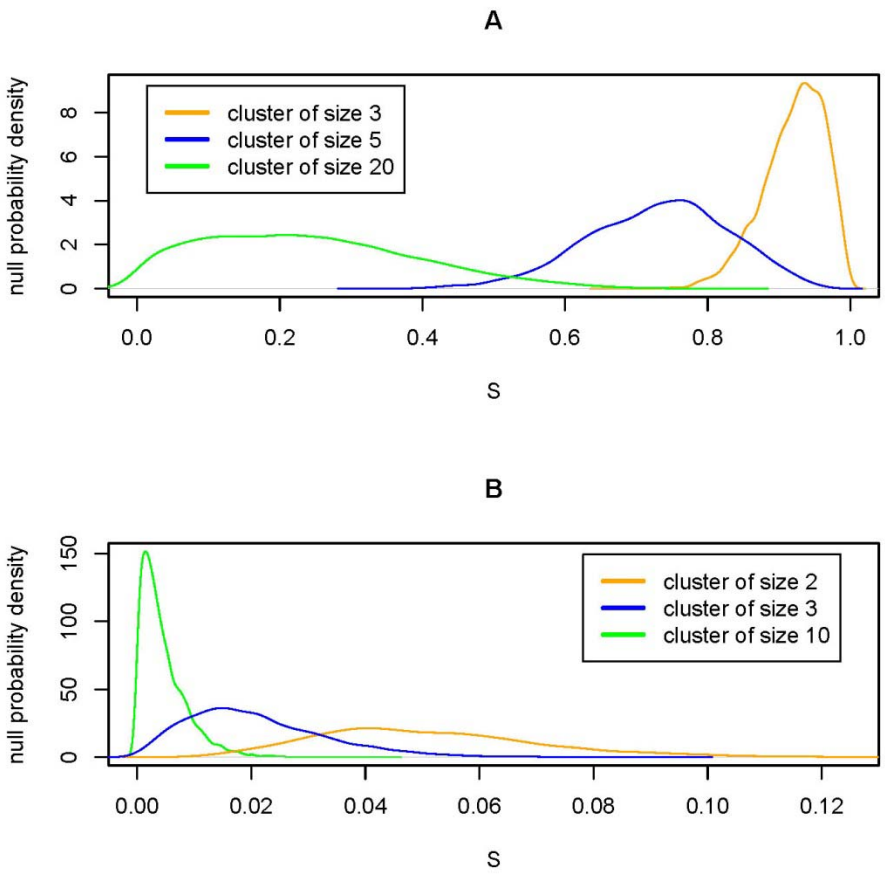

Fig. 2. The null distribution of node tightness S depends on the number of leaves. The empirical probability density distributions for the one-dimensional example (**A**, *cf* Fig. 1.) and for the Leukemia set (**B**) are shown, for three different values of the number of leaves in each case. Each plot is based on 5000 randomizations of the respective data set.

As Figure 2 illustrates, it is a generic property of such distributions to be skewed towards larger values of tightness for nodes with smaller numbers of leaves. The identity $S(n) = 1$ for single-leaf nodes is consistent with this observation. We therefore conclude that, for a given observed value of tightness, the appropriate null distribution is sampled by repeated randomization of the data, growing a tree for each randomization, selecting among its nodes the ones with the numbers of leaves matching the observation, and determining the tightness of these nodes. However, it is not guaranteed that, in any tree grown from randomized data, there will be a unique node with a number of leaves exactly equal to that of the observed node. To resolve this difficulty conservatively, we adopt the following procedure. If, for a given data randomization, the tree contains nodes with the number of leaves exactly as observed, the highest $S(n)$ computed for these nodes is added to the sample. Otherwise we consider all the nodes with the number of leaves nearest the observed one from above and all those with the number of leaves nearest the observed one from below and add to the sample the highest $S(n)$ of any of these nodes.

With the sampling procedure specified, tests for statistical significance of tightness can be conducted for all the internal nodes of the observed tree, excluding the root, since the latter has no parent. The number of tests is therefore two less than the number of leaves. Due to this multiplicity of tests, higher levels of significance are required for rejection of the null hypotheses for trees with larger numbers of leaves. A straightforward way to handle this requirement would be to increase the size of the sample from the null distribution by performing more randomizations. However, for trees with large numbers of leaves this simple-minded approach may be rendered impractical by computational cost. Instead, higher levels of significance may be accessed by using the extreme-value theory (EVT) to approximate the tail of the null distribution, thereby permitting considerable economy of computational effort (Knijnenburg, Wessels et al. 2009). We use the EVT-based method alongside the traditional approach in our benchmark studies in the following.

