View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Kent Academic Repository

OPEN ACCESS Freely available online

PLoS one

Hierarchical Information Clustering by Means of Topologically Embedded Graphs

Won-Min Song[1], **T. Di Matteo**[1,2], **Tomaso Aste**[1,3]*

1 Applied Mathematics, Research School of Physics and Engineering, The Australian National University, Canberra, Australia, **2** Department of Mathematics, King's College London, London, United Kingdom, **3** School of Physical Sciences, University of Kent, Kent, United Kingdom

Abstract

We introduce a graph-theoretic approach to extract clusters and hierarchies in complex data-sets in an unsupervised and deterministic manner, without the use of any prior information. This is achieved by building topologically embedded networks containing the subset of most significant links and analyzing the network structure. For a planar embedding, this method provides both the intra-cluster hierarchy, which describes the way clusters are composed, and the inter-cluster hierarchy which describes how clusters gather together. We discuss performance, robustness and reliability of this method by first investigating several artificial data-sets, finding that it can outperform significantly other established approaches. Then we show that our method can successfully differentiate meaningful clusters and hierarchies in a variety of real data-sets. In particular, we find that the application to gene expression patterns of lymphoma samples uncovers biologically significant groups of genes which play key-roles in diagnosis, prognosis and treatment of some of the most relevant human lymphoid malignancies.

Citation: Song W-M, Di Matteo T, Aste T (2012) Hierarchical Information Clustering by Means of Topologically Embedded Graphs. PLoS ONE 7(3): e31929. doi:10.1371/journal.pone.0031929

Editor: Ying Xu, University of Georgia, United States of America

Received August 25, 2011; **Accepted** January 17, 2012; **Published** March 9, 2012

Copyright: © 2012 Song et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: No current external funding sources for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tomaso.aste@anu.edu.au

Introduction

Filtering information out of complex datasets is becoming a central issue and a crucial bottleneck in any scientific endeavor. Indeed, the continuous increase in the capability of automatic data acquisition and storage is providing an unprecedented potential for science. However, the ready accessibility of these technologies is posing new challenges concerning the necessity to reduce data-dimensionality by filtering out the most relevant and meaningful information with the aid of automated systems. In complex datasets information is often hidden by a large degree of redundancy and grouping the data into clusters of elements with similar features is essential in order to reduce complexity [1]. However, many clustering methods require some *a priori* information and must be performed under expert supervision. The requirement of any prior information is a potential problem because often the filtering is one of the preliminary processing on the data and therefore it is performed at a stage where very little information about the system is available. Another difficulty may arise from the fact that, in some cases, the reduction of the system into a set of separated local communities may hide properties associated with the global organization. For instance, in complex systems, relevant features are typically both local and global and different levels of organization emerge at different scales in a way that is intrinsically not reducible. We are therefore facing the problem of catching simultaneously two complementary aspects: on one side there is the need to reduce the complexity and the dimensionality of the data by identifying clusters which are associated with local features; but, on the other side, there is a need of keeping the information about the emerging global organization that is responsible for cross-scale activity. It is therefore essential to detect clusters together with the different hierarchical gatherings above and below the cluster levels. In the literature there exist several methods which can be used to extract clusters and hierarchies [1–3] and the application to biology and gene expression data has attracted a great attention in recent years [4–7]. However, in these established approaches, to extract discrete clusters, one must input some a priori information about their number or define a thresholding value. This introduces other potential difficulties because complex phenomena are often associated with multi-scaling signals which cannot be trivially thresholded. In this paper, we propose an alternative method that overcomes these limitations providing both clustering subdivision and hierarchical organization without the need of any prior information, without demanding supervision and without requiring thresholding.

In recent years, several network based approaches have been proposed to describe complex data-sets and applied to several fields from biology [8,9] to social and financial systems [10,11]. Indeed, networks naturally reflect in their set of vertices the variety of elements in the system, they reflect in their edges the plurality of the interrelations between elements and they encode in their dynamics the complex evolution and adaptation of the system [12–16]. In this paper we apply the network paradigm to the study of complex data-structures. In our approach a graph with constrained complexity is built by means of a deterministic construction inserting recursively the most relevant links. In this construction, complexity is constrained by embedding the graph on an hyperbolic surface of genus g (where the genus is the number of handles of the surface) [17,18]. The Ringel-Youngs theorem ensures that for n vertices the complete graph, K_n, can be always embedded on a surface with large enough genus ($g \simeq O(n^2)$) [19]. Any graph is a sub-graph of K_n and therefore any graph can be embedded on a surface. In this paper we are interested in the limit where graphs are sparse and they are embedded on simple surfaces. The simplest case is $g=0$ and the resulting graph is called Planar Maximally Filtered Graph (PMFG) and it is a triangulation of a topological sphere. Topologically embedded graphs on planar surfaces ($g=0$) have a relatively small number of edges ($O(n)$) but they have high-clustering coefficients, they can display various kinds of degree distributions, from exponential to power-law tailed, and they can be used as a platform for modeling other systems [17,20–23]. It has been shown that PMFG graphs are efficient filtering tools having topological properties associated to the properties of the underlying system [18,24]. This makes the PMFG a desirable tool to extract clusters and hierarchies from complex data-sets.

Methods

The general idea at the basis of our method is to use the topological structure of PMFG graphs to investigate the properties of the data-sets. The PMFG is a weighted graph where edges uv have weights $w_{u,v}$ which, in general, are similarity measures (a larger weight $w_{u,v}$ of edge uv corresponds to a stronger similarity between u and v). Furthermore, a distance $d_{u,v}$, or more generally, a non-negative dissimilarity measure is also associated to the edges. Specifically, the PMFG is a graph $G(V,E,W,D)$ where V is the vertex set, E the edge set, W the edge-weight set and D the edge-distance set. A hierarchy in G can be built from a simple consequence of planarity which imposes that any cycle (a closed simple path with the same starting and ending vertex) must be either separating or non-separating [25]. If we detach from the graph the vertices belonging to a separating cycle then two disjoint and non-empty subgraphs are produced. The simplest cycle is the 3-clique which is a key structural element in PMFGs. An example of PMFG is shown in Fig. 1 where the separating 3-cliques are highlighted. By definition, each separating 3-clique, k_p, divides the graph G into two disconnected parts, the *interior* G_p^{in} and the *exterior* G_p^{ex}, that are joined by the clique itself. The union of one of these two parts and the separating clique is also a maximally planar graph. Such a presence of cliques within cliques provides naturally a hierarchy. The subdivision process can be carried on until all separating 3-cliques in G have been considered. The result is a set of planar graphs, that we call "bubbles", which are connected to each other via separating 3-cliques, forming a tree [26]. In Fig. 1(iv) the "bubble tree", denoted hereafter H_b, and its construction are shown. In the bubble tree vertices b_i represent bubbles and edges $b_i b_j$ represent the separating 3-clique, k_p, which is connecting the two bubbles. A direction can be associated to each edge in H_b by comparing the sums over the weights of the edges in the PMFG connecting the 3-clique k_p with the two bubbles. Specifically, a direction can be associated to the edge $b_i b_j$ by comparing the connections of k_p with the interior sub-graph G_p^{in} and the exterior sub-graph G_p^{ex} and considering the two weights

$$W_p^{in/ex} = \sum_{v \in k_p, u \in G_p^{in/ex}} A_G(v,u) \qquad (1)$$

where $A_G(v,u) = w_{vu}$ is the adjacency matrix of G. The direction is given toward the side with largest weight obtaining $\overrightarrow{H_b}$. (In the case of equal weights in the two directions, the two bubbles are joined into a single larger bubble.) In $\overrightarrow{H_b}$ there are three different kinds of bubbles: (1) *converging bubbles* where the connected edges are all incoming to the bubble; (2) *diverging bubbles* where the connected edges are all outgoing from the bubble; (3) *passage bubbles* where there are both inwards and outwards connected edges. An example is provided in Fig. 2 where we have two converging bubbles (b_1 and b_4), one diverging bubble (b_3) and one passage bubble (b_2). Converging bubbles are special being the end points of a directional path that follows the strongest connections and we consider them as the centers of clusters. Any bubble b_i connected by a directed path in $\overrightarrow{H_b}$ to a converging bubble b_α belongs to cluster α. By construction, bubbles in cluster α form a subtree $\overrightarrow{h_\alpha}$ which has only one converging bubble b_α and all edges are directed toward b_α. This is a non-discrete clustering of bubbles because there can be multiple directed paths between b_i and two or more converging bubbles b_α, b_β,… . In Fig. 2(ii) the two subtrees converging toward $b_\alpha = b_1$ and $b_\beta = b_4$ are highlighted, it is clear that in this example bubbles b_2 and b_3 are shared by the two subtrees. A non-discrete clustering of the vertex set $V(G)$ can now be obtained by assigning to each vertex v the cluster memberships of the bubbles that contain it. In order to obtain a *discrete* clustering for $V(G)$, we uniquely assign each vertex to the converging bubble which is at the smallest shortest path distance (see Fig. 2 for a schematic overview). This is achieved in two steps. *First*, we consider the vertices in the converging bubbles. Some vertices belong to only one converging bubble and, in this case, they are assigned to it (e.g. in Fig. 2 vertices v_1 and v_2 are assigned to $b_\alpha = b_1$ and vertices v_6, v_8 are assigned to $b_\beta = b_4$). Other vertices instead belong to more than one converging bubble (e.g. vertices v_3 and v_4 in Fig. 2) and in this case we look at the 'strength' of attachment

$$\chi(v,b_\alpha) = \frac{\sum_{u \in V(b_\alpha)} A_G(v,u)}{3(|V(b_\alpha)|-2)}, \qquad (2)$$

and assign each vertex to the bubble with largest strength. (The notation $|V(b_\alpha)|$ in Eq.2 indicates the number of vertices in the vertex set of b_α and $3(|V(b_\alpha)|-2)$ is the number of edges in the bubble.) After this assignment, each converging bubble α has a unique set of vertices $V^0(\alpha)$. (There can be converging bubbles with an empty set of vertices and, in this case, there will be no clusters associated to them.) *Second*, we consider all the other remaining vertices (e.g. vertices v_5, v_7 and v_9 in Fig. 2). A vertex v may belong to more than one subtree $\overrightarrow{h_\alpha}$, $\overrightarrow{h_\beta}$… and, in this case, it is assigned to the converging bubble that has the minimum mean average shortest path distance

$$\bar{L}(v,\alpha) = mean\{l(v,u) | u \in V^0(\alpha) \wedge v \in V(\overrightarrow{h_\alpha})\} \qquad (3)$$

with respect to all other converging bubbles. Here $l(v,u)$ is the shortest path distance on G from v to u (the smallest sum of distances $d_{r,s}$ over any path between v and u). We have now obtained a discrete partition of the vertex set $V(G)$ into a number of sub-sets $V(\alpha)$, $V(\beta)$,… each respectively associated to the converging bubbles b_α, b_β,… .

Once a unique partition of the vertex set into discrete clusters has been obtained, we can investigate how each of these clusters is internally structured and how different clusters gather together into larger aggregate structures. This can be achieved with a specifically tailored linkage procedure that builds the hierarchy at three levels.

