Overview

Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis

David A. Morrison*

Exploratory data analysis (EDA) involving both graphical displays and numerical summaries of data, is intended to evaluate the characteristics of the data as well as providing a form of data mining. For multivariate data, the best-known visual summaries include discriminant analysis, ordination, and clustering, particularly metric ordinations such as principal components analysis. However, these techniques have limiting mathematical assumptions that are not always realistic. Recently, network techniques have been developed in the biological field of phylogenetics that address some of these limitations. They are now widely used in biology under the name phylogenetic networks, but they are actually of general applicability to any multivariate dataset. Phylogenetic networks are fast and relatively easy to calculate, which makes them ideal as a tool for EDA. This review provides an overview of the field, with particular reference to the use of what are called splits graphs. There are several types of splits graph, which summarize the multivariate data in different ways. Example analyses are presented based on the neighbor-net graph, which seems to be the most generally useful of the available algorithms. This should encourage the more widespread use of these networks whenever a summary of a multivariate dataset is required. © 2014 John Wiley & Sons, Ltd.

How to cite this article:
WIREs Data Mining Knowl Discov 2014, 4:296–312. doi: 10.1002/widm.1130

INTRODUCTION

Data mining is the analysis step of knowledge discovery, and the oldest form of data mining is known as exploratory data analysis (EDA). EDA has traditionally been used to evaluate the characteristics of a dataset, using graphical displays as well as numerical summaries. Originally, it involved mainly statistical techniques (e.g., counting, cross-tabulation, regression), but was later developed to include neighborhood methods (e.g., nearest-neighbor algorithms, ordination) and clustering (e.g., hierarchical clustering, *k*-means clustering). Data mining has now expanded beyond these classical techniques, of course, to include next-generation techniques such as decision trees, artificial neural networks, and rule induction. Nevertheless, there is ongoing development in the classical area as well, as discussed in this overview.

For multivariate data, in which many characteristics (or descriptors) have been measured for each sample object (i.e., there are no missing data), the EDA summaries usually apply to the relationships between the objects. This is sometimes called Q-mode analysis; the alternative R-mode analysis looks at relationships among the descriptors. Here, I am restricting myself to Q-mode analysis, in which the complex relationships (potentially one or each characteristic) are reduced to some manageable subset of the most important ones (variously defined). The best-known visual summaries of multivariate relationships, as used in EDA, include both ordination and clustering.

Ordination embraces methods such as factor analysis, principal components analysis,

*Correspondence to: David.Morrison@ebc.uu.se

Department of Biomedical Sciences and Veterinary Public Health, Swedish University of Agricultural Sciences, Uppsala, Sweden

Conflict of interest: The author has declared no conflicts of interest for this article.

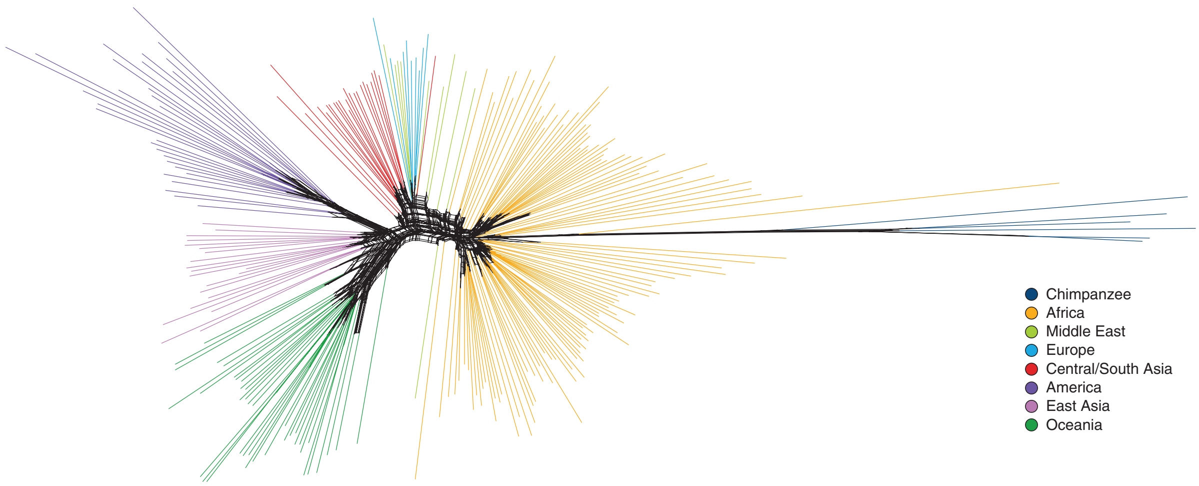


FIGURE 1 | A phylogenetic network of the genetic relationships (measured using microsatellite data) among 255 populations of humans and chimpanzees. The edges leading to the leaf nodes (which are unlabeled) are color-coded by their source. The network was constructed using the neighbor-net algorithm, with the similarity measured as the proportion of shared alleles. (The data are available from Ref 1.)

correspondence analysis, and nonmetric multidimensional scaling, while clustering comprises a multitude of hierarchical algorithms (both agglomerative and divisive), as well as nonhierarchical techniques such as k-means clustering and fuzzy clustering. Here, I provide an introduction to a new class of visual multivariate summaries called *phylogenetic networks*, which in some ways owes a debt to both clustering and ordination. An example is shown in Figure 1 (it is discussed further below).

Algorithmically, phylogenetic networks are a type of agglomerative hierarchical clustering, but in terms of their objective they have much in common with fuzzy clustering, in the sense that each object can simultaneously be a member of one to many clusters. These networks can be used for various of the common purposes of data mining, but most notably: (1) the automatic extraction of previously unknown patterns with regard to groups of objects, without using known structures in the data; (2) the detection of anomalous objects in the dataset; and (3) providing a compact representation of the dataset, which can be easily visualized as a connected graph.

CLUSTERS AND NETWORKS

Phylogenetic networks are so named because they were originally developed in the biological field of phylogenetics, starting in the early 1990s. They are now widely used in biology, but they are of general applicability to any multivariate dataset. Phylogenetic networks are fast and relatively easy to determine, which makes them ideal as a tool for EDA.[2]

Phylogenetic techniques, including phylogenetic trees and networks, are starting to be used outside of biology, as their more general applicability becomes more widely known. This has been particularly so in anthropology,[3–5] including the study of languages,[6] written texts,[7] folk tales,[8] and cultural artifacts.[9] There is much wider applicability, as well, as I am arguing here.

The Different Types of Network

Mathematically, networks are connected graphs, with nodes (representing the objects) connected by edges (representing some form of relationship). The edges may be undirected or directed, with the direction indicating some sort of asymmetrical relationship between the objects. There are now many types of empirical graphs called 'networks', and therefore it is important at the outset to place phylogenetic networks clearly within this context.

Most networks are directly connected networks based on empirical observations. In these networks, labeled nodes are connected directly to each other, and the edges (or arcs, if they have a specific direction) represent some sort of observed connection between the nodes. For example, in biology[10] the organisms within a local population may be genetically related (e.g., parent to offspring), and this can be represented by a directly connected network,[11] although not all of the organisms need to be connected to each other (i.e., the network is not necessarily fully connected). Alternatively, a protein interaction network has proteins as the nodes, and the edges represent pairwise interactions, although there is no specific direction to these interactions. Some of the edges in these types of

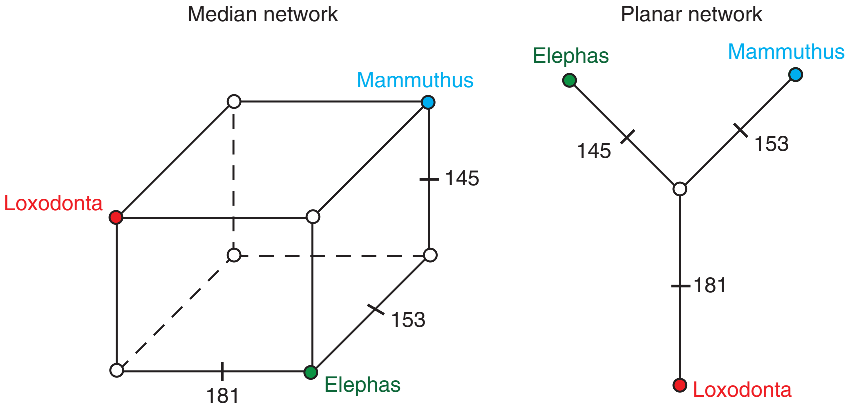


FIGURE 2 | Two phylogenetic networks of the same dataset concerning the genotypes of three species. There are three observed nodes (labeled by the species name) in both cases (filled circles), but five and one inferred nodes (open circles), respectively. The edges represent genetic similarity between the species; the numbers count the observed character differences between them. (The data are available from Ref 14.)

network may be inferred, which will then involve a model-based network inference procedure,[12] although the labeled nodes usually remain observed rather than inferred. In artificial neural networks, the inferred edges connect the observed input and output nodes via (unlabeled) hidden nodes.

Phylogenetic networks differ from these types of networks,[13] as shown in Figure 2. These graphs have labeled leaf nodes representing the objects, internal nodes that are usually unlabeled, and edges connecting all the nodes (i.e., the network is fully connected). The internal nodes are inferred rather than observed, as also are all of the edges. However, the inferred nodes are not hidden, as they are in an artificial neural network. In phylogenetic analysis, two types of graphs have been developed, although only one of them is relevant to this overview: (1) rooted evolutionary networks, in which the internal nodes represent ancestors of the leaf nodes, and the directed edges represent pathways of inferred historical relationship; and (2) unrooted affinity networks, in which the internal nodes do not represent ancestors, and the undirected edges represent similarity relationships among the leaf nodes. In this overview, I am specifically concerned with the second type, the similarity diagrams (in the algorithmic literature, these have sometimes been called ‘implicit phylogenetic networks’).