3 RESULTS

We evaluated the performance of TBEST in comparison to three published methods of identifying distinct subsets of observations, namely, DTC, SC and SLB. The four datasets used in the evaluation share two common features: they originate in biological experiments and in

each case there is an independently known, biologically meaningful partition of observations into types. We call this known partition "truth", and the corresponding types the true types, henceforth. The essential properties of the benchmark datasets are summarized in Table 3.

To better judge the performance of TBEST in comparison to the other three algorithms, we considered, for each dataset, more than one combination of dissimilarity and linkage methods used for hierarchical clustering. With the exception of the 2nd benchmark case, randomization of the input data, as required for both TBEST and SLB, consisted of randomly permuting the observed values, independently for each variable. The degree of agreement between a computed partition of the data and the truth is quantified in terms of corrected-for-chance Rand index, or cRI in the following (Hubert and Arabie 1985). It should be noted that the subsets of the data identified as distinct by TBEST and the other three techniques by necessity correspond each to a branch of a tree. This, however, is not necessarily the case for the true types, some of which do not correspond to a single branch. As a result, a perfect match between any computed partition and the truth may not be possible, and the maximal attainable value of cRI may be below 1. For this reason, to evaluate the performance of TBEST and the published methods across benchmark datasets, we also identify, for each tree considered, a partition into branches that best matches the truth and determine cRI between that partition and the computed partitions for each of the methods.

Table 3. Properties of the four benchmark datasets.

Dataset	Origin	Number of items	Number of variables	True number of classes
Leukemia	mRNA levels from microarray analysis	38	999	3
T10	DNA copy number analysis, sequencing	100	354	4
Organelles	Proteomic analysis, using mass spectrometry	24	4768	4
Chondrosarcoma	Flow cytometry analysis of surface markers from fluorescence intensity	32	11	4

3.1 Leukemia

The original Leukemia dataset (Golub, Slonim et al. 1999) contained mRNA level values for approximately 3000 genes; this number was reduced to 999 by feature selection (Monti, Tamayo et al. 2003). The truth is a partition of patient cases into those of acute myeloid leukemia (AML,

11 cases) and of acute lymphoblastic leukemia (ALL), and a further partition of the ALL subset into the B-cell lineage (19 cases) and the T-cell lineage (8 cases) types.