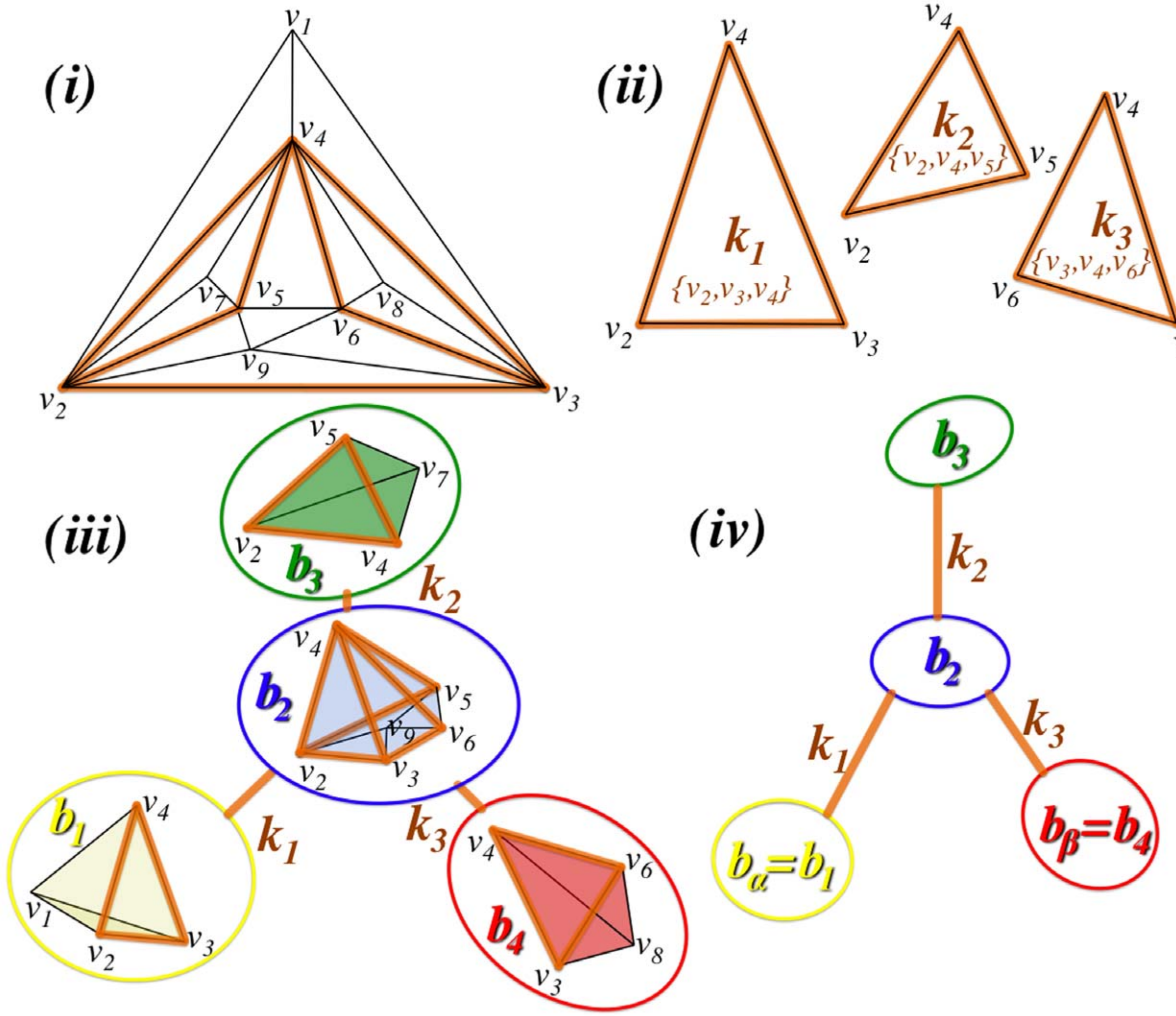


Figure 1. A schematic overview of the construction of the bubble tree. (i) An example of PMFG graph made of nine vertices $V(G)=\{v_1,v_2,v_3,v_4,\ v_5,v_6,v_7,v_8,v_9\}$ and containing three separating 3-cliques: k_1, k_2 and k_3. (ii) The separating 3-cliques have vertex sets: $V(k_1)=\{v_2,v_3,v_4\}$, $V(k_2)=\{v_2,v_4,v_5\}$, and $V(k_3)=\{v_3,v_4,v_6\}$. (iii) The separating 3-cliques identify four planar sub-graphs called ''bubbles'': b_1, b_2, b_3 and b_4 with vertex sets $V(b_1)=\{v_1,v_2,v_3,v_4\}$, $V(b_2)=\{v_2,v_3,v_4,v_5,v_6,v_9\}$, $V(b_3)=\{v_2,v_4,v_5,v_7\}$ and $V(b_4)=\{v_3,v_4,v_6,v_8\}$. (iv) The graph can be viewed as a ''bubble tree'' made of four bubbles connected through three separating 3-cliques.
doi:10.1371/journal.pone.0031929.g001

1. *Intra-bubble hierarchy:* we first assign each vertex $v \in V(\alpha)$ to a bubble b_i in the subtree $\overrightarrow{h_\alpha}$. Vertices in the converging bubbles have been already assigned to the sets $V^0(\alpha)$. For all remaining vertices, the ones belonging to only one bubble are assigned to such bubble (e.g. vertices v_7 and v_9 in Fig. 2). Whereas, vertices that belong to more than one bubble (e.g. vertex v_5 in Fig. 2) are assigned to the bubble that maximizes the strength $\chi(v,b_i)$ (Eq.2). In this way for every cluster α and for each bubble b_i in $\overrightarrow{h_\alpha}$ we have a unique vertex set $V^\alpha(b_i)$ on which we can now perform a complete linkage procedure [27] by using the shortest path distances $l(u,v)$ as distance matrix.
2. *Intra-cluster hierarchy:* we perform a complete linkage procedure between the bubbles in $\overrightarrow{h_\alpha}$ by using the distance matrix

$$d^I_\alpha(b_i,b_j)=\max\{l(u,v)|u \in V^\alpha(b_i) \wedge v \in V^\alpha(b_j)\}. \qquad (4)$$

3. *Inter-cluster hierarchy:* we perform a complete linkage procedure between the clusters by using the distance matrix

$$d^{II}(\alpha,\beta)=\max\{l(u,v)|u \in V(\alpha) \wedge v \in V(\beta)\}. \qquad (5)$$

With this procedure we obtain a novel linkage that starts from the discrete clusters and at higher level joins the clusters into super-clusters and, instead, at lower level splits the clusters into a hierarchy of bubbles and splits the bubbles into a hierarchy of elements. For brevity, in the rest of the paper, we will refer to our clustering and linkage method as the *DBHT technique.*

The computational complexity of this method is smaller than $\mathcal{O}(|V|^3)$ (with $|V|$ the number of vertices, which is equal to the number of variables in the dataset) and it is dominated by the construction of PMFG. Indeed, the Boyer-Myvold algorithm to check planarity [28] runs in $\mathcal{O}(|V|)$ and it might have to be run for each couple of vertices (i.e. $V(V-1)/2$ times). However, typically, the algorithm terminates before the exhaustive scanning of all edges. From empirical tests, performed on various datasets, we

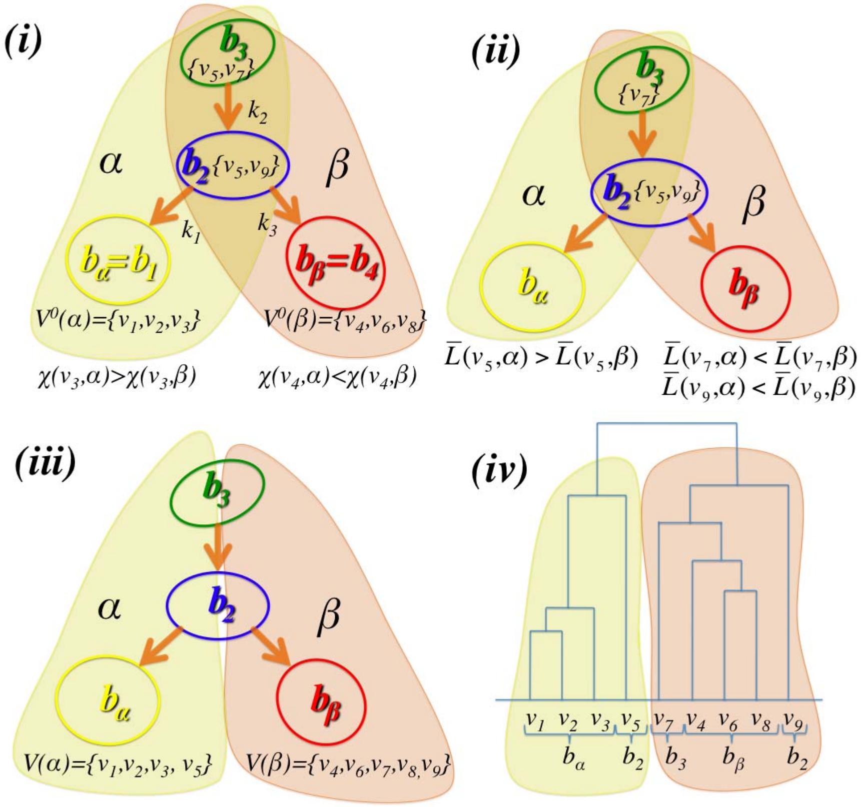

(i)
b_3
$\{v_5,v_7\}$
k_2
α
$b_2\{v_5,v_9\}$
β
k_1
k_3
$b_\alpha=b_1$
$b_\beta=b_4$
$V^0(\alpha)=\{v_1,v_2,v_3\}$
$V^0(\beta)=\{v_4,v_6,v_8\}$
$\chi(v_3,\alpha)>\chi(v_3,\beta)$
$\chi(v_4,\alpha)<\chi(v_4,\beta)$
(ii)
b_3
$\{v_7\}$
α
$b_2\{v_5,v_9\}$
β
b_α
b_β
$\bar{L}(v_5,\alpha) > \bar{L}(v_5,\beta)$
$\bar{L}(v_7,\alpha) < \bar{L}(v_7,\beta)$
$\bar{L}(v_9,\alpha) < \bar{L}(v_9,\beta)$
(iii)
b_3
α
b_2
β
b_α
b_β
$V(\alpha)=\{v_1,v_2,v_3, v_5\}$
$V(\beta)=\{v_4,v_6,v_7,v_8,v_9\}$
(iv)
v_1 v_2 v_3 v_5 v_7 v_4 v_6 v_8 v_9
b_α b_2 b_3 b_β b_2

Figure 2. Illustration of the DBHT technique. (i) Construction of the directed bubble tree where directions are given to the 3-cliques k_1, k_2 and k_3 (from Fig. 1) accordingly with the largest weight W_p^{in} and W_p^{out} (see Eq.1). In this example we have two converging bubbles: $b_\alpha = b_1$ and $b_\beta = b_4$. A unique set of vertices can be associated to each of the two converging bubbles b_α and b_β where vertices shared by both the converging bubbles (i.e. the vertices v_3 and v_4) are assigned accordingly with the largest strength χ (Eq.2). (ii) All the other non-assigned vertices (i.e. v_5, v_9 and v_7) are associated to the cluster with minimum average shortest path length $\bar{L}$ (Eq.3). (iii) The vertex set is uniquely divided into two clusters respectively associated to the two converging bubbles: $V(\alpha)=\{v_1,v_2,v_3,v_5\}$ and $V(\beta)=\{v_4,v_6,v_7,v_8,v_9\}$. (iv) The hierarchical organization and the clustering structure can be represented with a dendrogram.
doi:10.1371/journal.pone.0031929.g002

measured an overall runtime of $\mathcal{O}(|V|^\alpha)$ with $\alpha \sim 2.7$. (See Supporting Information S1 and S2.)