Relationship to Other Multivariate Methods

Multivariate datasets consist of a set of objects (real or imagined) that have all been measured for the same characteristics. Therefore, Q-mode analyses assume that we are comparing like with like across the objects. What creates that likeness is important for the interpretation of the analysis outcome, but is not important for the mathematical algorithm itself. In phylogenetics, for example, the likeness is often assumed to have been created by evolutionary homology, and the networks are interpreted in that light; but this assumption is not necessary for other datasets. The cause of the likeness may well vary from dataset to dataset, and this is important for interpretation, but not otherwise, provided that the dataset does, indeed, compare like with like within each characteristic.

Multivariate data-summary analyses try to discover structure in the dataset without using any previously known structures in those data. Of the main data-summary techniques, ordination creates a scatter plot with each object as a point on the graph. The relative distance between the points on the graph summarizes the relationships among the objects. That is, the ordination summarizes the multi-dimensional neighborhoods of the different objects. This is useful for showing continuous relationships, for example, or showing several relationships simultaneously (e.g., one for each axis of the graph).

On the other hand, clustering tries to discover groups in the data that are in some way similar. These clusters may be hierarchically arranged (each group nested within another group), in which case a tree is used for display (i.e., a connected graph). This tree has a root, so that the edges are directed away from the root. If the clusters are not hierarchically arranged, then the boundaries between groups can be ‘fuzzy’, so that each object can simultaneously be a member of more than one cluster (associated with a different probability for each cluster). The clusters can be formed by agglomeration, where the objects are sequentially added to a cluster, or division, where the dataset is progressively divided into smaller clusters. Clustering is useful for showing discrete relationships, but it essentially only displays one relationship (the one that relates to the grouping).

Phylogenetic networks (usually) proceed in the same manner as an agglomerative hierarchical clustering analysis, except that each object is allowed to simultaneously be a member of many clusters (if necessary). Instead of using a tree for display, a network is used, which is basically a tree with reticulations.

However, the network is unrooted, so that the edges are undirected. Like hierarchical clustering, if there are reasonably well-defined groups in the data then the analysis will detect them, because each object will be in only one group (or a small number of groups with very unequal probabilities). Like fuzzy clustering, if there are no strong cluster patterns then each object will be equally part of many groups. Like ordination, the relative position of the objects in the graph represents their neighborhood, so that several relationships can be displayed simultaneously.

A phylogenetic network is a fully connected graph, with labeled leaf nodes representing the observed objects, unlabeled internal nodes that are inferred by the analysis, and undirected edges connecting the nodes, representing the inferred relationships between those nodes. The length of the edges represents the amount of support in the data for the relationship indicated by that edge. That is, objects that are closely connected in the network are similar to each other based on the observed characteristics, and those that are further apart are progressively more different from each other.

For the example network shown in Figure 1, there is a large collection of internal nodes (in black) with apparently little hierarchical structure, and yet the leaf nodes (in color) do clearly show a degree of large-scale geographical clustering (i.e., the different colors are mostly grouped in unique places on the graph). This indicates that the human genotypes are similar to each other within each geographical area, although there is little distinction between the Middle Eastern (light green) and European (blue) genotypes.

As with any multivariate data-display technique, the phylogenetic network algorithm involves only the visualization of the data (i.e., constructing the graph). There is one prior step that should be considered: whether the descriptors need to be adjusted to make them directly comparable, such as by normalization (e.g., to a standard normal), standardization (e.g., range standardized), or transformation (e.g., log transformed). This decision will be made on a case by case basis, and it can determine which characteristics of the data dominate the summary.

For phylogenetic networks, a summary of the character data can be displayed directly (as shown, for example, in Figure 2). However, for complex data this is unwieldy, as it becomes impossible to display the network in only two or three dimensions (i.e., on a piece of paper or computer screen). Under these circumstances it is possible to visualize the data by first calculating an association coefficient (e.g., resemblance, similarity) or distance (e.g., difference, dissimilarity) among pairs of objects based on the characteristics (as was done, for example, in Figure 1). This similarity or distance summarizes the relationships among the descriptors, and it is this summary that is then displayed in the graph.

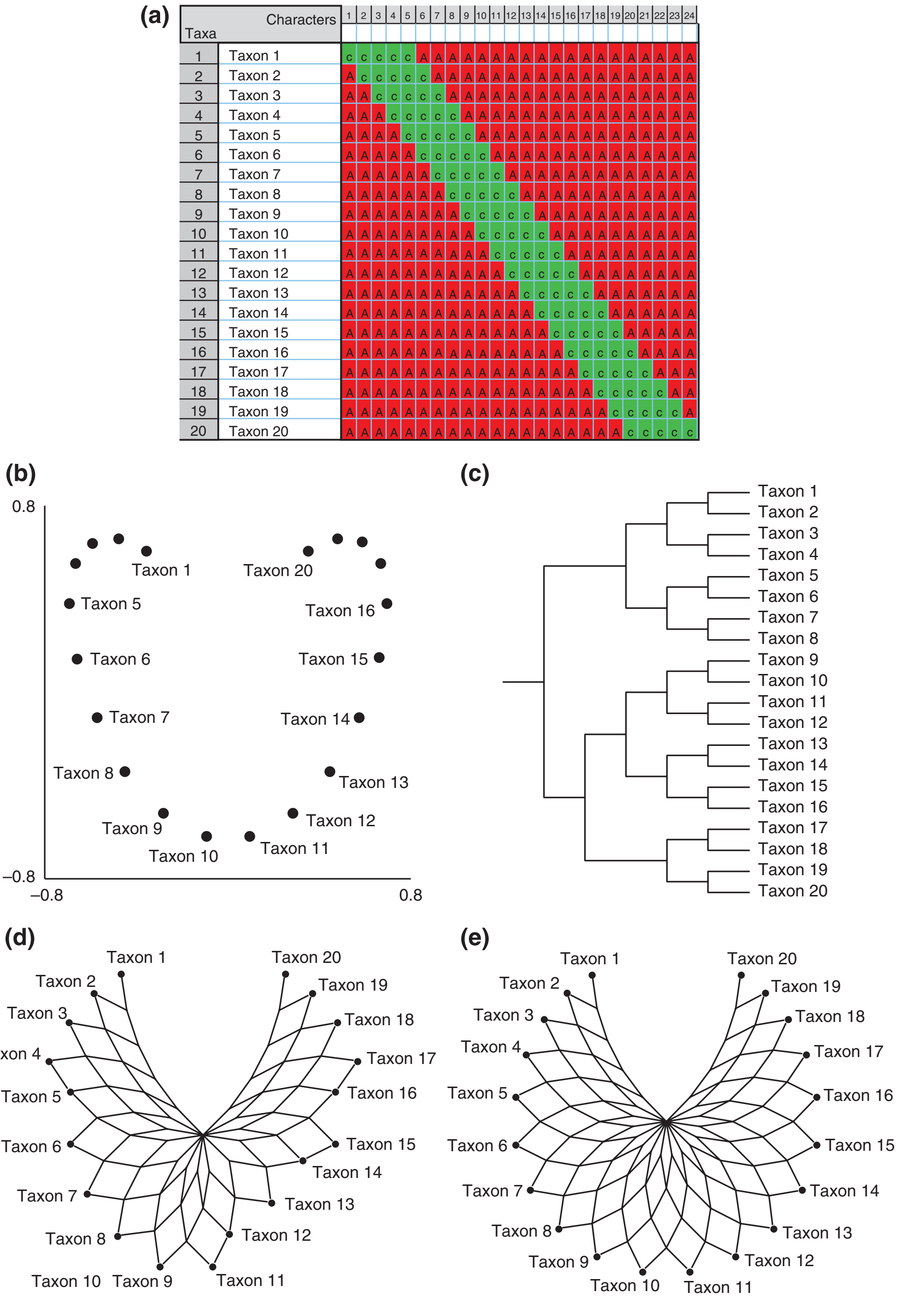
The simplest distance is the Hamming distance, which is a count of the number of characteristics that differ between the pairs of objects, followed by the Manhattan distance, which is the sum of the differences in attribute values between the pairs of objects. These are often the best distances to use for data mining, as they impose very little structure on the data. However, far more sophisticated distances have also been developed, many of which are quite general but some of which have been devised for specific purposes in different disciplines. The choice of distance (or similarity) needs to be made on a case by case basis.

There can be no missing data if the characters are to be displayed directly in the network. If distances are being used, then a distance measure is required that correctly accounts for the missing data; otherwise, there will be bias in the resulting network.

Limitations of Current Multivariate Summaries

Ordination does a good job of displaying multivariate neighborhoods, although in practice it is limited to displaying the scatterplot using only two or three dimensions. However, it has long been recognized that, in spite of its popularity, there can be serious distortions in the graphical display.[15–17] In particular, ordinations based on eigenanalysis (such as principal components analysis, correspondence analysis, and principal coordinates analysis) try to display the graph using one more dimension than actually exists in the data. For example, if there is a single dominant gradient among the objects in the dataset, which could therefore be displayed using only one graph dimension, then in practice this will be spread over two dimensions in the graph, instead, forming what has been called a horseshoe or arch (see Figure 3(b)). This artifact is seen quite frequently in the empirical literature.

Hierarchical clustering does a good job of displaying clustered relationships, although it effectively displays only one set of relationships in its connected graph. However, it has the limitation of exclusivity, where each object is forced into one group only. Nonhierarchical clustering can avoid this limitation by recognizing over-lapping fuzzy clusters, where group membership is probabilistic. However, this then exhibits the limitation that it does not clearly show the relationships between the clusters, especially if they do not overlap.