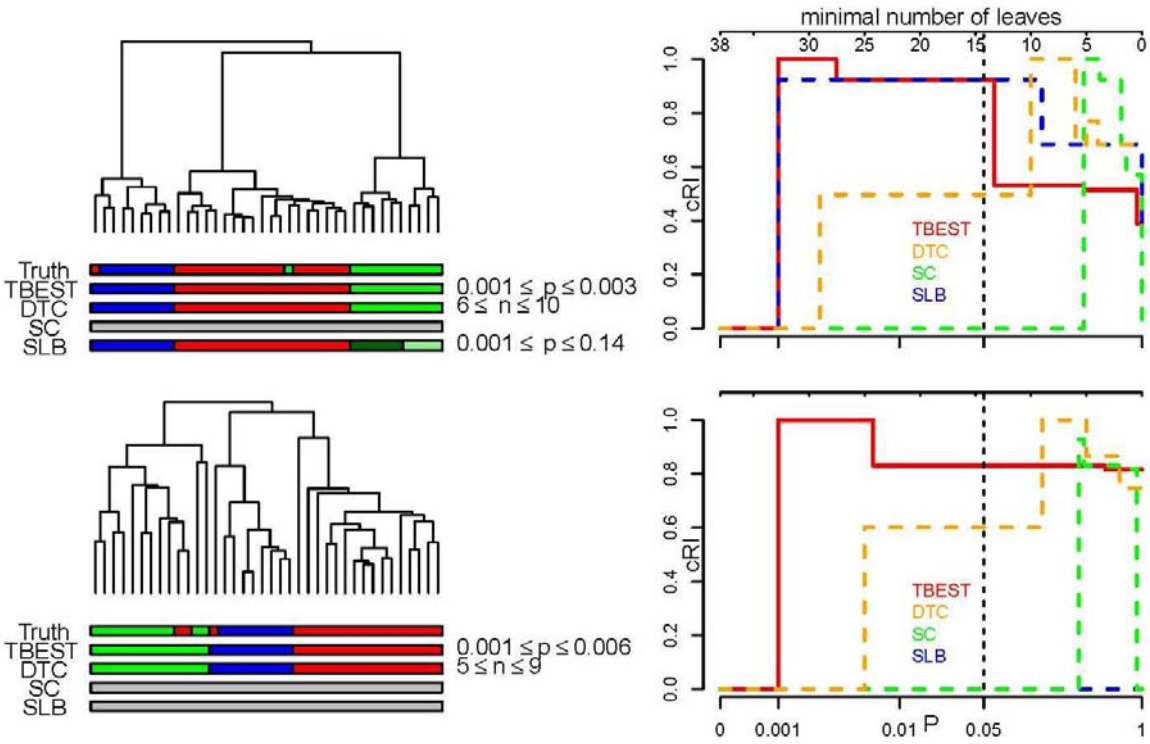

Fig. 3. Performance comparison of TBEST and the three published methods in Leukemia dataset for the Euclidean dissimilarity – Ward linkage combination (top) and for the (1 - Pearson correlation) dissimilarity – average linkage (bottom). In each case, the left portion shows the corresponding dendrogram, under which then true partition and the partition best matching the truth for each of the methods are shown as color bars. In the right portion, the relative corrected Rand index of the computed partition is plotted against the required level of significance p for each of the significance-based methods and against the minimal allowed number of leaves for DTC.

The comparison between the four algorithms is displayed graphically in Figure 3. For the Ward linkage, two of the significance-based methods, SC and TBEST, attain the highest possible value of the cRI. However, SC only does so with low significance ($p > 0.33$), while TBEST achieves it best performance with high significance ($p \approx 2\times10^{-3}$) and maintains performance close to optimal in a wide range of p-values. The performance of SLB in this case is similar to that of TBEST, but SLB does not attain the optimum. With the average linkage, TBEST outperforms both SC and SLB throughout the entire range of p-values considered and attains optimal performance at high significance. In both cases the performance of DTC is highly sensitive to the minimal allowed size of a branch, especially so for the Ward linkage, where this algorithm

attains top performance for sizes between 6 and 10, but performs substantially below the optimum outside this range.

3.2 T10

The second benchmark dataset originates from DNA copy number analysis of 100 individual cells harvested from a breast tumor (Navin, Kendall et al. 2011). The true partition in this case is four-way, with the subsets differing from each other by ploidy as determined by cell sorting. The rows of the data matrix correspond each to a cell, the columns correspond each to a pre-defined genomic region of recurrent copy number variation called a core, specified by the sign of variation (gain or loss) and the endpoint positions of the region. The entries in the matrix quantify the extent to which copy number alterations observed in the cells match the cores (Krasnitz, Sun et al. 2013).