Results

In this section, we apply the DBHT technique to various data sets ranging from artificial data with known clustering and hierarchical structures to real gene expression data. Comparisons are made between the results retrieved by the DBHT technique and some of state-of-the-art cluster analysis techniques such as k-means++[29], Spectral clustering via Normalized cut on k-nearest neighbor graph (kNN-Spectral) [30,31], Self Organizing Map (SOM) [32] and Q-cut [33]. Let us here stress that all these techniques –except DBHT– are non-deterministic and require some *a priori* information in order to setup the initial parameters. To compare with the DBHT technique, we run the other techniques for a broad range of parameters and pick the set of parameters that are best performing in average. This is an important negative bias against the DBHT technique that however, as we shall see shortly, can still outperform consistently the state-of-the-art counterparts. We also tested the capability of DBHT technique to correctly detect the hierarchical organization by applying it to known synthetic datasets and comparing the results with the outcomes from average and complete linkage techniques. Furthermore, we explored the meaningfulness of the hierarchical gathering of clusters and the significance of their subdivision in sub-clusters by looking at the functional properties of these gatherings and splittings in real datasets.

Tests DBHT clustering on synthetic data

We have evaluated performance of the clustering techniques by comparing their outcomes with the known artificial clustering structure by using a popular external validity index: the adjusted Rand index [34] which returns 1 for a perfect match and in average 0 for a random guess. Specifically, we have generated correlated data-series by using a multivariate Gaussian generator (MVG) [35] that produces N stochastic time series $y_i(t)$ of length $T = 10 \times N$ with zero mean and Pearson's cross-correlation matrix R that approximates an input correlation structure R^* which is a block-diagonal matrix where the blocks represent the clusters and may have different sizes. The matrix R^* has all ones on the diagonal, it has zero correlations outside the blocks ($\rho^{ou*} = 0$) and it has a correlation value ρ^{in*} inside the blocks. Furthermore, we have added a number N_{ran} of random correlations unrelated to the cluster structure. We have also generated multivariate Log-Normal distributions by taking the exponential of MVG series generated by using reference correlation R^*_{log} which is devised to retrieve the correct approximation of R^* with log-normal statistics [36]. To these correlated series we have added a noise $\eta_i(t)$ obtaining $y_{i'}(t) = y_i(t) + c\sigma_i\eta_i(t)$, where σ_i is the standard deviation of $y_i(t)$ and c is a constant that can be used to tune the relative amplitude of noise. We have tested normally distributed ($p(\eta)\propto\exp(-\eta^2/2)$), log-normally distributed ($p(\eta)\propto\exp(-\log(\eta)^2/2)$) or power-law distributed ($p(\eta)\propto 1/\eta^{\alpha+1}$) noises. We have used different values for the relative amplitude of noise c and, in the case of power-law distributed noise, we have also varied the exponent α. By increasing the effect of noise and/or the number of random elements, the Pearson's cross-correlation matrix R passes from a very well defined structure similar to R^* to a less defined structure where the difference between the average measured intra- and inter-cluster correlations in R, $<\rho^{in}> - <\rho^{ou}>$, becomes negligible.

Figure 3 compares the performance of the DBHT technique with k-means++, SOM, kNN-Spectral and Q-cut for correlated synthetic datasets consisting of 129 data series generated both with normal and log-normal statistics, with normal or power law noise with $\rho^{in*}=0.9$, $\rho^{ou*}=0$ and $N_{ran}=25$. This example refers to a rather extreme case where the clusters have highly dis-homogeneous sizes with one large cluster with 64 elements and eight clusters with 5 elements each. As one can see from Fig. 3 in this case the DBHT technique is strongly outperforming the other methods. In the Supporting Information S1, we report on a large number of cases where we demonstrate that consistently the DBHT technique is better, or at least equivalent, to the best performing counterparts for a very broad range of combinations of different kinds of artificial data. Let us here note that stochastic techniques such as k-means++ and SOM are particularly sensitive to noise distributions and tend to perform poorly with fat-tailed distributed noise. On the other hand, the Qcut technique carries an inherent resolution limit that over-shadows small clusters [37]. The DBHT technique instead is less affected by these factors and it consistently delivers good performances across the range of parameters.

Tests DBHT hierarchy on synthetic data

We have tested the capability of the DBHT technique to detect hierarchies by simulating data with hierarchical structure such that smaller clusters are embedded inside larger clusters making a nested structure with different intra-cluster correlations. An example is shown in Fig. 4(a) where we report an input correlation R^* which is a nested block-diagonal matrix with zero inter-cluster correlation and with a structure of 4 'large' clusters (64 elements each) with intra-cluster correlation of $\rho_1^{in*}=0.7$. Each of the large clusters contains inside two 'medium' clusters (8 in total with 32 elements each) with $\rho_2^{in*}=0.8$ that contain inside two 'small' clusters (16 in total with 16 elements each) with $\rho_3^{in*}=0.95$. We have simulated 30 different sets of data series of length $T=10\times N$ by using MVG from R^* with added power law noise with $\alpha=1.5$ and $c=0.1$. We have tested the efficiency of the DBHT technique by moving through the hierarchical levels varying the number of clusters from only one at the top hierarchy to the number of elements at the lowest hierarchy. Fig. 4(b) shows the dendrogram retrieved with the DBHT technique. By following the hierarchy from top to bottom, one can see that a structure with 4 main clusters rapidly emerges and its partition coincides exactly with the 'true' partition in R^*. Then these clusters correctly split into two parts each making 8 clusters in total scoring a value of 0.97 for the adjusted Rand index with respect to the 'true' partition at this level. Finally, these 8 clusters split again producing a partition that has an adjusted Rand index of 0.94 with respect to the 'true' partition at this level. The partition into discrete clusters identified by the DBHT is almost identical with this last one having 17 clusters instead of the 16 'true' clusters and achieving also an adjusted Rand index of 0.94 (see Supporting Information S1). One can see from Fig. 4(c,d) that, instead, the complete and average linkages give a less clear hierarchical structure. Several other examples are reported in the Supporting Information S1. The better performance of the DBHT technique over linkage methods can be explained by the fact that linkage techniques suffer from the greedy nature of the algorithm, where a misclassification of an element in an early stage of clustering can never be remedied [1,3]. The rate of misclassification depends on the type of linkage distance, with the average linkage optimized for isotropic clusters, and complete linkage optimized for compact and well-defined clusters. On the other hand, DBHT hierarchy is based on a combination of linkage distance and topological constraints at multiple hierarchical levels: bubbles, clusters, bubble tree. This reduces the error rate with respect to the complete linkage distance.

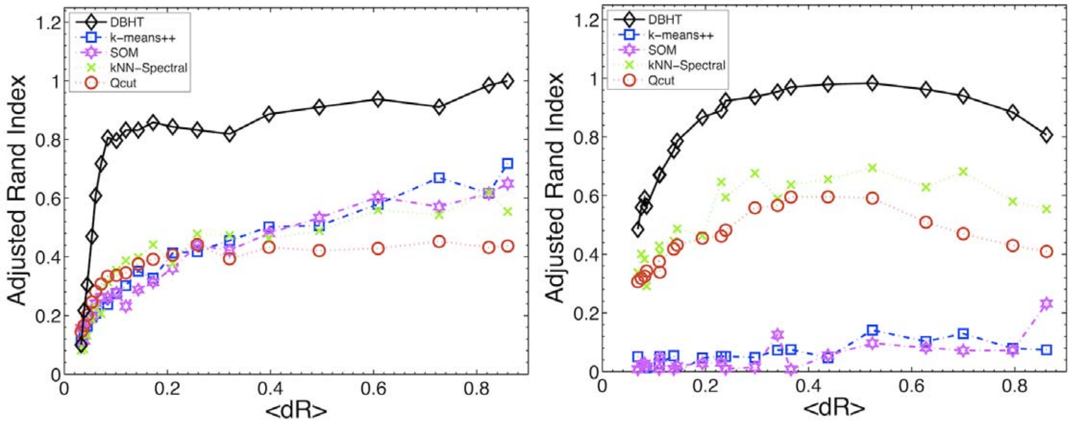


Figure 3. Demonstration that the DBHT technique can outperform other state-of-the-art clustering techniques, namely: k-means++[29], Spectral clustering via Normalized cut on k-nearest neighbor graph (kNN-Spectral) [30,31], Self Organizing Map (SOM) [32], and Q-cut [33]. The figures report the adjusted Rand indexes [34] for the comparison between the the 'true' partition embedded in the artificially generated data and the partition retrieved by the clustering methods. In these examples we have eight clusters of size 5 elements and one cluster of size 64 elements with $\rho^{in*}=0.9$, $\rho^{ou*}=0$ and $N_{ran}=25$. The plots report average values over a set of the 30 trials. The horizontal-axis reports the gap between average intra- and inter-cluster correlations $dR=<\rho^{in}>-<\rho^{ou}>$ that becomes smaller when the noise c increases. (**a**) Normally distributed correlated datasets with added Normal noise with c varying from 0 to 4. (**b**) Log-Normally distributed correlated datasets with added power law noise with $\alpha=1.5$ and c varying from 0 to 0.1.
doi:10.1371/journal.pone.0031929.g003