	Characters Taxa	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Taxon 1	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
2	Taxon 2	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
3	Taxon 3	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
4	Taxon 4	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
5	Taxon 5	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
6	Taxon 6	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A	A
7	Taxon 7	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A	A
8	Taxon 8	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A	A
9	Taxon 9	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A	A
10	Taxon 10	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A	A
11	Taxon 11	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A	A
12	Taxon 12	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A	A
13	Taxon 13	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A	A
14	Taxon 14	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A	A
15	Taxon 15	A	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A	A
16	Taxon 16	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A	A
17	Taxon 17	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A	A
18	Taxon 18	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A	A
19	Taxon 19	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c	A
20	Taxon 20	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	c	c	c	c	c

FIGURE 3 | The effect of a single-gradient dataset on multivariate data summaries. (a) The dataset, with 20 objects (Taxon 1–Taxon 20) and 24 characters, each of which has two possible states (A or C). (b) The principal components ordination of the data (not all of the objects are labeled). (c) The UPGMA hierarchical clustering of the data. (d) The median network analysis of the data. (e) The neighbor-net analysis of the data.

Phylogenetic networks try to balance these various strengths and weaknesses.[18] Such a network can display clusters (as for hierarchical clustering), if there are reasonably distinct groupings of the objects, but otherwise it will display neighborhoods (as for ordination), if there are more-or-less continuous relationships. It can simultaneously display alternative clusterings (as for fuzzy clustering), if there are several contradictory patterns in the data. Thus, in many ways, phylogenetic networks are a compromise between the three alternative multivariate data analyses. They have considerably fewer restricting assumptions than do the other data summaries, and so they can adapt themselves to different data patterns.

As far as the dimensional distortions are concerned, phylogenetic networks have not yet been thoroughly evaluated. However, it is known that different techniques respond in different ways to the multivariate patterns. For instance, the neighbor-net algorithm seems to be very effective for displaying one-dimensional data patterns, while median networks are more effective for two-dimensional patterns. As an example, Figure 3(a) shows a simple dataset with one-dimensional patterns, where the data patterns at the two ends of the gradient have nothing in common. Figure 3(b) shows the horseshoe effect produced by the principal components analysis, which displays the gradient using two graphical dimensions instead of one, falsely giving the impression that Taxon 1 and Taxon 20 are similar to each other. Furthermore, the dots are not equally spaced, which they should be given the continuous pattern in the original data. Figure 3(c) shows that the hierarchical clustering analysis groups the objects in a somewhat arbitrary manner, first aggregating them in adjacent pairs, but failing to maintain a symmetrical hierarchy with further aggregation, as would be required for a gradient. Figure 3(d) shows the median network, which correctly associates adjacent objects while separating the two ends of the gradient. Unfortunately, it fails to retain the symmetry of the gradient, by superimposing two of the objects in the network (Taxon 9 and Taxon 10). Figure 3(e) shows the neighbor-net analysis, which correctly displays the data as a single gradient. However, the latter network is much more complex than is necessary, since the dataset could be summarized graphically by arranging the objects equally spaced along a single straight line.

PHYLOGENETIC NETWORKS

There are many types of phylogenetic network, including splits graphs, parsimony networks, and reticulograms.[19,20] These are based on somewhat different mathematical criteria,[21] but in practice the algorithms will often produce similar networks for any given dataset. For the purposes of data mining, it seems that the family of methods that produce splits graphs are the most promising; and so I will focus on them in this overview. Thus, all the examples shown here are different forms of splits graph.

Splits Graphs and Their Interpretation

A splits graph[22–24] is actually a separation network, in the sense that the edges separate the objects into groups rather than connecting them together. In a splits graph, each edge represents a bipartition of the objects based on one or more characteristics. That is, each edge splits the graph into two. This is a straightforward generalization of a tree (as used in hierarchical clustering), as each edge in a tree also represents a bipartition of the objects. If an edge in a tree is ‘cut’ (i.e., removed) then the two resulting parts of the graph connect the objects forming each of the two partitions (i.e., nonoverlapping subsets). Each bipartition will be supported (in some way) by the differing characteristics of the objects.

A splits graph shows the bipartitions that are supported by the dataset, and only those bipartitions. If there is no conflict in the data then each bipartition is represented by a single edge in the graph; and if there are contradictory patterns then the each bipartition is represented by a set of parallel edges. Internal nodes appear whenever edges or sets of edges intersect. The edge lengths represent the relative amount of support among the descriptors in the whole dataset for each of the splits.

A simple example of how to interpret a splits graph is shown in Figure 4. The data to be summarized concern the five internationally released compilation albums of the musical duo of Paul Simon & Art Garfunkel: *Simon & Garfunkel’s Greatest Hits* (1972); *The Simon & Garfunkel Collection* (1981); *The Concert in Central Park* (1982); *The Definitive Simon & Garfunkel* (1992); and *The Essential Simon & Garfunkel* (2003). The data are the best-selling chart positions of these albums in eight different countries. The objective is to summarize the similarities among the countries with respect to how popular these albums have been.

Figure 4(a) shows the neighbor-net analysis of the data, which in this case is based on the Manhattan distance. Note that the graph is to be interpreted as showing relationships only along the edges of the network. The graph is not very tree-like, indicating that there are a number of incompatible patterns in the dataset. However, the strongest patterns indicate four weak clusters in the network: Sweden, New Zealand + France, the Netherlands + Japan, and Germany + the United Kingdom + Australia.

The remaining four parts of the figure highlight some of the features of a splits graph. Figure 4(b) illustrates a simple split of the data for which there is no contradiction among the descriptors, so that it requires only a single edge in the network. Indeed, in this dataset each country has an edge of its

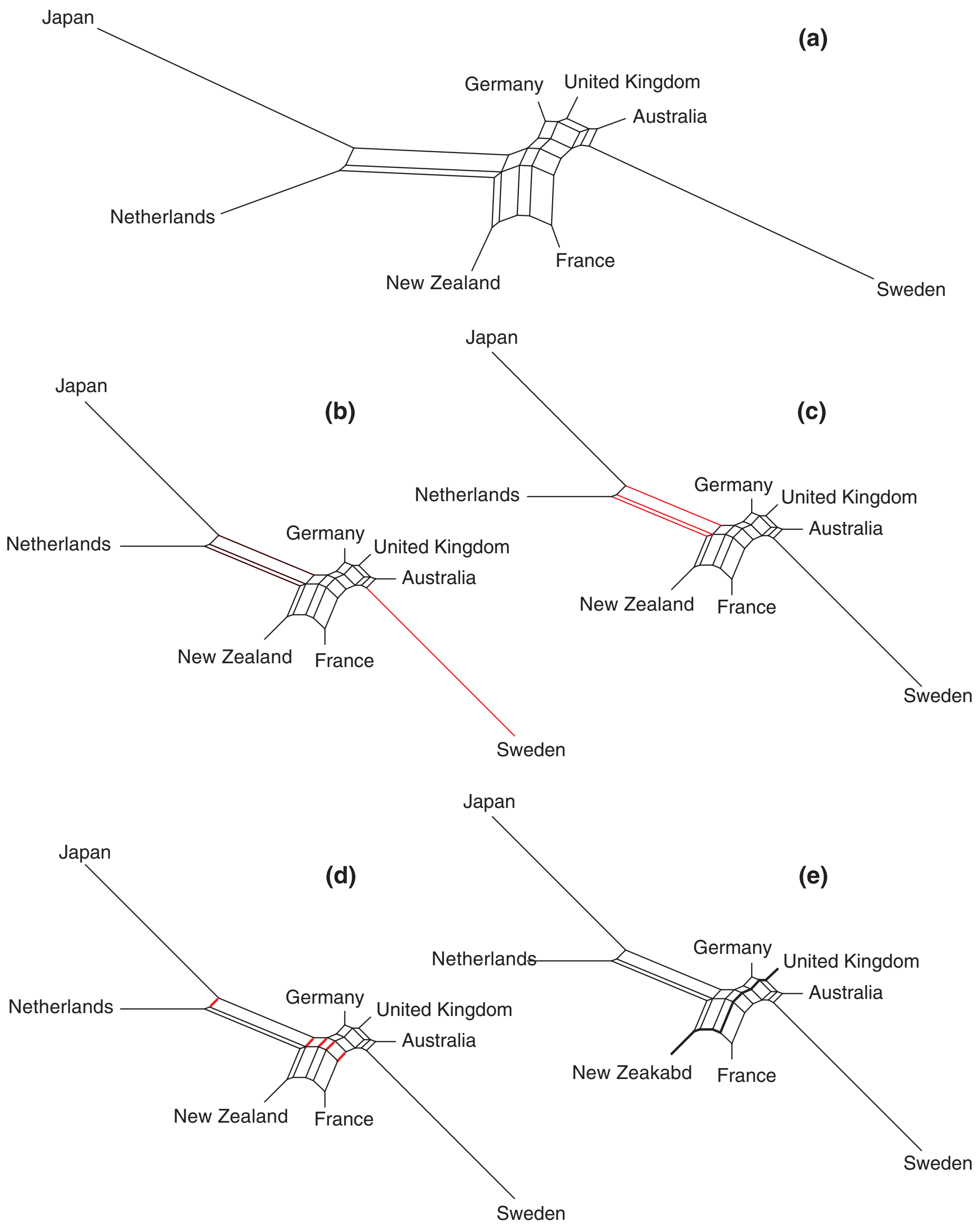


FIGURE 4 | Phylogenetic network of the chart positions of five Simon & Garfunkel albums in eight countries. (a) The neighbor-net analysis based on the Manhattan distance. (b) The split separating Sweden from the other countries. (c) The split (in red) separating Japan + Netherlands from the other countries. (d) The split (in red) separating Netherlands + New Zealand + France from the other countries. (e) The network distance (in bold) separating New Zealand from the United Kingdom. (The data are available from Ref 25.)

own, representing its unique characteristics. However, Figure 4(c) and (d) shows two splits that are incompatible with each other. In Figure 4(c) the Netherlands is shown clustered with Japan while in Figure 4(d) the Netherlands is shown clustered with New Zealand + France. Both clusters can be displayed in the network, but this requires each of these splits to be represented by a set of parallel edges rather than a single edge each. Figure 4(e) illustrates the way in which the edge lengths relate to the original distance or character data. The shortest pathlength between objects (i.e., the sum of the edge lengths separating them) should represent the distance in the original dataset. Note that in a network there are many possible shortest paths between each pair, unlike in a tree where there is only one shortest path.