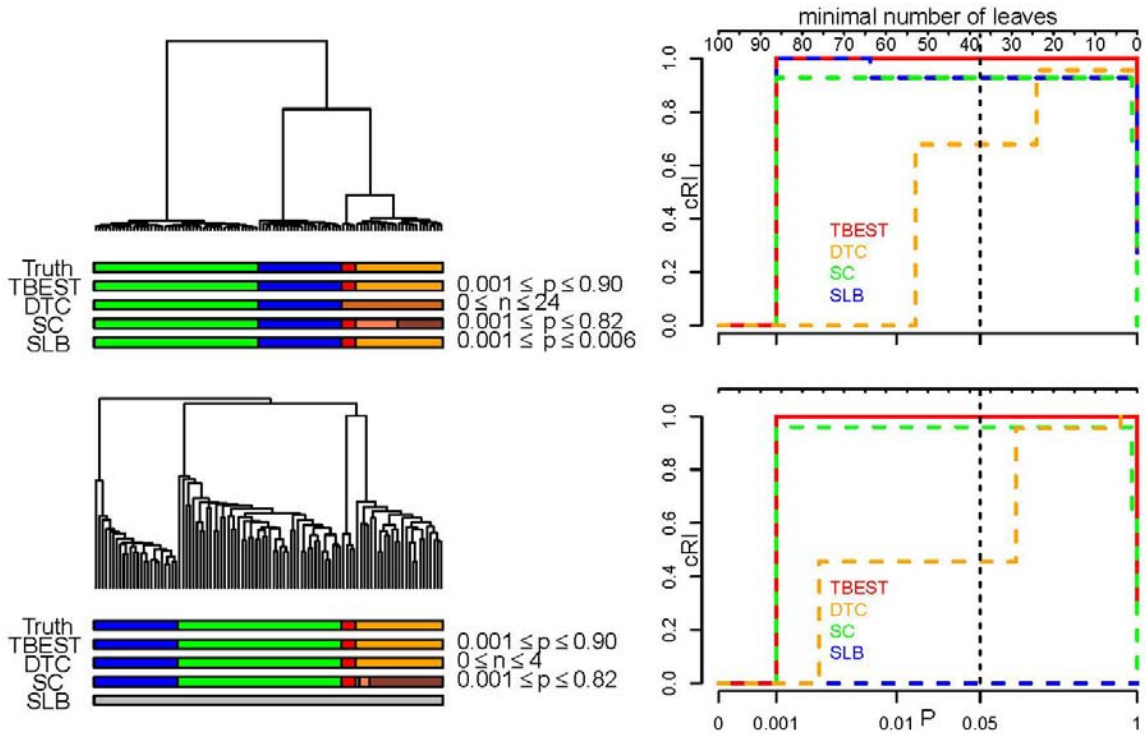

Fig. 4. Performance comparison of TBEST and the three published methods in T10 dataset for the Euclidean dissimilarity – Ward linkage combination (top) and for the (1 - Pearson correlation) dissimilarity – average linkage (bottom). In each case, the left portion shows the corresponding dendrogram, under which then true partition and the partition best matching the truth for each of the methods are shown as color bars. In the right portion, the relative corrected Rand index of the computed partition is plotted against the required level of significance p for each of the significance-based methods and against the minimal allowed number of leaves for DTC.

There are multiple instances of strong geometric overlap between cores. As a result, the corresponding columns in the data matrix exhibit strong pairwise correlations, positive for cores of equal sign (both gains or both losses), and negative for cores of opposite signs. Consistent with these geometric constraints, the null distribution in this case is generated as follows: the data matrix is divided into sub-matrices by the chromosome number (1,2,...,22,X), and rows are permuted independently within each sub-matrix.

The results are illustrated in Figure 4. For the Ward linkage only TBEST and SLB identify the true partition, with TBEST succeeding in a broader range of p-values. For the average linkage TBEST outperforms the other two significance-based algorithms and matches the truth perfectly in a broad range of p-values, while DTC matches the truth if the minimal allowed number of leaves is 4 or less.

3.3 Organelles

Next, we consider a dataset derived from proteomic analysis of the content of four cellular compartments in a number of mouse tissues. Data were log-transformed and normalized by proteins before clustering (Kislinger, Cox et al. 2006). The true partition of the data is by the cellular compartment, and the two hierarchical clustering methods considered here both have the branch structure organized by the compartment label, to a good approximation. Of the three significance-based methods compared, only TBEST reproduces the truth to the maximal extent possible for both combinations of dissimilarity and linkage, and it does so stably in the broadest range of the levels of significance (Figure 5).

DTC achieves top performance for the (1 - Pearson correlation) dissimilarity – Ward linkage combination if its minimal allowed number of leaves does not exceed that of the smallest compartment-associated branch of the tree. However, this property is lost for the (1 – Pearson correlation) dissimilarity – average combination where a cluster with two leaves is identified by DTC if the minimal number of leaves is set at or below 2.