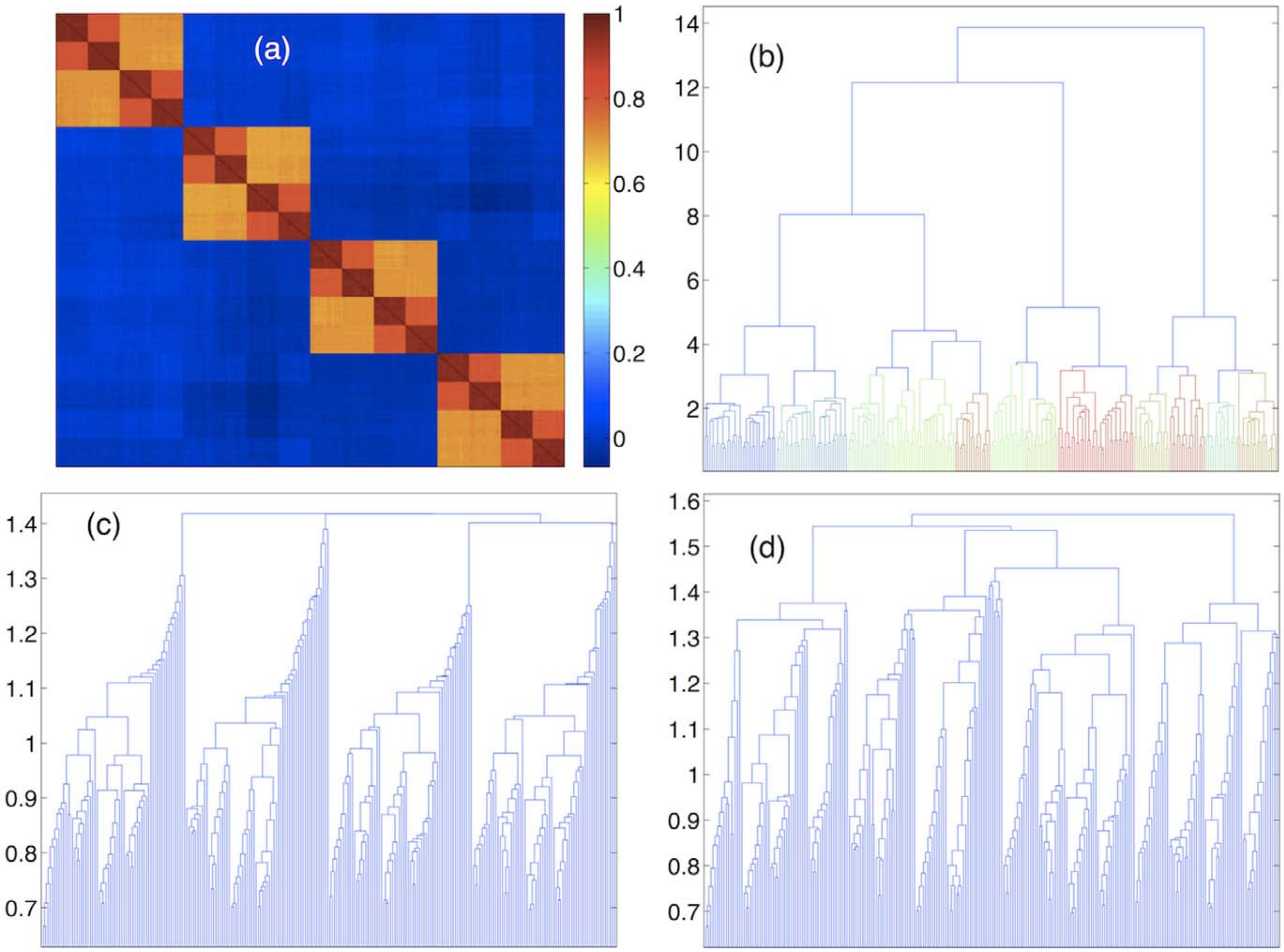


Figure 4. Demonstration that the DBHT technique can detect clusters at different hierarchical levels outperforming other established linkage methods. The synthetic data are generated via a multivariate Gaussian generator with added power law noise with exponent $\alpha=1.5$ and $c=0.1$. (**a**) Input correlation R^* for a synthetic data structure with nested hierarchical clustering with 4 'large' clusters, containing 8 'medium' clusters, containing 16 'small' clusters. (**b**) Dendrogram associated with the DBHT hierarchical structure. (**c**) Dendrogram associated with the Average linkage. (**d**) Dendrogram associated with the Complete linkage.
doi:10.1371/journal.pone.0031929.g004

Application of DBHT technique to Fisher's Iris Data

One of the typical benchmark referred in clustering analysis literature is the iris flower data set from Fisher [38]. Briefly, the data set contains the measure of four features (i) sepal length; (ii) sepal width; (iii) petal length; (iv) petal width, for 50 iris plants from three different types of iris, namely (1) Iris Setosa; (2) Iris Versicolour; (3) Iris Virginica. The data set is available from UCI Machine Learning Repository website [39]. It is known that, the clustering structure of the data set linearly separates one type of Iris from the other two. The remaining two types are instead not linearly separable and their subdivision is a classical challenge for any clustering technique [39]. Here, in order to compute clustering and hierarchies we have used the pair-wise Euclidean distance $\mathbf{D}_{\mathbf{euc}}(i,j)=\|x_i-x_j\|$ as dissimilarity matrix and $\mathbf{R}_{\mathbf{euc}}(i,j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$ as similarity matrix [31], where σ is the standard deviation of $\mathbf{D}_{\mathbf{euc}}(i,j)$ for all pairs of (i,j). From these measures, we directly computed clustering and hierarchies via DBHT technique obtaining the graph structure shown in Fig. 5(a) where one can see that all the three iris types are rather well separated occupying different parts of the graph. By extracting three clusters from the DBHT hierarchy we observe that the first flower type (Iris Setosa) is fully separated and the other two are rather well divided with only a few misplacements. The DBHT results are compared with other two graph-based techniques, Qcut and kNN-Spectral techniques computed using $\mathbf{R}_{\mathbf{euc}}$ for a range of $kNN=2,\ldots,(N-1)$. These methods are non deterministic and we retained only the best partitions which give the highest adjusted Rand score which are shown in Fig. 5(b,c). We can observe that Qcut and kNN-Spectral techniques provide a poorer separation of the last two flower types (Iris Versicolour and Iris Virginica). This is quantified by the adjusted Rand index computed by comparing with the true partition that gives 0.89 for DBHT and 0.85 for both Qcut and kNN-Spectral. Indeed, these last two techniques both misplace 8 elements of the two groups whereas DBHT misplaces only six. Other two clustering techniques, k-means++ and SOM, have been run over 30 iterations with an input number of clusters $k=3$, yielding to poorer partitions with the largest adjusted Rand indexes respectively of 0.73 and 0.80 which are well below the score achieved by the DBHT technique. The iris flower data set and the codes to reproduce the result in Fig. 5(a) are provided in the Supporting Information S2.

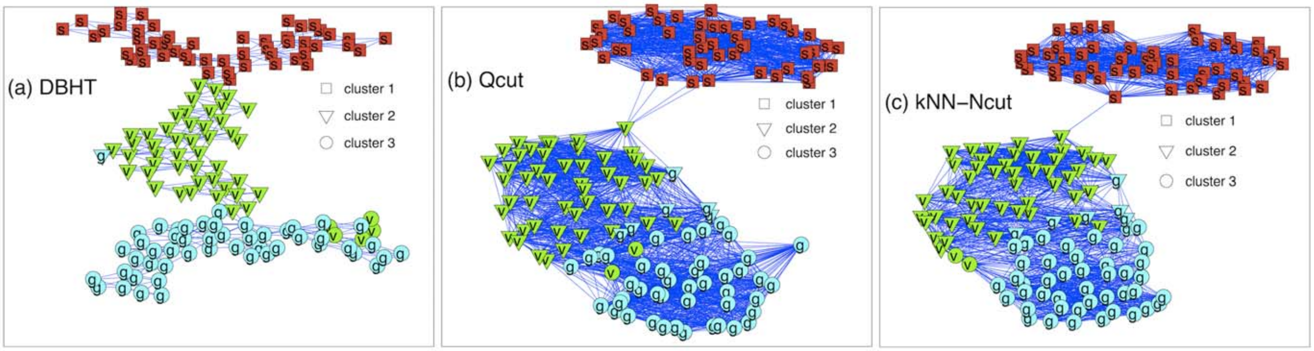


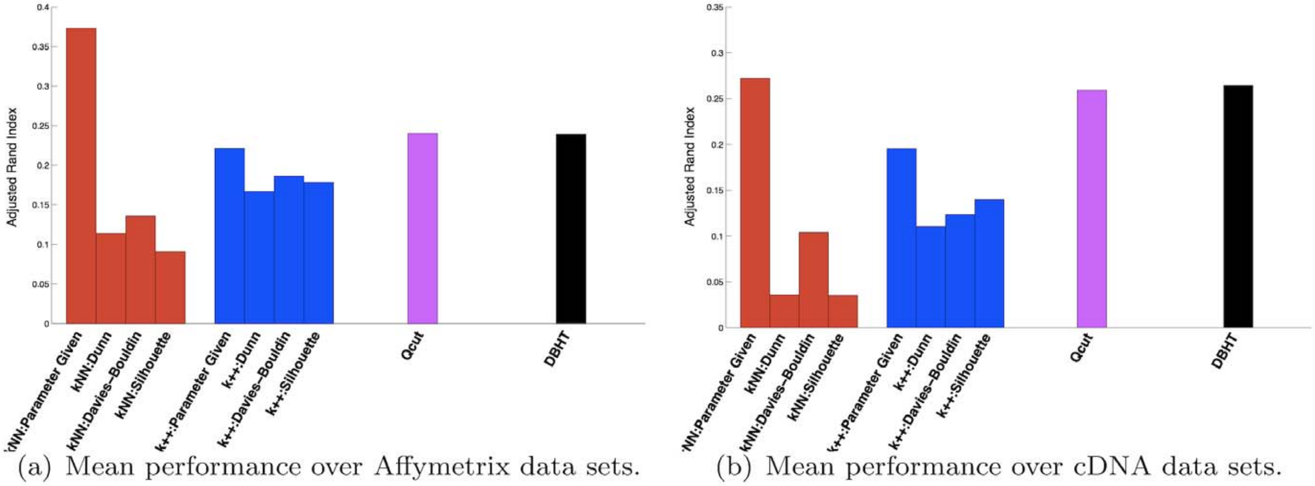
Figure 5. Comparison between the clustering obtained via: (a) DBHT technique, (b) best Qcut and (c) best kNN-Spectral on iris flower data set from Fisher [38]. The labels inside the symbols correspond to the three different types of flowers: (s) Iris Setosa; (v) Iris Versicolour; (g) Iris Virginica. The shapes of the symbols correspond to the clusters retrieved by the different clustering techniques.
doi:10.1371/journal.pone.0031929.g005

Application of DBHT technique to a benchmark gene expression dataset

In order to validate the applicability of DBHT technique on gene expression data, we have used a benchmark dataset collected by de Souto *et al* [40] which contains 21 Affymetrix and 14 cDNA gene expression patterns from different cancer types. For this dataset we have compared clusters computed via DBHT technique with clusters computed with k-means++, kNN-Spectral, and Qcut evaluating the respective performances by using the Adjusted Rand index [34]. Differently from DBHT that requires no prior parameters, k-means++ and kNN-Spectral require instead information on the number of clusters, and we therefore tested two cases: (i) clustering with benchmark number of clusters given *a priori*; (ii) clustering with an internal validity measure to estimate the optimal number of clusters (namely: Dunn index [41], Davies-Bouldin [42] and Silhouette width [43], see Ref. [43]). Another requirement for kNN-Spectral is the number of nearest neighbors kNN. We have used $kNN \in [2,8]$ (as indicated in Ref. [31]) and picked the case with best mean performance. Also Qcut requires to choose the value of kNN. In this case, we have used $kNN \in [3,5]$ (as suggested by Ref. [33]) and selected the value which yields to the best Q.