Types of Splits Graph

Conceptually, the simplest form of splits graph is a median network.[26,27] A median network is a splits graph that displays *all* of the splits present in the dataset, based on the descriptors themselves. That is, an edge is added to the network for every split formed by every character. If two (or more) characters form the same split, then the relevant edge length is increased, instead. An example is shown in Figure 2 (left), where three splits are represented by three sets of parallel edges of differing length.

A median network can be produced only from binary (two-state) data (e.g., presence–absence). The equivalent for multi-state data is called a quasi-median network.[28–30] Strictly speaking, the latter is not a splits graph because triangles can appear instead of parallelograms, and splits graphs contain only parallelograms and/or single edges.

The basic limitation of a median network is that increasing complexity of the patterns in a dataset leads directly to increasing complexity in the splits graph. Real data are often far too complex to be displayed in a median network, as the graph can produce undisplayable hypercubes (and beyond a cube the diagram is uninterpretable, anyway). So, the family of splits-graph methods is formed by various approaches to simplifying the network representation of the data, thus forming subgraphs of the full median network.

For example, reduced-median networks,[31] greedily reduced-median networks,[27] local Buneman graphs,[32] and quartet window analysis[33] all attempt to remove some of the incompatibilities in the data during the construction of the median network. Median-joining networks[28] do the same thing to form a subgraph of the full quasi-median network. Alternatively, pruned median networks[34] and pruned quasi-median networks[35] attempt to thin the network after it has been created (i.e., removing 'unwanted' edges).

Another approach has been to display only what are called weakly compatible splits, which also form a subgraph of the median-network hypercubes. These can usually be displayed in a plane without edge crossings (i.e., can be presented in two dimensions), as illustrated in Figure 2 (right). If the algorithm uses the character data then it is called parsimony splits,[36] while it is called split decomposition[37] if it is based on a distance measure. Unfortunately, increasing complexity of the patterns in a dataset often leads these two methods to fail, when they produce an unresolved graph (called a star tree) instead of an informative network.

The neighbor-net method[38,39] was developed as a compromise between the hypercubes produced by median networks and the unresolved networks produced by split decomposition. It does this by producing what are called circular splits, which will always be planar (i.e., can be displayed in two dimensions) without being unresolved. It is based solely on distances (not the character data). The graphs in Figures 1, 3, and 4 were all produced by this algorithm.

It is also possible to construct splits graphs from a collection of trees (i.e., the output of a hierarchical clustering algorithm). Consensus networks[40–42] include in the network all of the splits that occur in all of the trees, or (if preferred) only those splits occurring in some threshold percentage of the trees. Super-networks[43,44] are similar but require only that the trees have overlapping subsets of the objects, rather than requiring all of the trees to have the same set of objects.

In the empirical biological literature, the reduced-median and median-joining networks are the most popular for population-level studies, while the neighbor-net algorithm is the most popular for species-level studies. For more general data mining, outside of biology, neighbor-net seems to offer the greatest possibilities.

There are two computer programs that are most commonly used for producing splits graphs as phylogenetic networks: SplitsTree[45,46] (http://www.splitstree.org/), and Network[47] (http://www.fluxus-engineering.com/sharenet.htm).

EMPIRICAL EXAMPLES

In this section, I provide six examples of the use of phylogenetic networks to summarize multivariate data. There is a wide range of examples, which illustrate the diversity of application of these networks, as well as different possible forms of analysis (i.e., different combinations of standardization, choice of similarity measure, and network algorithm). All of the networks were produced with the SplitsTree v.4.11.3 program, except for Figure 8 which was produced with Network v.4.6.1.1.

Type 1 Interferons

The first example is a biological one, although it is a somewhat unusual use of phylogenetic networks, as they were originally intended to be used for studying the relationships among species rather than among protein classes.

It concerns the DNA sequences of 231 type-I interferon genes, taken from the genomes of various mammal species. Interferons are proteins that function as part of a mammal's immune system, helping to

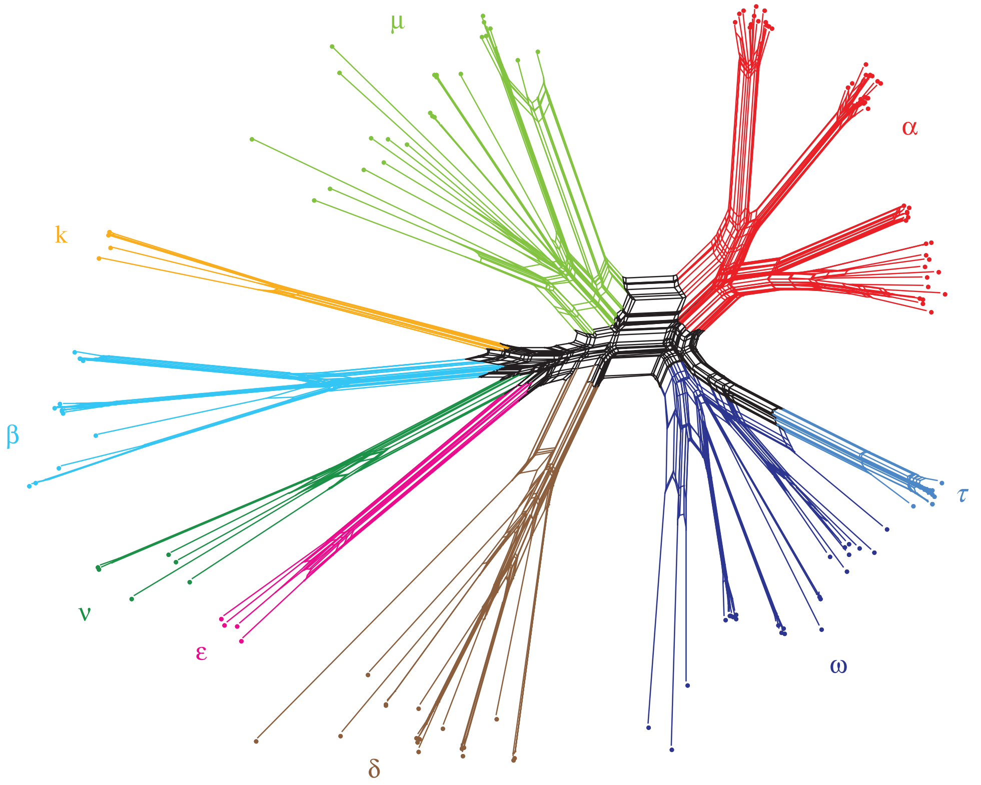


FIGURE 5 | A phylogenetic network of the amino acid sequences of type-I interferon of various mammal species. The nine different recognized subtypes are labeled (with Greek letters), although the individual samples are not labeled, and the relevant part of the network is color-coded for each subtype. The network was constructed using the neighbor-net algorithm, with the Hamming distance. (The data are available from Ref 48.)

protect the animals from disease-causing pathogens. These genes code for at least nine different classes (or subtypes) within the type-I interferon family, each class having its own molecular function within the animals' immune systems. Each mammal species has one or more gene copies of each of these type-I interferon classes (sometimes plus some pseudogenes). So, in the data analysis each object (or leaf node in the network) represents one gene copy, with the characteristic descriptors being the amino acids of their aligned DNA sequences.

The phylogenetic network of these data (Figure 5) shows that these nine interferon classes cluster quite clearly. That is, while each mammal species has its own unique copies of the genes, the protein classes are still recognizably distinct—for example, a β gene is still recognizably a β gene no matter which species it is in. Nevertheless, some of these subtypes have considerable within-class variation in their protein sequences, indicating the variability that exists both within and between the different mammal species.

The relationships between the clusters are indistinct in most cases, but there is nevertheless a well-supported bipartition associated with each cluster (as indicated by the different colors). The only exception concerns the splits separating the ω and τ classes, which are not as large as that for the other classes. Indeed, it is usually considered that, biologically, the τ class is actually a subset of the ω class.

So, this is an example where the phylogenetic network indicates a strongly clustered pattern within the dataset. This was important in the published data analysis in relation to the class labeled μ. This had previously been recognized as an indeterminate subtype provisionally called $\alpha\omega$, but the network analysis shows that it is quite distinct from both subtype α and subtype ω, and thus is worthy of recognition in its own right. The data mining in this case thus involved an important piece of knowledge discovery.

FIFA World Cup Soccer

We can now move on to the uses of phylogenetic networks outside biology.

The Fédération Internationale de Football Association (FIFA) World Cup™ soccer competition has been held every 4 years since 1930 (except 1942 and 1946). The finals competition is reported to be the most widely viewed sporting event in the world, surpassing even the Olympic Games. The number of national teams accepted into the finals has varied from 13 to the current 32; and at the end of the finals FIFA provides an ordering of these teams based on their success in the finals series. In the analysis presented here, each national team is an object, and the 19 characteristics are the ranking results of each World Cup competition.