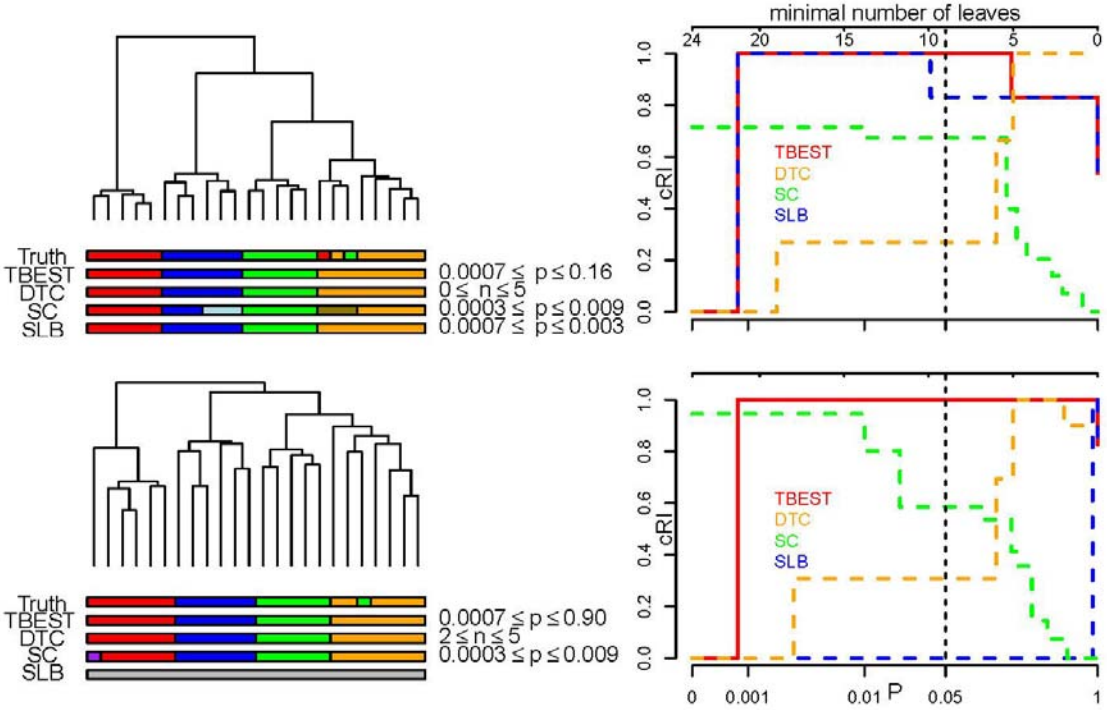

Fig. 5. Performance comparison of TBEST and the three published methods in Organelles dataset for the (1 - Pearson correlation) dissimilarity – Ward linkage combination (top) and for the (1 - Pearson correlation) dissimilarity – average linkage (bottom). In each case, the left portion shows the corresponding dendrogram, under which then true partition and the partition best matching the truth for each of the methods are shown as color bars. In the right portion, the relative corrected Rand index of the computed partition is plotted against the required level of significance p for each of the significance-based methods and against the minimal allowed number of leaves for DTC.

3.4 Chondrosarcoma

Finally, we discuss the performance of the four methods on a dataset generated by flow cytometry analysis of cells harvested from human tissues and cell lines. Among 34 samples, two samples were identified as multivariate outliers and removed before clustering (Diaz-Romero, Romeo et al. 2010). The truth is a four-way partition, with three parts corresponding each to a different tissue of origin and the fourth part formed by cells from tumor cell lines.

We have identified three combinations of dissimilarity and linkage for which the tree structure is fully consistent with the true partition and performed comparative analysis for all three. For two of these combinations ((1 - Spearman correlation) dissimilarity – Ward linkage and (1 - Kendall correlation) dissimilarity – average linkage) partition by TBEST matches the truth in a range of acceptable levels of significance. SLB only does so for the first combination, while SC fails to match the truth. Note the data dimension in this case is 11, and it is smaller than 32, the

number of observations. This dataset is therefore outside the range of applicability of SC. Manhattan dissimilarity – Ward linkage is the only combination for which TBEST fails to match the truth at an acceptable level of significance. DTC performs well for the first and third combinations, but only matches the truth in a restricted range of numbers of leaves in the second case.