The results are shown in Fig. 6. We can see that DBHT and Qcut achieve the best average performances when the number of clusters is not given as input. Instead, when the benchmark number of clusters is supplied, then kNN-Spectral shows the best mean performances, followed by DBHT and Qcut that perform similarly, and finally k-means++. Let us stress that the true number of clusters is an important piece of information that is not available in most practical cases and therefore an high performance in this case may not be of practical relevance. However, we note that, even in this unfavorable case, the DBHT can perform extremely well. Indeed, if we look at the performances for each sample (see Fig. 7) we see that DBHT can achieve the best performance for many cDNA data and for some Affymetrix data.

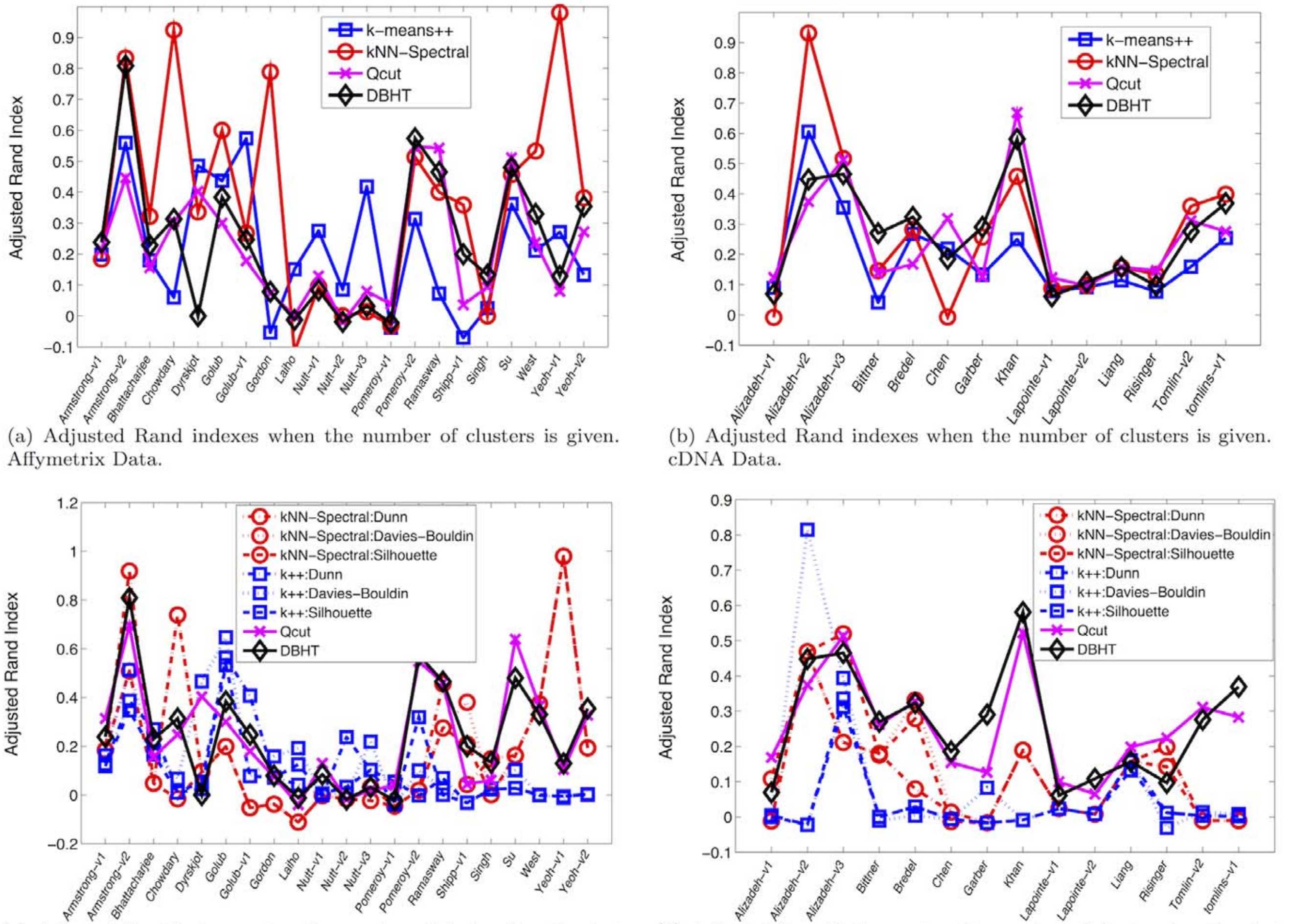
Let us remark that, the 'golden standard' clusters provided by de Souto *et al* do not necessarily represent the true and meaningful underlying structure of the gene expression data. For example, in Fig. 8, we have analyzed in details the special case of Yeoh-v1 Affymetrix data which gives outstanding performance for kNN-Spectral technique and poorer performance for the DBHT technique (see Fig. 7). Fig. 8(a) shows the correlation structure R of the data set visualized according to the known golden standard



(a) Mean performance over Affymetrix data sets.

(b) Mean performance over cDNA data sets.

Figure 6. Average Adjusted Rand index to compare performances of clustering algorithms: k-means++, Qcut, kNN-Spectral and DBHT for the benchmark data sets collected by de Souto *et al* [40] (k++ indicates k-means++). The relatively high performing "Parameter given" results refer to cases when the true number of cluster is given to the algorithm as input. In all the other cases the number of cluster is computed by using internal validity measures. (a) Affymetrix data; (b) cDNA data.
doi:10.1371/journal.pone.0031929.g006



(a) Adjusted Rand indexes when the number of clusters is given. Affymetrix Data.

(b) Adjusted Rand indexes when the number of clusters is given. cDNA Data.

(c) Adjusted Rand indexes when the number of clusters is estimated. Affymetrix Data.

(d) Adjusted Rand indexes when the number of clusters is estimated. cDNA Data.

Figure 7. Adjusted Rand indexes for each sample in the de Souto *et al* [40] datasets. (Top) Performances for each dataset when the true number of cluster is given as input. (Bottom) Performances for each dataset when the true number of cluster is computed by using internal validity measures. (Left) Affymetrix data; (Right) cDNA data.
doi:10.1371/journal.pone.0031929.g007

cluster structure. One can see that, beside the golden standard clusters, there is also a finer meaningful structure that is not detected by kNN-Spectral but it affects instead the DBHT clustering. In this case, the high performance of the kNN-Spectral is a consequence of a coarse-grained picture which is not necessarily best reflecting all the features of the dataset. In general, in practical cases, the real clustering structure is often ambiguous and the availability of clustering methods based on different criteria is a key ingredient to properly explore these structures. Often, the subdivision into distinct clusters is not well defined and the information provided by the DBHT thechnique, concerning the hierarchical way in which clusters split into sub-parts and in which they merge into larger aggregates, can become essential.

Application of DBHT technique to gene expression data set from human cancer samples

We have applied the DBHT technique to analyze gene expression data sets collected by Alizadeh *et al* [44] concerning 96 malignant and normal lymphocyte samples belonging to the three most relevant adult lymphoid malignancies, namely: Diffuse Large B-Cell Lymphoma (DLBCL); Follicular Lymphoma (FL); Chronic Lymphocytic leukemia (CLL); together with other 13 kinds of samples from normal human tonsil, lymph node, Transformed Cell Line, Germinal Centre B, Activated Blood B, and Resting Blood B. This data set has already served as a benchmark to evaluate performance of clustering techniques on gene expression data [33,45] and this is why we have chosen to test our method on this referential dataset. Patients with DLBCL cancer type have variable clinical courses and different survival rates and there are strong indications that DLBCL classification includes more than one disease entity [44]. The challenge for a clustering algorithm is therefore to analyze the DLBCL genetic profiles and individuate different subtypes of DLBCL to be associated with different clinical courses. Indeed, various studies have attempted to highlight genetically significant genes that can be of clinical significance to improve the DLBCL patients' diagnosis and clinical treatments [44,46–51]. In particular, it is understood that DLBCL is a very heterogeneous type of Lymphoma and there are at least three distinct subtypes which differ in treatment methods for improved survival of the patients [44,46,52].

We have first applied the DBHT technique on the gene expression data by using Pearson's correlation as similarity measure, and correlation distance as the dissimilarity measure.