The FIFA rank-order data for each Cup were range-scaled to vary from 1 (last in the order) to 2 (first in the order), to deal with the varying number of finalists (i.e., range standardization of the data). Absence from the finals was coded as 0, which could be due to not competing that year, or to competing but not qualifying for the finals (only the Brazilian team has made it to the finals every time). Adjustments were made to deal with changes in the political entities that the teams represent (i.e., the breakup of Czechoslovakia, Germany, Yugoslavia, and the USSR).

The resulting phylogenetic network (Figure 6) is not very tree-like, and there are no clear clusters of countries based on their competition success. However, the data mining does reveal neighborhoods of similarity among the countries, so that some countries have had somewhat similar degrees of success in the World Cup finals. In particular, I have colored two of the largest splits in the data red and blue. These splits break the data into four neighborhoods (rather than clusters), and I have colored their leaf labels differently. In two of the neighborhoods there are also recognizable subsets, which have been given different color shades.

The black-colored neighborhood contains those teams that have usually been unsuccessful when they have appeared in the finals (e.g., they have been eliminated in the first round). There are two subsets here (colored black and gray), with the teams in the gray subset appearing in the finals mainly from 1990 to 2002. The purple neighborhood contains those teams who have usually been moderately successful whenever they have qualified for the finals (e.g., they have made it to the second round). The orange neighborhood recognizes those teams whose finals results have varied from very good to very poor.

The green neighborhood contains those teams who have been successful on most of those occasions when they have appeared in the finals (e.g., they have made it to the quarter-finals). Note that the most successful teams (Brazil, Germany, Italy) do not stand out within this group—this is a by-product of the data standardization. There are also two subsets here (colored green and lime green), with the lime-green subset all making it to the finals in 1970, 1994, and 1998, but otherwise appearing only sporadically.

So, this is an example where the phylogenetic network indicates neighborhoods rather than clusters within the dataset. There are no strong clusters within the World Cup results because no group of teams has dominated (unlike some other sports, such as ice hockey). For example, the team who has won most often, Brazil, has also on occasion done poorly, and the most consistent team, Germany, has not won as often as the Brazilians.

Opera House Acoustics

There are many opera houses and concert halls in the world, intended for acoustic (i.e., nonamplified) orchestral and vocal performances. These performance halls vary considerably in their perceived acoustic desirability. For example, Beranek[50] provides a ranking of many of these halls, based on the judgment of professional performers and music lovers. However, the architectural design of such halls is based on measured physical quantities involving the overall dimensions, reverberation time, sound distribution, and sound diffusion. It is therefore of interest to ask to what extent these acoustic parameters are reflected in the human subjective judgment about quality.

In the analysis presented here, each of the 52 performance halls is an object, and the 10 characteristics are measurements of some of their acoustic variables. Once again, the resulting phylogenetic network (Figure 7) is not very tree-like but does reveal neighborhoods of multivariate similarity among the performance halls. That is, halls near each other in the network have similar acoustic properties. However, the interest here lies in comparing these acoustic characteristics to the subjective judgments of the audience and performers.

The top 11 ranked halls are highlighted in the network (in red), showing that most of them are in the same network neighborhood. That is, the halls that are preferred by the performers and audiences mostly have similar acoustic qualities. Indeed, all of the great concert halls are shoe-box shaped, with the exception of the Teatro Colón, which is horse-shoe shaped, and Cardiff Hall, which is a surround hall. Of these 11 halls, eight were built before 1908, which shows you how little we have learned recently about designing concert halls.

More interestingly, however, the network indicates another set of eight halls that are in the same network neighborhood (colored purple). As a technical note, there are no splits in the network that uniquely separate the red and purple labels into different partitions. There are three large splits that include some of

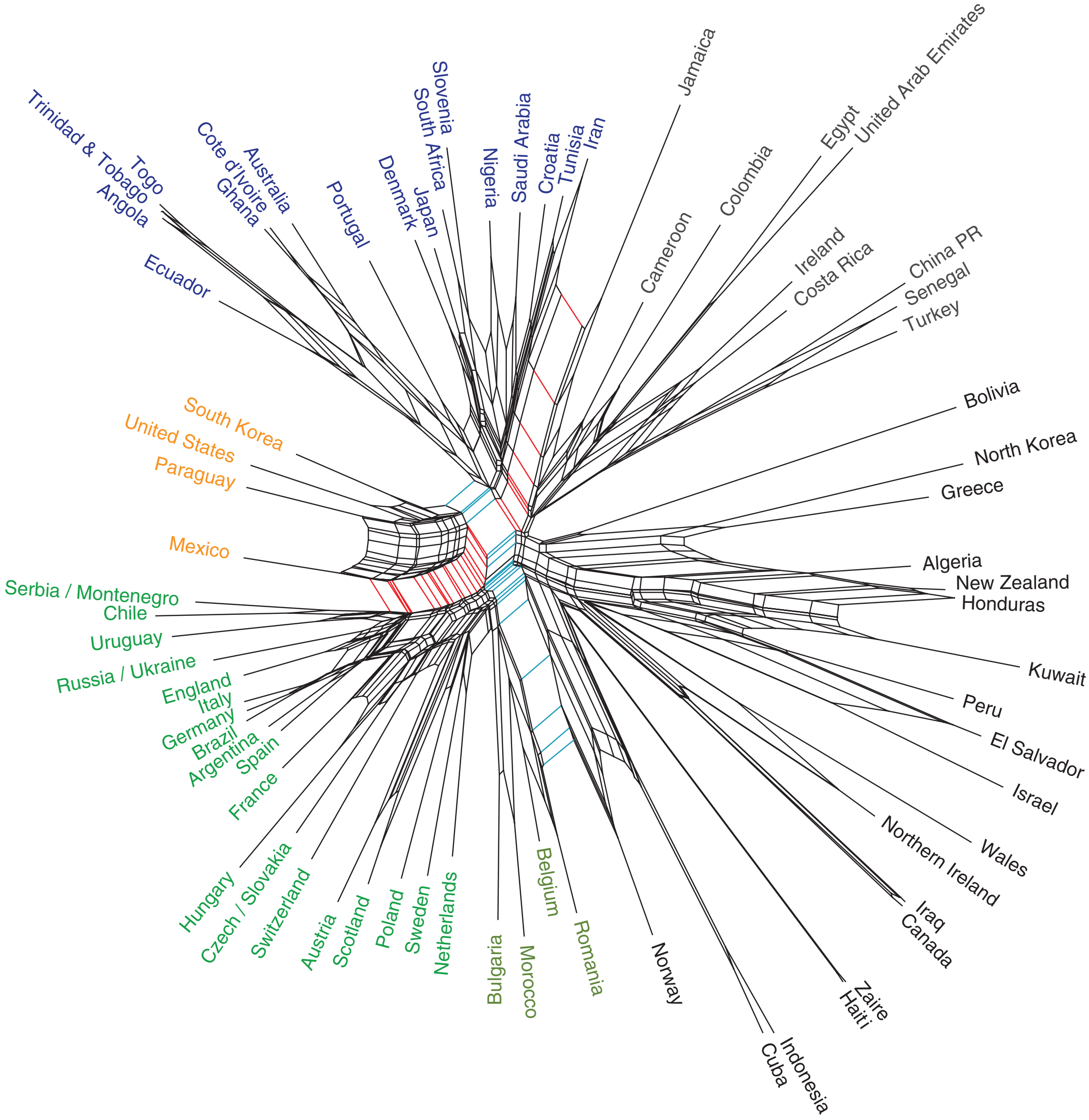


FIGURE 6 | A phylogenetic network of the results from the FIFA World Cup soccer competition, 1930–2010. The network was constructed using the neighbor-net algorithm, with the distance measured as the Steinhaus dissimilarity, which ignores the so-called negative matches. The countries are color-coded by neighborhood within the network. (The data are available from Ref 49.)

the red and purple labels in both partitions, and one small split that separates the purple halls from all of the other halls.

This network arrangement means that the purple-colored halls have similar acoustic qualities to the top-ranked halls (in red), but they have much lower rankings, which means that they are subjectively detected as being less desirable in terms of acoustic performances. Apparently, there is more that meets the ear than can be measured by acoustic instruments. Thus, in this case the data mining reveals a fascinating phenomenon, which is worthy of more detailed study.