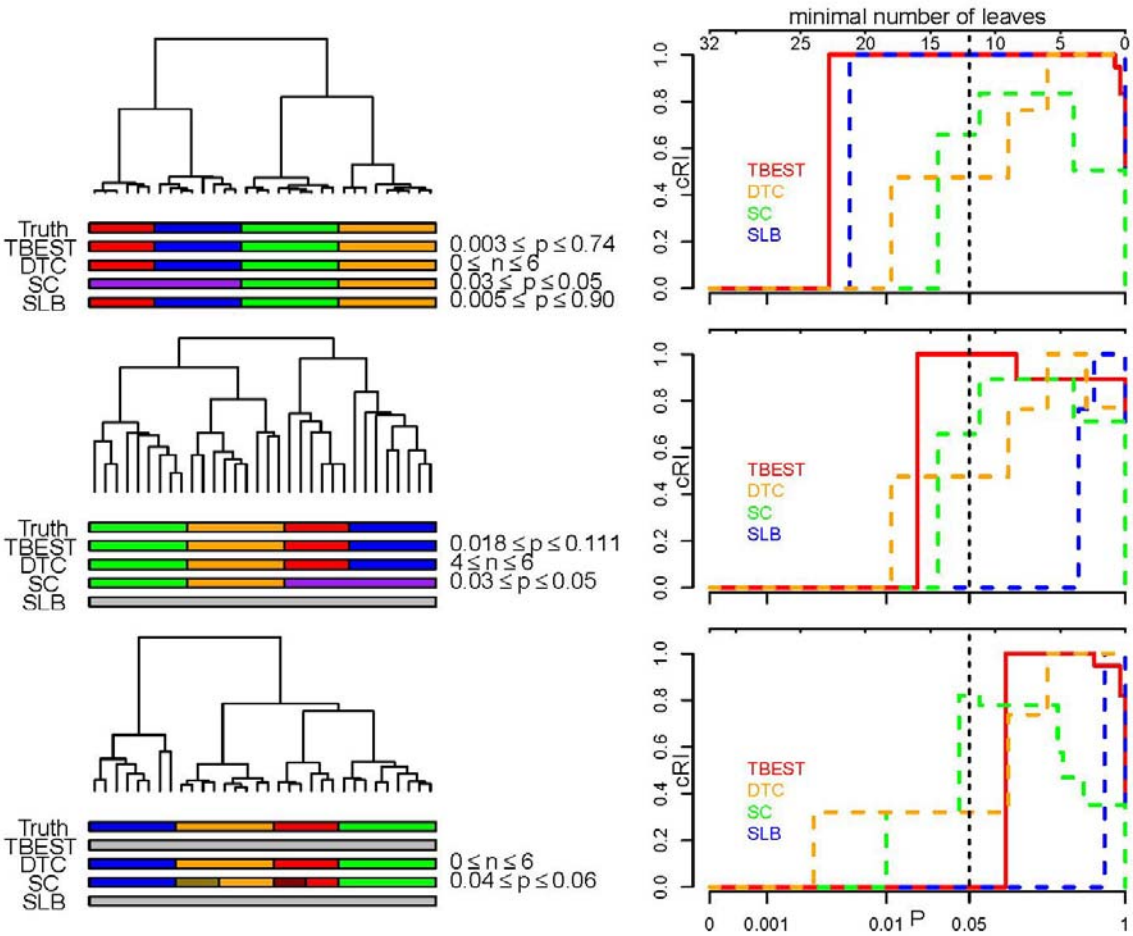

Fig. 6. Performance comparison of TBEST and the three published methods in Chondrosarcoma dataset for the (1 - Spearman correlation) dissimilarity – Ward linkage combination (top), (1 - Kendall correlation) dissimilarity – average linkage combination (middle), and Manhattan dissimilarity – average linkage (bottom). In each case, the left portion shows the corresponding dendrogram, under which then true partition and the partition best matching the truth for each of the methods are shown as color bars. In the right portion, the relative corrected Rand index of the computed partition is plotted against the required level of significance p for each of the significance-based methods and against the minimal allowed number of leaves for DTC.

4 DISCUSSION

As our test results demonstrate, the performance of TBEST as a tool for data partitioning is equal or superior to that of similar published methods in a variety of biology-related settings. This is

true in particular for datasets with underlying tree-like organization, such sets of genomic profiles of individual cancer cells, of the same type as our second benchmark case above. In a work presently in progress we are applying TBEST systematically to a number of datasets of a similar nature.