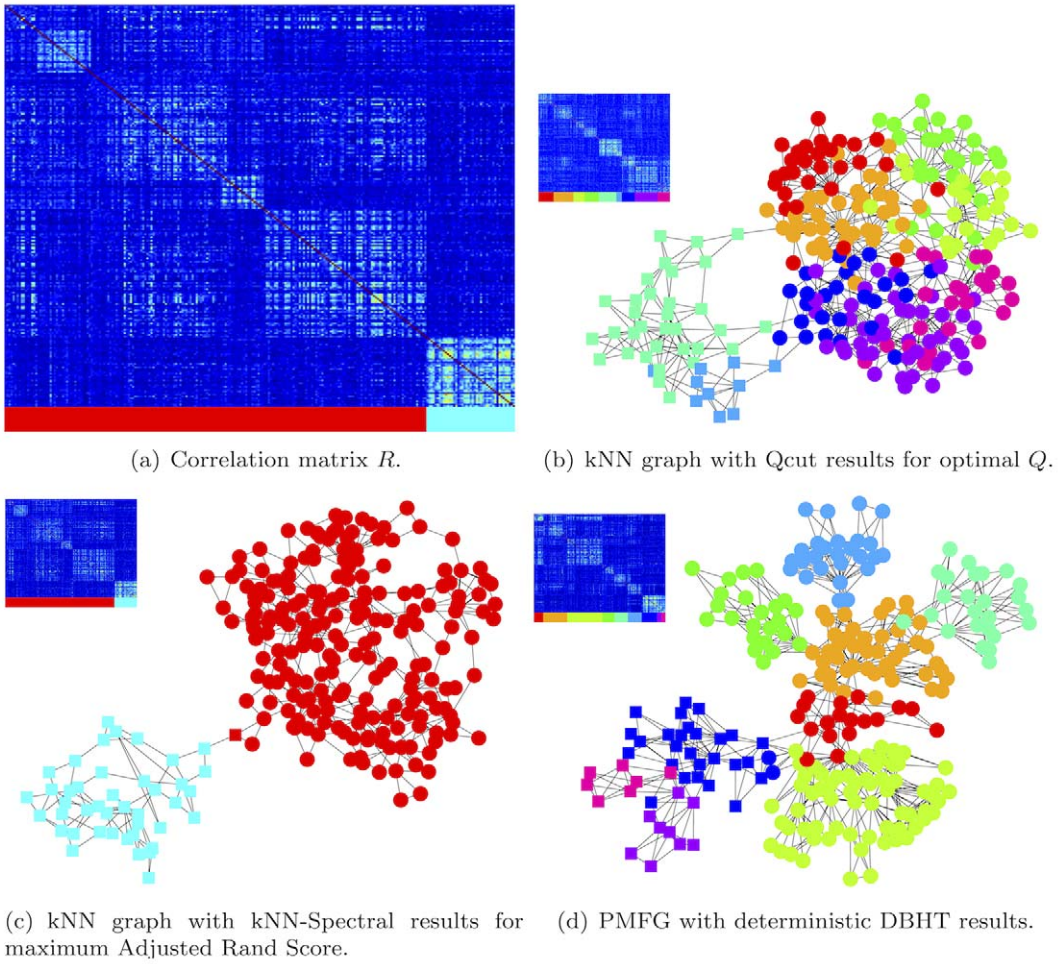


Figure 8. Comparison between the clusters obtained with the DBHT method and the clusters obtained from kNN graph with Qcut results for optimal Q for the dataset Yeoh-v1 Affymetrix [40]. (a) Correlation matrix structure R, which are ordered accordingly with the 'known' clustering structure of Yeoh-v1 data. (b, c, d) *Insets*: correlation matrices R ordered accordingly with the Qcut, kNN-Spectral and DBHT respectively. The clusters are indicated on the bottom with color bars. (b, c, d) *Main plots*: results for Qcut, kNN-Spectral and DBHT respectively where the 'golden standard' clusters for Yeoh-v1 data (as by de Souto *et al* [40]) are depicted in vertices of different shapes: square or circle. The computed clusters are instead depicted in different colors, shown both in the graphs and in the color bars on the bottom of the Correlation matrix. One can note that, despite kNN-Spectral technique gives a very good agreement with the 'golden standard' provided by de Souto *et al*, the structure extracted by the DBHT method gives a very clean clustering partition that is clearly revealed in the visualization of the relative correlation matrix in the inset of (d).
doi:10.1371/journal.pone.0031929.g008

The DBHT clustering yielded to 11 sample-clusters, which are shown in Fig. 9. One can immediately note that all FL samples are gathered together in one cluster that also contains the DLCL-0009 sample, which has been associated to FL in other studies on the same data [33,44]. Transformation of FL to DLBCL is common [53], and this cluster suggests that DLCL-0009 may have derived from FL, sharing therefore common gene expression patterns. We also observe in Fig. 9 that all, except one, the CLL samples occupy a single cluster. The missing CLL sample is attached to this cluster and it is included in a cluster containing Resting Blood B samples which have indeed similar expressions patterns and clinical similarity to CLL and are often merged together by other clustering techniques [33]. DLBCL cancer types appear in four different sample-clusters which are however lying together in a branch of the PMFG graph. Significantly, these clusters also include some other GCB-like samples. Remarkably, if we look at the patient survival rates (Table 1), we see that these four sample-clusters are extracting DLBCL cancer subtypes with very different clinical courses. Indeed, if we consider separately the patients with DLBCL type of Lymphoma accordingly with the subdivision into the four sample-clusters '1', '5', '7' and '9' (from bottom to top of the Fig. 9), they respectively have survival rates 100%, 56%, 15% and 29% (see Table 1 for details). In the work of Alizadeh *et al* [44] survival rate differentiation in DLBCL patients was associated with two main cancer subtypes, namely GCB-like and ABC-like, with the latter considered more fatal than the former. We can note that, in our clustering, sample-cluster '1' contains GCB-like DLBCL, and it also includes other GCB samples such as tonsil GCB, tonsil GC fibroblast, and high survival rates are common in GCB-like cancer types (see Supporting Information S1). Cluster '5' is also characterized by GCB-like DLBCL samples, however its proximity to ABC-like clusters (see Supporting Information S1), may be the clue to relatively low survival rate in comparison to cluster '1'. Cluster '9' is characterized by a majority of ABC-like DLBCL to which we may attribute its relatively low survival rate [44]. On the other hand, cluster '7', which shows a surprisingly low survival

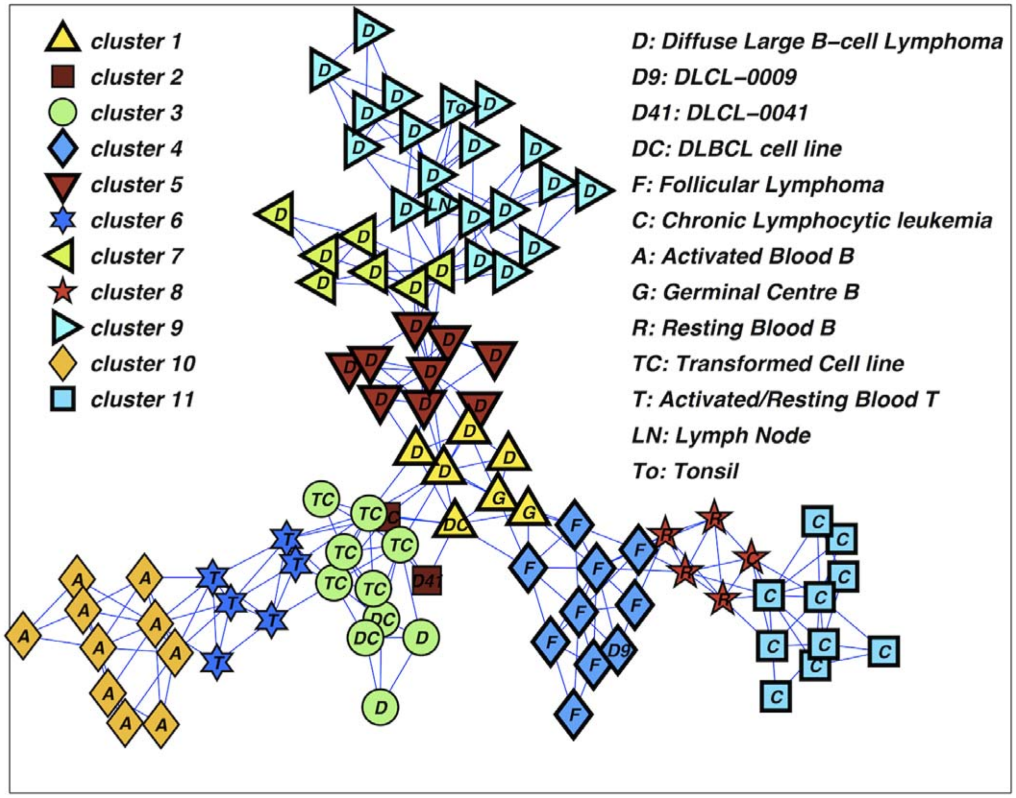


Figure 9. Sample-cluster structure for 96 malignant and normal lymphocyte samples from Alizadeh *et al* 2000 [44], the labels inside the symbols correspond to the different sample types as listed in the legend. The DBHT technique retrieves 11 sample-clusters here represented with different symbols (see legend). The underlying network is the PMFG from which the clustering has been computed.
doi:10.1371/journal.pone.0031929.g009

rate, has instead a significant number of GCB-like DLBCL samples, this might signal the existence of another relevant DLBCL subtype. The gene expression data and the codes to reproduce the result in Fig. 6 are provided in the Supporting Information S2.

In order to functionally validate these sample-clusters, we have analyzed the expression profiles for 6 groups of genetic clones with known physiological roles, namely: GCB- Germinal Center B cell (111 clones), LyN- Lymph Node (136 clones), PBC- Pan B Cell (81 clones), Pr- Proliferation (312 clones), TC- T Cell (111 clones) and ABC- Activated B Cell (86 clones) [44]. The significance of regulation patterns has been evaluated by one-tailed T tests with cut-off p-value of 0.01. The number of up-/down-regulated profiles for each group of clones is shown in Table 2. Significant up-/down-regulation patterns of the expression profiles in the sample-clusters reflect the biological relevance the group of gene-clones in each sample-cluster. We first observe that sample-clusters containing DLBCL cancer types (e.g. cluster '1', '5', '7' and '9') distinguish from other samples by up-regulating more clones from Pr, hence reflecting higher proliferative index. Sample clusters associated to DLBCL are also differentiating among themselves, for instance, sample-clusters '1' and '5' both up-regulate GCB clones but they differ significantly in the up-regulation of LyN clones, supporting the subdivision of GCB-like DLBCL by these sample clusters. Similarly, sample-cluster '7' shows a unique expression signature that highlights a strong up-regulation of LyN clones in comparison to other clones. Given that this sample-cluster is a mixture of ABC-like and GCB-like DLBCLs, and it shows distinctively low survival rate, this again suggests that sample-cluster '7' is a different subtype of DLBCL outside of GCB-/ABC-like classification. Overall, these results indicate that DBHT clustering technique is able to reveal a meaningful

Table 1. Survival rates of cancer patients with DLBCL type of Lymphoma. The patients are divided in four groups corresponding to the four sample-clusters containing DLBCL obtained with DBHT technique (see Fig. 6).

	Sample Cluster '1'	Sample Cluster '5'	Sample Cluster '7'	Sample Cluster '9'
Cluster Size	7	9	7	20
# of DLBCL	4	9	7	17
# Survived over 5 yrs	3 (100%)	5 (56%)	1 (14%)	5 (29%)
# Died in 5 yrs	0	4	6	12

doi:10.1371/journal.pone.0031929.t001

Table 2. Number of up-regulated (on the left) and/down-regulated (on the right) expression profiles for each group of clones with known physiological roles as reported in Ref. [44].

	GCB	LyN	PBC	Pr	TC	ABC
Sample Cluster '1'	**61/0**	0/2	27/0	**115/0**	1/15	4/12
Sample Cluster '2'	2/0	0/2	0/2	7/3	0/1	0/3
Sample Cluster '3'	0/35	2/37	0/15	259/0	0/38	4/3
Sample Cluster '4'	83/0	0/97	48/0	1/193	3/12	0/37
Sample Cluster '5'	**21/2**	97/0	7/3	**119/0**	2/4	0/11
Sample Cluster '6'	7/27	1/47	0/61	6/126	86/0	32/0
Sample Cluster '7'	4/6	**111/0**	0/24	**17/4**	14/3	13/1
Sample Cluster '8'	0/2	0/41	17/1	0/199	6/4	2/7
Sample Cluster '9'	1/13	133/0	7/1	**70/0**	14/4	24/2
Sample Cluster '10'	0/37	3/48	1/14	44/68	1/20	61/0
Sample Cluster '11'	20/43	0/110	27/12	0/303	20/16	1/56

The sample-clusters are obtained by DBHT technique and labels are as in Fig. 9. Some significant up-/down-regulation patterns, commented in the text, are highlighted by boldface font.
doi:10.1371/journal.pone.0031929.t002

classification of biologically significant DLBCL subtypes which is richer than what proposed in the original study by Alizadeh *et al* [44].