Textual Analysis of Genesis 1:3

Textual analysis is an interesting recent use of phylogenetic networks. Stemmatology is the discipline that attempts to reconstruct the transmission history of a written text on the basis of relationships between the various extant versions (e.g., manuscripts or printings). That is, variations in word order, spelling, and punctuation can reveal which editions of particular books are likely to be copies of which other editions. In the analysis presented here, the objects are 26 English-language versions of the Christian Bible, published from 1382 to 2011. The characteristics are

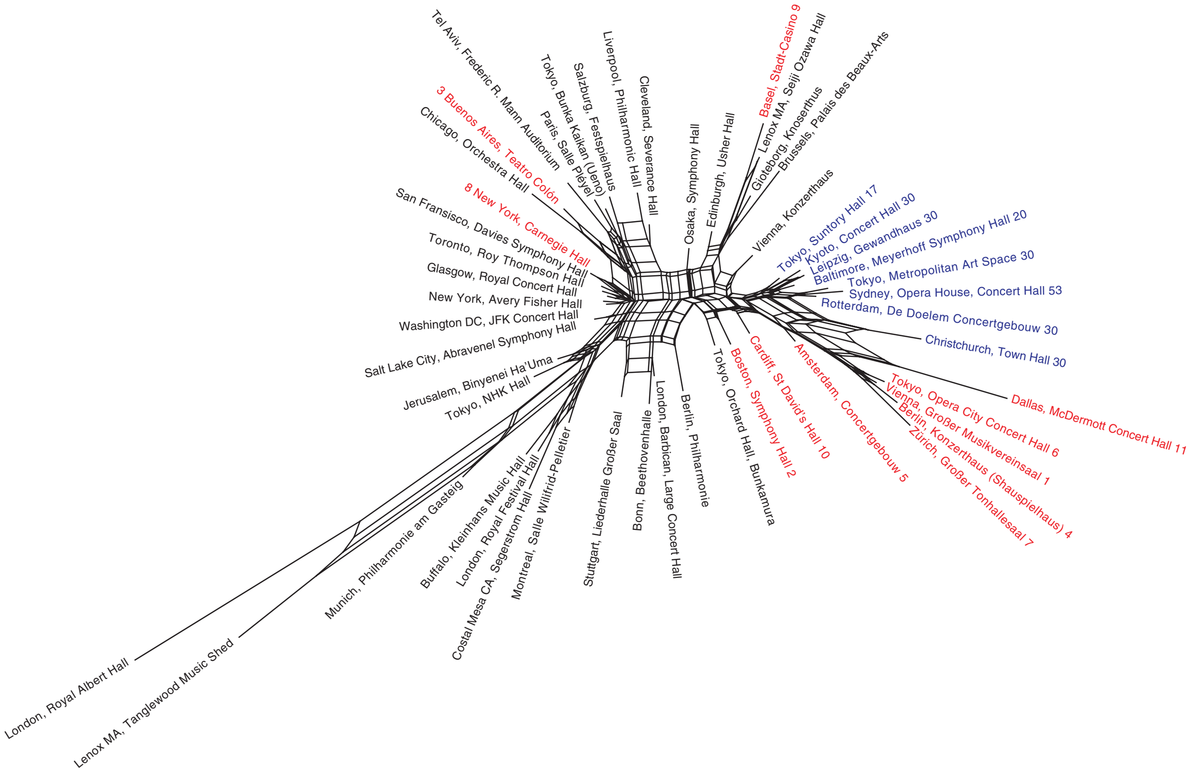

Tel Aviv, Frederic R. Mann Auditorium
3 Buenos Aires, Teatro Colón
Chicago, Orchestra Hall
Tokyo, Bunka Kaikan (Ueno)
Paris, Salle Pléyel
Salzburg, Festspielhaus
Liverpool, Philharmonic Hall
Cleveland, Severance Hall
Osaka, Symphony Hall
Edinburgh, Usher Hall
Basel, Stadt-Casino 9
Lenox MA, Seiji Ozawa Hall
Gioteborg, Knoserthus
Brussels, Palais des Beaux-Arts
Vienna, Konzerthaus
Tokyo, Suntory Hall 17
Kyoto, Concert Hall 30
Leipzig, Gewandhaus 30
Baltimore, Meyerhoff Symphony Hall 20
Tokyo, Metropolitan Art Space 30
Sydney, Opera House, Concert Hall 53
Rotterdam, De Doelem Concertgebouw 30
Christchurch, Town Hall 30
Dallas, McDermott Concert Hall 11
Tokyo, Opera City Concert Hall 6
Vienna, Großer Musikvereinsaal 1
Berlin, Konzerthaus (Shauspielhaus) 4
Zürich, Großer Tonhallesaal 7
Amsterdam, Concertgebouw 5
Cardiff, St David's Hall 10
Boston, Symphony Hall 2
Tokyo, Orchard Hall, Bunkamura
Berlin, Philharmonie
London, Barbican, Large Concert Hall
Bonn, Beethovenhalle
Stuttgart, Liederhalle Großer Saal
Montreal, Salle Wilifrid-Pelletier
Costal Mesa CA, Segerstrom Hall
London, Royal Festival Hall
Buffalo, Kleinhans Music Hall
Munich, Philharmonie am Gasteig
Lenox MA, Tanglewood Music Shed
London, Royal Albert Hall
Tokyo, NHK Hall
Jerusalem, Binyenei Ha'Uma
Salt Lake City, Abravenel Symphony Hall
Washington DC, JFK Concert Hall
New York, Avery Fisher Hall
Glasgow, Royal Concert Hall
Toronto, Roy Thompson Hall
San Fransisco, Davies Symphony Hall
8 New York, Carnegie Hall

FIGURE 7 | A phylogenetic network of the acoustic characteristics of selected concert halls and opera houses. The network was constructed using the neighbor-net algorithm, with the Manhattan distance. The top 11 ranked performance halls are highlighted in red (and their rank is shown), and the other halls in the same network neighborhood are marked in purple. (The data are available from Ref 51.)

the words and punctuation of the third sentence of the Old Testament (Genesis 1:3).

To vary things, this time I have used the reduced-median network algorithm rather than the neighbor-net algorithm to produce the phylogenetic network (Figure 8). This does not produce a planar graph in this particular case, but instead involves a series of interconnected cubes that represent the various bipartitions of the data. It is also a more condensed summary compared to that produced by the neighbor-net algorithm (with Hamming distance).

In this example, the data mining does not even reveal clear neighborhoods in the network, let alone clusters. There is a general separation of the older Genesis texts on the left of the graph and the more recent texts on the right, indicating the gradual change in English through the centuries. However, there are no clear relationships to known copying between versions of the Bible.

Historically, we would expect the Tyndale Bible, Coverdale Bible, Matthew Bible, and Great Bible texts to be closely related, but the Great Bible seems not to fit this expectation. Additionally, we would expect a similarity between the Geneva Bible and the Bishop's Bible, which is also not reflected in the study sentence; nor is the acknowledged debt of the King James Version to the Tyndale Bible.

However, the fact that the Wycliffe Bible and Later Wycliffe were written in Middle English rather than Modern English is clear from their distant network relationship to the other texts; and the close historical relationship of the Challoner Revision and the Douay-Rheims Bible is also clear.

Several texts show isolated relationships. The Knox Bible, for example, is unique among the modern texts in being taken from the Latin Vulgate rather than the original Hebrew text, while the Common English Bible is unusual in trying to balance two translation principles (Dynamic Equivalence and Formal Equivalence) rather than using only one. On the other hand, the New International Version is clearly a very traditional version of the text, given its relationships as shown in the graph, which perhaps explains its modern popularity (it is apparently the most widely used current Bible).

The close association of the Good News Bible with Young's Literal Translation is interesting, given that the former is an (often criticized) free paraphrase of the original Hebrew text while the latter is a literal

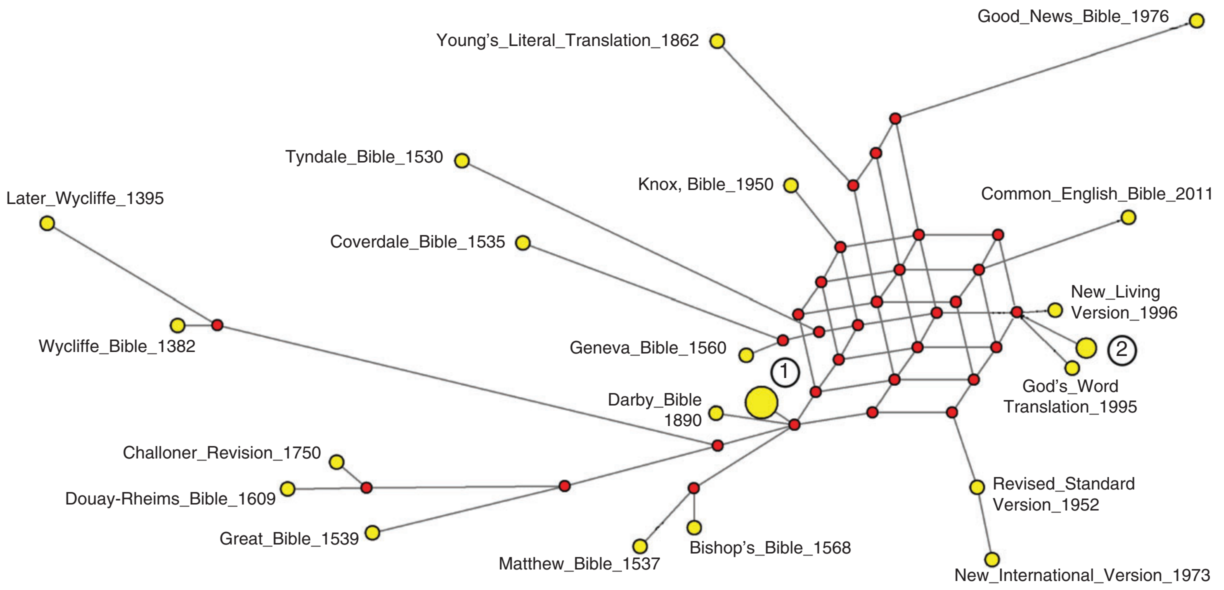


FIGURE 8 | A phylogenetic network of the text of the third sentence of the Bible (Genesis 1:3) as it appears in various published editions. The network was constructed using the reduced-median network algorithm (based on $r=2$). There are five versions of the Bible superimposed at label ①(Webster's Bible 1833, English Revised Version 1885, American Standard Version 1901, King James Version 1611, Blayney Revision 1769), and two are superimposed at label ②(New King James Version 1982, New American Standard Bible 1971). The data were collated from various sources on the Internet.

translation of that same text—you can't get more different translation principles.

Single-Malt Scotch Whiskies

I have been reliably told, by people with extensive experience of the matter, that each and every Scotch single-malt whiskey is unique, and that therefore personal preference for one over another is entirely justified. This claim can be assessed by comparing the whiskies based on quantitative measurement of their various sensory characteristics. In the analysis presented here, the objects are the products of 109 whisky distilleries, from throughout Scotland. The 68 characteristics include the standard beverage-assessment features: nose (12), color (14), body (8), palate (15), and finish (19).