TBEST can both be applied and formulated more broadly. The applicability of TBEST is not limited to data partitioning that has been our focus here. TBEST can be used for finding all significantly distinct branches of a hierarchical tree, regardless of whether these form a full partition. Further, alternatives to the test statistic of Equation 1 can easily be devised, For example, for any non-leaf node *n* we can introduce

$$\Sigma(n) \equiv \frac{h(n)-\frac{1}{2}[h(c_1(n))+h(c_2(n))]}{h(n)} \qquad (2)$$

where $c_1(n), c_2(n)$ are the two children of n. While the discussion of these extensions is beyond the scope of this work, an implementation of TBEST as an R language package (Sun and Krasnitz 2013) provides a number of options, both for the definition of tightness and for annotation of significantly distinct branches.

Finally, we note that tightness of tree branches is complementary to another important notion in clustering, namely, cluster stability under re-sampling of the input data. The latter notion can be analyzed in a number of ways, such as bootstrap analysis of trees (Felsenstein 1985, Efron, Halloran et al. 1996, Shimodaira 2002) or methods not directly related to trees (Dudoit and Fridlyand 2002, Monti, Tamayo et al. 2003). Existing work provides examples where both distinctness and stability under resampling are prerequisites of a meaningful partition (Cancer Genome Atlas Research Network 2011). Incorporation of TBEST into such combined analysis will be addressed in the future.

ACKNOWLEDGEMENTS

We are grateful to M. Wigler for contributing to the early stages of this work and numerous subsequent discussions; to S. Yoon for reading and commenting on the manuscript; to M. Akerman, B. Meunier and J.F. Hoquette for generously sharing their data with us; to K.A Schlauch for generously providing software.

Funding: This work was supported by the National Institutes of Health grant NIH/1UO1CA168409-01 and by grant 125217 from the Simons Foundation.

REFERENCES

Cancer Genome Atlas Research Network (2011). "Integrated genomic analyses of ovarian carcinoma." Nature **474**(7353): 609-615.

Diaz-Romero, J., S. Romeo, J. V. Bovee, P. C. Hogendoorn, P. F. Heini and P. Mainil-Varlet (2010). "Hierarchical clustering of flow cytometry data for the study of conventional central chondrosarcoma." J Cell Physiol **225**(2): 601-611.

Dudoit, S. and J. Fridlyand (2002). "A prediction-based resampling method for estimating the number of clusters in a dataset." Genome Biol **3**(7): RESEARCH0036.

Efron, B., E. Halloran and S. Holmes (1996). "Bootstrap confidence levels for phylogenetic trees." Proc Natl Acad Sci U S A **93**(14): 7085-7090.

Felsenstein, J. (1985). "Confidence limits on phylogenies: An approach using the bootstrap." Society for the Study of Evolution **39**: 783-791.

Hubert, L. and P. Arabie (1985). "Comparing Partitions." Journal of Classification **2**(2-3): 193-218.

Kislinger, T., B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, J. Rossant, T. R. Hughes, B. Frey and A. Emili (2006). "Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling." Cell **125**(1): 173-186.

Knijnenburg, T. A., L. F. A. Wessels, M. J. T. Reinders and I. Shmulevich (2009). "Fewer permutations, more accurate P-values." Bioinformatics **25**(12): I161-I168.

Krasnitz, A., G. Sun, P. Andrews and M. Wigler (2013). "Target inference from collections of genomic intervals." Proc Natl Acad Sci U S A **110**(25): E2271-2278.

Langfelder, P., B. Zhang and S. Horvath (2008). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R." Bioinformatics **24**(5): 719-720.

Liu, Y., D. N. Hayes, A. Nobel and J. S. Marron (2008). "Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data." Journal of the American Statistical Association **103**(483): 1281-1293.

Monti, S., P. Tamayo, J. Mesirov and T. Golub (2003). "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data." Machine Learning **52**(1-2): 91-118.

Munneke, B., K. A. Schlauch, K. L. Simonsen, W. D. Beavis and R. W. Doerge (2005). "Adding confidence to gene expression clustering." Genetics **170**(4): 2003-2011.

Navin, N., J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks and M. Wigler (2011). "Tumour evolution inferred by single-cell sequencing." Nature **472**(7341): 90-94.

Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Syst Biol **51**(3): 492-508.

Sun, G. and A. Krasnitz (2013). "TBEST: Tree branches evaluated statistically for tightness."The Comprehensive R Archive Network:http://cran.rproject.org/web/packages/TBEST/index.html.