Let us now move a step further and use the DBHT technique to identify significant groups of genes that are of relevance for particular cancer samples. Indeed, an accurate identification of significant genes is crucial in treating the tumor cells as there are a large number of different genetic mechanisms from which these tumor cells originate, hence they require different treatments [54,55]. We have therefore performed a two-way clustering: on the samples and genes simultaneously. In this way, we can cross-tabulate the samples against genes obtaining a simple and effective picture of significant gene expression patterns. Let us note that with conventional clustering techniques, the two-way clustering adds another dimension of complexity. Indeed, samples and gene expression profiles have different dimensions and scales and therefore it is necessary to tune the clustering parameters separately for each clustering way. On the other hand, the DBHT technique has no adjustable parameters and it is deterministic providing therefore a unique cross classification without any increase in complexity. The DBHT technique identifies 180 gene-clusters from which we have extracted 6 clusters which are significantly differentiating for sample-clusters associated to FL, CLL and DLBCL, accordingly with a p-value threshold of 0.01 with Bonferroni correction. The expression profiles of these significant gene-clusters are reported in Fig. 10. We have then validated functional significance of these gene-clusters by performing a gene-ontology (GO) analysis to identify significant GO terms for biological processes [56]. (See Supporting Information S1 for the statistical analysis methods and GO results.) Let us here report on some relevant genes, from each of the 6 significant gene-clusters, selected by choosing the most frequently appearing genes in the GO terms. Interestingly, these genes reveal some of biologically significant mechanisms that regulate growth of tumor cells, and that affect survival of respective lymphoma malignancy. In particular:

- Gene cluster '44' (significant for sample-cluster '1'): This gene-cluster is up-regulated for sample-cluster '1' in comparison to the expressions in other sample-clusters associated to lymphoma. Significantly, one of its key genes is CDK1, which is a key player in cell cycle. It has been indicated that over-expression of CDK1 is common in DLBCL cancer types, and it is therefore a potential therapeutic target [57].
- Gene cluster '4' (significant for sample-cluster '4'): This gene-cluster particularly expresses for sample-cluster '4', which consists mostly of FL samples. Among the genes in this gene-cluster there is SYK which -indeed- has been indicated as a promising target gene for antitumor therapy for treating FL, where inhibition of SYK expression increases the chance of survival [58].
- Gene cluster '1' (significant for sample-cluster '5'): Gene cluster 1 is particularly down-regulated for sample-cluster '5'. This gene-cluster contains TGF-B1 which is a well-known transcription factor to regulate proliferation, in particular a negative regulator of B-cell lymphoma which induces apoptosis of the tumor cells via NF-κB/Rel activity [59]. This suggests that suppression of the tumor cells by TGF-B1 would be lessened in sample-cluster '5' due to the down-regulation, and this may contribute to the decreased chance of survival observed in sample-cluster '5' in comparison to that of sample-cluster '1'.
- Gene cluster '4' (significant for sample-cluster '7'): This gene-cluster is slightly down-regulated for sample-cluster '7', and GO analysis extracts two genes, CDKN1B/p27^{Kip1} and CDKN2D/p19, which are key tumor suppressor genes for aggresive neoplasms [60,61]. The inhibited tumor suppressive role of these genes might have led to aggressive growth of tumor cells suggesting a plausible explanation for the poorest survival rate, observed for sample-cluster '7', with respect to the other DLBCL sample-clusters (see Table 1). Indeed, it has been suggested that p27 is associated to lymphomagenesis through Skp2 [61] and Skp2 has been indicated as an independent marker to predict survival outcome in DLBCL [61,62].
- Gene cluster '125' (significant for sample-cluster '9'): This gene-cluster shows distinct up-regulation pattern for sample cluster '9', and it includes an interesting gene 'IL-6'. IL-6 is known to be a central target gene in a synergistic crosstalk between NF-κB and JAK/STAT pathway, which is a unique feature for some DLBCL [55]. It is suggested that, these have implications for targeted therapies by blocking STAT3 expression, a gene that is activated by IL-6 [55,63].
- Gene cluster '102' (significant for sample-cluster '11'): This gene-cluster particularly down-regulates the CLL sample-cluster among all lymphoma-related clusters. Though it does not indicate a particularly significant GO term (see Supporting Information S1), it includes a number of genes related to regulating tumor cell growth for CLL (see Supporting Information S1 for the list of genes). Among these genes, let us note IRF1, which is a well-known mediator for cell fate by facilitating apoptosis, and it is also a tumor suppressor [64]. As the expression of IRF1 is slightly down-regulated, we suspect that this may contribute to the growth of CLL tumor cells.

In conclusion let us stress that these results strongly indicate that the DBHT technique can detect relevant differentiations and aggregations in both cancer-samples and gene-clones revealing important relations that can be used for diagnosis, for prognosis and for treatment of these human cancers.

Discussion

In summary, we have introduced a novel approach, the DBHT technique, to extract cluster structure and to detect hierarchical

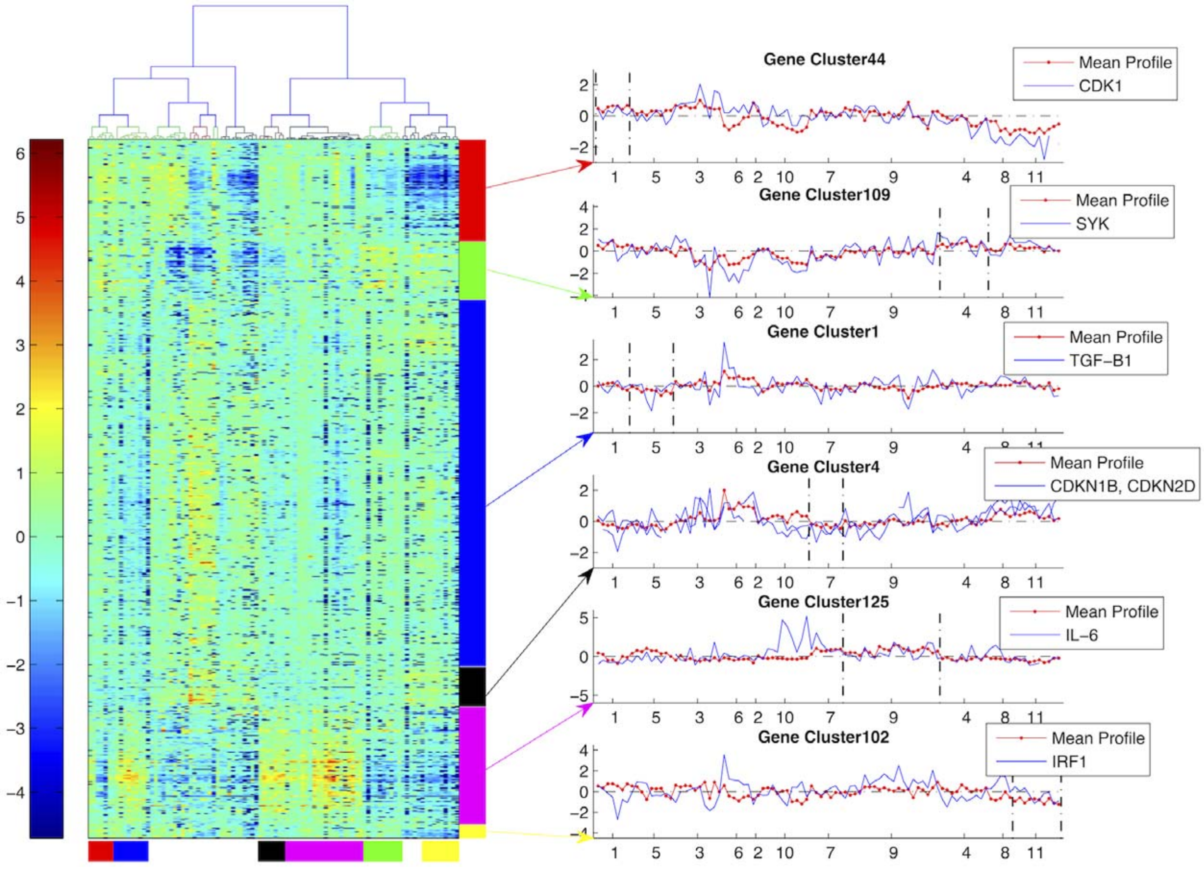


Figure 10. Expression profiles for six significant gene-clusters obtained by the DHBT method. *Left*: Heat map of gene expression profiles for the clusters of genes. Each row represents the expression profile from a clone, and each column represents a sample. The samples are organized according to the DBHT hierarchy as shown on the dendrogram on the top. Significant gene-clusters are highlighted with different colors as follows (from top to bottom, colours online): Red - gene-cluster '44' (significant for sample-cluster '1'); Green - gene-cluster '109' (significant for sample-cluster '4'); Blue - gene-cluster '1' (significant for sample-cluster '5'); Black - gene-cluster '4' (significant for sample-cluster '7'); Magenta - gene-cluster '125' (significant sample-cluster '9'); Yellow - gene-cluster '102' (significant for sample-cluster '11'). The same color scheme is used on the bottom of the heat-map to denote the corresponding sample-clusters. *Right*: Mean expression profile for each gene-cluster together with the expression profiles of note-worthy gene for each sample-cluster. The x-axes report the gene clusters. The boundaries of the relevant sample-cluster for each gene-cluster are indicated with the vertical dashed lines.
doi:10.1371/journal.pone.0031929.g010

organization in complex data-sets. This approach is based on the study of the properties of topologically embedded graphs built from a similarity measure. The DBHT technique is deterministic, it requires no a-priori parameters and it does not need any expert supervision. We have shown that the DBHT technique can successfully retrieve the clustering and hierarchical structure both from artificial data-sets and from different kinds of real data-sets outperforming in several cases other established methods. The application of the DBHT technique to a referential gene-expression dataset [44] shows that this method can be successfully used in differentiating patients with different cancer subtypes from gene-expression data. In particular, we have correctly retrieved the differentiation into distinct clusters associated with cancer subtypes (FL, CL and DLBCL) along with a meaningful hierarchical structure. The DBHT technique provides a meaningful differentiation of the DLBCL cancer samples into four distinct clusters which turn out to correspond to different survival rates. The application of the DHBT clustering technique over the gene-clones identifies new groups of genes that play a relevant role in the differentiation of the cancer subtypes, and possibly in relevant genetic pathways which control survival/proliferation of the tumor cells. Differently from [44] which indicates GCB- and ABC-like DLBCL classification under thorough supervision with biological expertise, we have found instead, in a completely un-supervised manner, four subtypes of DLBCL with different expression signatures that differentiate significantly in their genetic mechanisms and biological features resulting in well distinct survival rates, hence providing a new perspective. It should be stressed that the DBHT technique is addressing the problem of data clustering and hierarchical study from a different perspective with respect to other approaches commonly used in the literature. It therefore provides an important alternative support in a field where the sensitivity of the results to the kind of approach is often crucial. The DBHT technique can be extended to more complex measures of dependency which may be also asymmetric. In our graph theoretic approach this can be handled by constructing topologically embedded directed graphs. Another extension may concern the use of graph-embedding on surfaces of genus larger than zero that will provide more complex networks and a richer data filtering [17].