A weighted similarity measure was used, to give the five types of characteristic equal influence (note that the 68 characteristics are not equally distributed among the five assessment features). The similarity coefficient ignores so-called negative matches, so that only shared characteristics generate similarity (not shared lack of a characteristic).

The network is not tree-like (Figure 9), although some neighborhoods are indicated. This contrasts with the results of the agglomerative clustering analysis performed by Lapointe and Legendre,[52] who produced a classification of the whiskies based on the clusters that they detected. They concluded that there is, indeed, a weak but detectable relationship between their classification and the geographical location of the various distilleries. The network calls into question any such classification scheme, and there is very little evidence of geographical patterns.

This example contrasts the results of clustering and networks for data mining. The clustering puts the objects into exclusive groups irrespective of whether such groups exist, whereas the network allows the objects to be members of many groups, if that better represents the data patterns. In this case, where there is little evidence of grouping, the network delivers a more reasonable summary of the multivariate data.

Thai Buddha Images

Anthropological data are very likely to involve horizontal flows of historical information (between contemporaneous cultures) as well as vertical ones (from generation to generation within cultures). One way to assess the balance between these flows is to analyze the data using a phylogenetic network—if the network method produces a tree-like diagram then we can safely conclude that vertical descent has had a larger influence on the transmission of the cultural information than has horizontal transfer.

In the analysis presented here, the objects are 42 cast metal Buddha statues from seven widely recognized chronological culture-historical groups in Thailand. The characteristics are 17 morphological

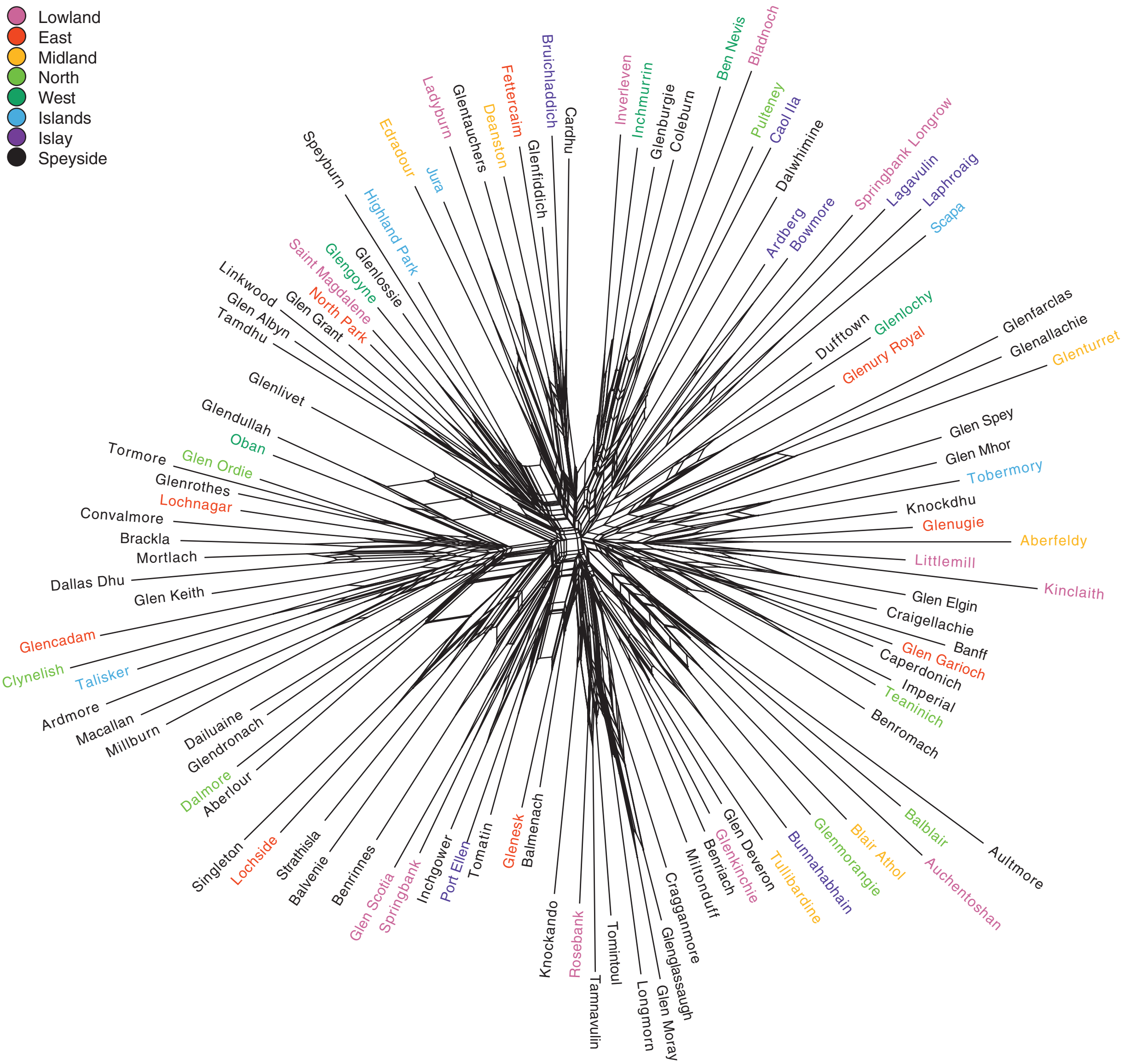


FIGURE 9 | A phylogenetic network of various sensory characteristics of single-malt Scotch whiskies. The network was constructed using the neighbor-net algorithm, with the weighted Bray-Curtis similarity which ignores so-called negative matches. The label colors represent different geographical regions within Scotland. (The data are available from Ref 52.)

features of the statues' heads. Clearly, the network is not very tree-like (Figure 10), and so we can infer that there has been a considerable influence of horizontal flow of cultural information, as well as the vertical flow through time.

There are, however, distinct temporal patterns in the network. The samples from the earliest three periods (Dvaravati, Khmer, Thirteenth Century) are at the right-hand side of the network, while the samples from the next period (Sukhothai) are at the bottom-left. This implies that a large stylistic change occurred between the Thirteenth Century and the Sukhothai periods. Furthermore, the Khmer period style is rather distinct from that of the immediately preceding period (Dvaravati) and the immediately following one (Thirteenth Century), which are themselves not as distinct. That is, there was no stylistic change between the first two periods, but there was a small change to the next period, and then a large change to the following period.

The samples from the latest two periods (Lan Na, Late Ayutthaya) are gathered mainly in two locations, at the bottom of the graph and at the top-left. This indicates that, although there are two distinct styles, they do not correlate with the two culture-historical periods. The samples from the Early

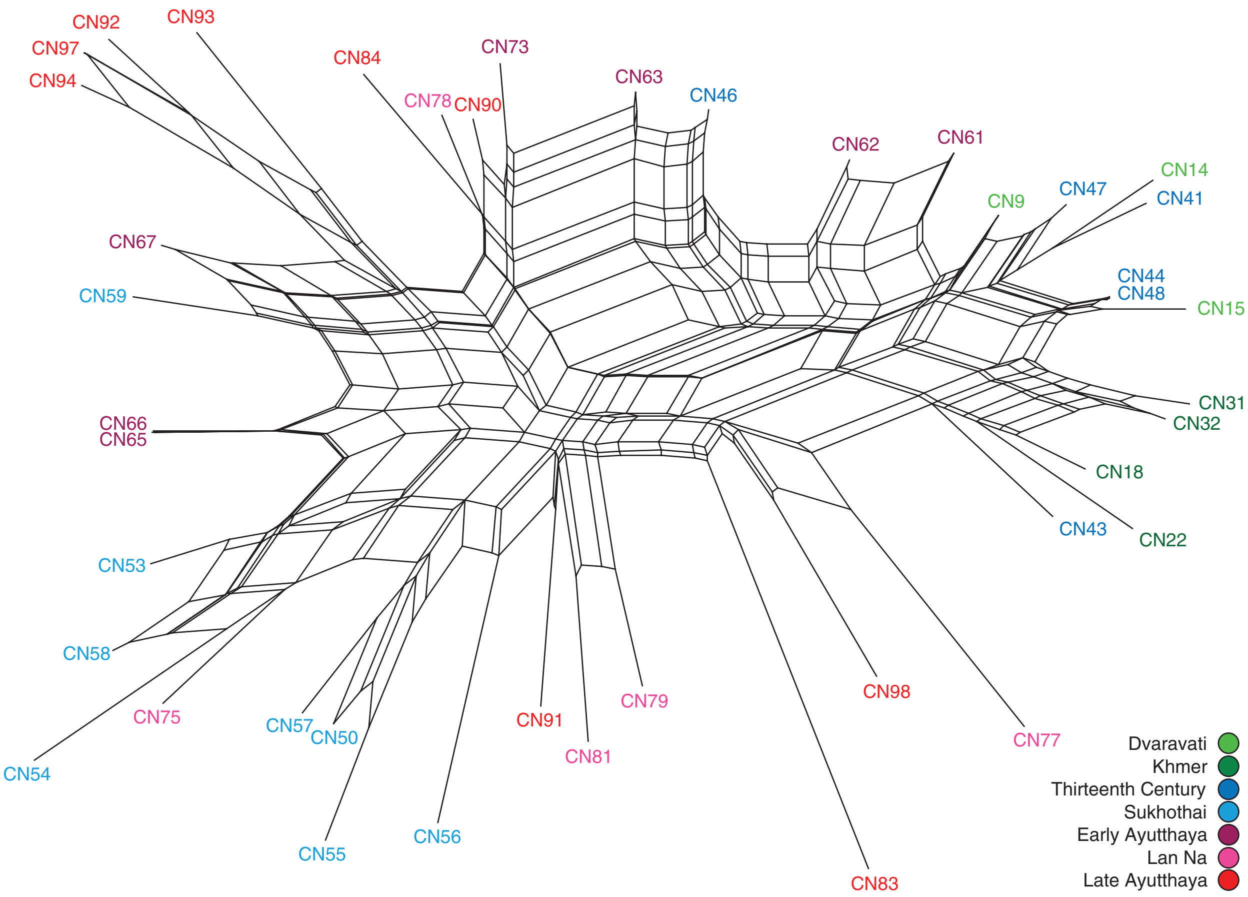


FIGURE 10 | A phylogenetic network of morphological features of Thai Buddha statues. The network was constructed using the neighbor-net algorithm, with the Hamming distance. The labels refer to the museum catalog numbers of the statues, and the colors represent different culture-historical groupings. (The data are available from Ref 53.)