Supporting Information

Supporting Information S1 The file *PaperSupporting_ver230112_PLoSOne.pdf* contains additional information to the

manuscript explaining methods, procedures and results in further details. It consists of 16 pages, 3 tables and 11 figures.
(PDF)

Supporting Information S2 The file *DHBT_codesAndData.zip* is a compressed achieve file containing the matlab code *DBHT.m* to compute the DBHT clusters and hierarchies, this code calls 8 other functions: *BubbleCluster8.m, CliqHierarchyTree2.m, BubbleCluster8.m, clique3.m, cRand1.m, DirectHb.m, doPMFG.m, DrawPMFG.m.* The achieve also contains the code *iris_demo.m* and the data *matlab_iris_demo.mat* which can be used to reproduce Fig. 5(a). Demo code and dataset to reproduce Fig. 9 are instead: *ymphoma_demo.m, matlab_DLBCL_demo.mat.* The *ReadMe.tex* file explains code usage and installation.
(ZIP)

Acknowledgments

Many thanks Dr. Rohan Williams for helpful discussions and advices on Gene Ontology analysis and COST MP0801 project.

Author Contributions

Conceived and designed the experiments: WMS TDM TA. Performed the experiments: WMS TDM TA. Analyzed the data: WMS TDM TA. Contributed reagents/materials/analysis tools: WMS TDM TA. Wrote the paper: WMS TDM TA.

References

1. Jain A, Murty M, Flynn P (1999) Data clustering: A review. ACM Comuting Surveys 31.
2. McQueen J (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: 281–297.
3. Xu R (2005) Survey of clustering algorithms. IEEE Transactions on Neural Networks 16: 645–678.
4. Eisen M, Spellman P, Brown P, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25): 14863–14868.
5. Rocke DM, Ideker T, Troyanskaya O, Quackenbush J, Dopazo J (2009) Papers on normalization, variable selection, classification or clustering of microarray data, Editorial. Bioinformatics 25: 701–702.
6. Rivera C, Vakil R, Bader J (2010) NeMo: Network Module identification in Cytoscape. BMC Bioinformatics 11: No. Suppl 1.
7. Quackenbush J (2001) Computational analysis of microarray data. Nature Review 2: 418–427.
8. Jonsson PF, Cavanna T, Zicha D, Bates PA (2006) Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. BMC Bioinformatics 7: 2.
9. Goh KII, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci USA 104: 8685–8690.
10. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99: 7821–7826.
11. Kitsak M, Riccaboni M, Havlin S, Pammolli F, Stanley HE (2010) Scale-free models for the structure of business firm networks. Phys Rev E 81: 1–9.
12. Amaral L, Scala A, Barthelemy M, Stanley H (2000) Classes of small-world networks. Proc Natl Acad Sci 97: 11149–11152.
13. Garlaschelli D, Capocci A, Caldarelli G (2007) Self-organized network evolution coupled to extremal dynamics. Nature Physics 3: 813–817.
14. Caldarelli G (2007) Scale-Free Networks: Complex Webs in Nature and Technology. Oxford: Univesity Press.
15. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. Nature 464: 1025–1028.
16. Hooyberghs H, Van Schaeybroeck B, Moreira A, Andrade J, Herrmann H, et al. (2010) Biased percolation on scale-free networks. Phys Rev E 81: 011102.
17. Aste T, Di Matteo T, Hyde S (2005) Complex networks on hyperbolic surfaces. Physica A 346: 20–26.
18. Tumminello M, Aste T, Di Matteo T, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci USA 102: 10421–10426.
19. Ringel G (1974) Map Color Theorem. Springer-Verlag, Berlin.
20. Andrade JSJ, Herrmann HJ, Andrade RF, da Silva LR (2005) Apollonian networks: Simultaneously scale-free, small world, euclidan, space filling and matching graphs. Phys Rev Lett 94: 1–4.
21. Di Matteo T, Aste T, Hyde S (2004) Exchanges in complex networks: Income and wealth distributions. In: F. Mallamace, HE. Stanley, eds. Physics of complex systems (new advances and perspectives). volume 155 of *Proceedings of the international school of physics Enrico Fermi* 435–442. International School of Physics Enrico Fermi on the Physics of Complex Systems - New Advances and Perspectives, Varenna, ITALY, JUL 01-11, 2003.
22. Di Matteo T, Aste T, Gallegati M (2005) Innovation flow through social networks: productivity distribution in France and Italy. Eur Phys J B 47: 459–466.
23. Pellegrini GL, de Arcangelis L, Hermann HJ, Perrone-Capano C (2007) Activity-dependent neural network model on scale-free netowkrs. Phys Rev E 76: 016107.
24. Di Matteo T, Pozzi F, Aste T (2010) The use of dynamical networks to detect the hierarchical organization of financial market sectors. Eur Phys J B 73: 3–11.
25. Diestel R (2005) Graph Theory ed. 3. Springer-Verlag.
26. Song WM, Di Matteo T, Aste T (2011) Nested hierarchies in planar graphs. Discrete Applied Mathematics 159: 2135–2146.
27. Sorensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biologiske Skrifter 5: 1–34.
28. Boyer JM, Myrvold WJ (2004) On the cutting edge: Simplified o(n) planarity by edge addition. Journal of Graph Algorithms and Applications 8: 2004.
29. Arthur D, Vassilvitskii S (2007) k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp 1027–1035.
30. Shi J, Malik J (2000) Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8): 888–905.
31. von Luxburg U (2007) A tutorial on spectral clustering. Technical report, Max-Planck-Institut für biologische Kybernetik.
32. Kohonen T, Schroeder MR, Huang TS (2001) editors. *Self-Organizing Maps.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition.
33. Ruan J, Dean A, Zhang W (2010) A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Systems Biology 4(1): 8+.
34. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193–218.
35. Hernádvölgyi IT (1998) Generating random vectors from the multivariate normal distribution. Technical Report TR-98-07, University of Ottawa.
36. Shaun S. Wang (2004) Casualty Actuarial Society Proc. Vol. LXXXV; and Available: http://www.mathworks.com/matlabcentral/fileexchange/6426. Accessed 2012 Jan 20.
37. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci USA 104(1): 36–41.
38. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals Eugen 7: 179–188.
39. UCI Machine Learning Repository. Iris data. Available: http://archive.ics.uci.edu/ml/datasets/Iris. Accessed 2012 Jan 20.
40. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A (2008) Clustering cancer gene expression data: a comparative study. BMC Bioinformatics 9: 1–14.
41. Dunn J (1974) Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics 4: 95–104.
42. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Machine Intell 1: 2224–227.
43. Handl J, Knowles J, Kell DB (2005) Computational cluster validation in post-genomic data analysis. Bioinformatics 21(15): 3201–3212.
44. Alizadeh A, Eisen M, Davis R, Ma C, Lossos I, et al. (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403: 503–511.
45. Wang J, Delabie J, Aasheim HC, Smeland E, Myklebost O (2002) Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics 3(1): 36.
46. Abramson JS, Shipp MA (2005) Advanced in the biology and therapy of diffuse large b-cell lymphoma: moving toward a molecularly targeted approach. Blood 106: 1164–1174.
47. Lenz G, Wright GW, Emre NCT, Kohlhammer H, Dave SS, et al. (2008) Molecular subtypes of diffuse large b-cell lymphoma arise by distinct genetic pathways. Proc Natl Acad Sci USA 105(36): 13520–13525.
48. Wada N, Kohara M, Ogawa H, Sugiyama H, Fukuhara S, et al. (2009) Change of cd20 expression in diffuse large b-cell lymphoma treated with rituximab, and anti-cd20 monoclonal antibody: A study of the osaka lymphoma study group. Case Rep Oncol 2(3): 194–202.
49. Nathalie AJ, Boyle M, Bashashati A, Leach S, Brooks-Wilson A, et al. (2009) Diffuse large b-cell lymphoma: reduced cd20 expression is associated with an inferior survival. Blood 113.
50. Zhao X, Lapalombella R, Joshi T, Cheney C, Gowda A, et al. (2007) Targeting cd37-positive lymphoid malignancies with a novel engineered small modular immunopharmaceutical. Blood 110(7): 2569–2577.

51. Filipits M, Jaeger U, Pohl G, Stranzl T, Simonitsch I, et al. (2002) Cyclin d3 is a predictive and prognostic factor in diffuse large b-cell lymphoma. Clinical Cancer Research 8(3): 729–733.
52. Chen L, Monti S, Juszczynski P, Daley J, Chen W, et al. (2008) Syk-dependent tonic b-cell receptor signaling is a rational treatment target in diffuse large b-cell lymphoma. Blood 111(4): 2230–2237.
53. Lossos IS, Alizadeh AA, Diehn M, Warnke R, Thorstenson Y, et al. (2002) Transformation of follicular lymphoma to diffuse large-cell lymphoma: Alternative patterns with increased or decreased expression of c-myc and its regulated genes. Proc Natl Acad Sci USA 99(13): 8886–8891.
54. Coffey GP, Rajapaksa R, Liu R, Sharpe O, Kuo C-C, et al. (2009) Engagement of cd81 induces ezrin tyrosine phosphorylation and its cellular redistribution with filamentous actin. Journal of Cell Science 122(17): 3137–3144.
55. Lam LL, Wright G, Davis RE, Lenz G, Farinha P, et al. (2008) Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor- pathways in subtypes of diffuse large b-cell lymphoma. Blood 111(7): 3701–3713.
56. BiNGO Available: http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html. Accessed 2012 Jan 20.
57. Zhao XF, Gartenhaus RB (2009) Phospho-p70s6k and cdc2/cdk1 as therapeutic targets for diffuse large b-cell lymphoma. Expert Opinion on Therapeutic Targets 13(9): 1085–1093.
58. Leseux L, Hamdi SM, al Saati T, Capilla F, Recher C, et al. (2006) Syk-dependent mtor activation in follicular lymphoma cells. Blood 108(13): 4156–4162.
59. Arsura M, Wu M, Sonenshein GE (1996) TGF-β1 inhibits NF-κb/rel activity inducing apoptosis of B cells: Transcriptional activation of iκbα. Immunity 5(1): 31–40.
60. Kamijo T, Zindy F, Roussel MF, Quelle Dawn E, Downing James R, et al. (1997) Tumor Suppression at the Mouse INK4a Locus Mediated by the Alternative Reading Frame Product p19 ARF. Cell 91(5): 649–659.
61. Seki R, Okamura T, Koga H, Yakushiji K, Hashiguchi M, et al. (2003) Prognostic significance of the f-box protein skp2 expression in diffuse large b-cell lymphoma. American Journal of Hematology 73(4): 230–235.
62. Saez AI, Saez AJ, Artiga MJ, Perez-Rosado A, Camacho F-I, et al. (2004) Building an outcome predictor model for diffuse large b-cell lymphoma. Am J Pathol 164(2): 613–622.
63. Ding BB, Yu JJ, Yu RY, Mendez LM, Shaknovich R, et al. (2008) Constitutively activated stat3 promotes cell proliferation and survival in the activated b-cell subtype of diffuse large b-cell lymphomas. Blood 111(3): 1515–1523.
64. Romeo G, Fiorucci G, Chiantore MV, Percario ZA, Vannucchi S, et al. (2002) Irf-1 as a negative regulator of cell proliferation. Journal of Interferon and Cytokine Research 22(1): 39–47.