Ayutthaya period are scattered throughout the top and left of the network, suggesting that this is an intermediate style between that of the immediately previous Sukhothai period and the earliest three periods, rather than being an innovative style leading to the succeeding Lan Na period.

In this case, the data mining has detected some patterns that are not strictly historical ones, for which we need to look for some other explanation. The knowledge discovery in this case is the realization that we need to search for further information, because things are not as simple as they might be.

CONCLUSION

Worthwhile methods of data mining will lead to knowledge discovery, and the graphical displays associated with EDA are often an important component of data mining. In this overview, I have shown that phylogenetic networks succeed in the objective of providing effective graphical summaries of multivariate data. They combine several of the good features of previous multivariate data-summary techniques, including ordination, hierarchical clustering, and fuzzy clustering. Moreover, they avoid some of the known mathematical limitations of these alternative methods, which are not always realistic. Phylogenetic networks thus allow multiple patterns to be displayed in a connected graph in a manner that effectively summarizes the information content of any multivariate dataset. They were developed in the biological field of phylogenetics, but they exhibit much wider applicability than this. Phylogenetic networks are fast and relatively easy to calculate, which makes them ideal as a tool for EDA.

I have provided an overview of the field, with particular reference to the use of splits graphs. There are various types of splits graph, which summarize the multivariate data in different ways, and the neighbor-net graph seems to be the most generally useful of the available algorithms. I have presented example analyses based on a wide variety of datasets, including biology, architecture, archeology, sports, stemmatology, music, and gastronomy. My intention has been to encourage the more widespread use of these networks whenever a multivariate summary of a dataset is required.

REFERENCES

1. Pemberton TJ, DeGiorgio M, Rosenberg NA. Population structure in a comprehensive genomic data set on human microsatellite variation. *Genes Genom Genet* 2013, 3:891–907.
2. Morrison DA. Using data-display networks for exploratory data analysis in phylogenetic studies. *Mol Biol Evol* 2010, 27:1044–1057.
3. Mace R, Holden CJ, Shennan SJ, eds. *The Evolution of Cultural Diversity: A Phylogenetic Approach*. London: UCL Press; 2005.
4. Forster P, Renfrew C, eds. *Phylogenetic Methods and the Prehistory of Languages*. Cambridge, UK: McDonald Institute of Archaeological Research; 2006.
5. Lipo CP, O'Brien MJ, Collard M, Shennan SJ, eds. *Mapping Our Ancestors: Phylogenetic Approaches in Anthropology and Prehistory*. New Brunswick, NJ: Aldine Transaction; 2006.
6. Forster P, Toth A. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proc Natl Acad Sci USA* 2003, 100:9079–9084.
7. Spencer M, Wachtel K, Howe CJ. Representing multiple pathways of textual flow in the Greek manuscripts of the Letter of James using reduced median networks. *Comput Humanit* 2004, 38:1–14.
8. Tehrani JJ. The phylogeny of Little Red Riding Hood. *PLoS One* 2013, 8:e78871.
9. Tëmkin I, Eldredge N. Phylogenetics and material cultural evolution. *Curr Anthropol* 2007, 48:146–153.
10. Raval A, Animesh RA. *Introduction to Biological Networks*. Boca Raton, FL: CRC Press; 2013.
11. Proulx SR, Promislow DE, Phillips PC. Network thinking in ecology and evolution. *Trends Ecol Evol* 2005, 20:345–353.
12. Omony J. Biological network inference: a review of methods and assessment of tools and techniques. *Ann Res Rev Biol* 2014, 4:577–601.
13. Morrison DA. Phylogenetic networks are fundamentally different from other kinds of biological networks. In: Zhang WJ, ed. *Network Biology: Theories, Methods and Applications*. Nova Science Publishers: New York; 2013, 23–68.
14. Orlando L, Hänni C, Douady CJ. Mammoth and elephant phylogenetic relationships: *Mammut americanum*, the missing outgroup. *Evol Bioinform* 2007, 3:45–51.
15. Gauch HG, Whittaker RH, Wentworth TR. A comparative study of reciprocal averaging and other ordination techniques. *J Ecol* 1977, 65:157–174.
16. Minchin PR. An evaluation of relative robustness of techniques for ecological ordinations. *Vegetatio* 1987, 69:89–107.
17. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 2008, 40:646–649.
18. Bryant D. Radiation and network breaking in Polynesian linguistics. In: Forster P, Renfrew C, eds. *Phylogenetic Methods and the Prehistory of Languages*. McDonald Institute of Archaeological Research: Cambridge, UK; 2006, 111–118.
19. Huson DH, Scornavacca C. A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol* 2011, 3:23–35.
20. Morrison DA. *Introduction to Phylogenetic Networks*. Uppsala, Sweden: RJR Productions; 2011.
21. Huson DH, Rupp R, Scornavacca C. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge, UK: Cambridge University Press; 2011.
22. Dress A, Huson D, Moulton V. Analyzing and visualizing sequence and distance data using SplitsTree. *Discrete Applied Mathematics* 1996, 71:95–109.
23. Dress AWM, Huson DH. Constructing splits graphs. *IEEE/ACM Trans Comput Biol Bioinform* 2004, 1:109–115.
24. Huber KT, Moulton V. Phylogenetic networks. In: Gascuel O, ed. *Mathematics of Evolution and Phylogeny*. Oxford, UK: Oxford University Press; 2005, 178–204.
25. Wikipedia. Available at: http://en.wikipedia.org/wiki/Simon_&_Garfunkel_discography. (Accessed January 25, 2013).
26. Bandelt H-J. Phylogenetic networks. *Verhand Naturwiss Vereins Hamburg* 1994, 34:51–71.
27. Bandelt H-J, Macauley V, Richards M. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogenet Evol* 2000, 16:8–28.
28. Bandelt H-J, Forster P, Röhl A. Median-joining networks for inferring intraspecies phylogenies. *Mol Biol Evol* 1999, 16:37–48.
29. Bandelt H-J, Huber KT, Moulton V. Quasi-median graphs from sets of partitions. *Discrete Appl Math* 2002, 122:23–35.

30. Bandelt H-J, Dür A. Translating DNA data tables into quasi-median networks for parsimony analysis and error detection. *Mol Phylogenet Evol* 2007, 42:256–271.

31. Bandelt H-J, Forster P, Sykes BC, Richards MB. Mitochondrial portraits of human populations using median networks. *Genetics* 1995, 141:743–753.

32. Huber KT, Watson EE, Hendy MD. An algorithm for constructing local regions in a phylogenetic network. *Mol Phylogenet Evol* 2001b, 19:1–8.

33. Bandelt H-J. Exploring reticulate patterns in DNA sequence data. In: Bakker FT, Chatrou LW, Gravendeel B, Pelser PB, eds. *Plant Species-Level Systematics: New Perspectives on Pattern and Process*. Königstein, Germany: Koeltz; 2005, 245–269.

34. Huber KT, Moulton V, Lockhart P, Dress A. Pruned median networks: a technique for reducing the complexity of median networks. *Mol Phylogenet Evol* 2001a, 19:302–310.

35. Ayling SC, Brown TA. Novel methodology for construction and pruning of quasi-median networks. *BMC Bioinform* 2008, 9:115.

36. Bandelt H-J, Dress AWM. A relational approach to split decomposition. In: Opitz O, Lausen B, Klar R, eds. *Information and Classification*. Berlin, Germany: Springer; 1993, 123–131.

37. Bandelt H-J, Dress AWM. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol* 1992b, 1:242–252.

38. Bryant D, Moulton V. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. *Lect Notes Comput Sci* 2002, 2452:375–391.

39. Bryant D, Moulton V. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* 2004, 21:255–265.

40. Holland B, Moulton V. Consensus networks: a method for visualizing incompatibilities in collections of trees. *Lect Notes Bioinform* 2003, 2812:165–176.

41. Holland BR, Delsuc F, Moulton V. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. *Syst Biol* 2005, 54:66–76.

42. Holland BR, Jermiin LS, Moulton V. Improved consensus network techniques for genome-scale phylogeny. *Mol Biol Evol* 2006, 23:848–855.

43. Huson DH, Dezulian T, Klöpper T, Steel MA. Phylogenetic super-networks from partial trees. *IEEE/ACM Trans Comput Biol Bioinform* 2004, 1:151–158.

44. Huson DH, Steel MA, Whitfield J. Reducing distortion in phylogenetic networks. *Lect Notes Bioinform* 2006, 4175:150–161.

45. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 1998, 14:68–73.

46. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 2006, 23:254–267.

47. Röhl A. Phylogenetische netzwerke. PhD Thesis, Department of Mathematics, University of Hamburg, Germany, 1999.

48. Detournay O, Morrison DA, Wagner B, Zarnegar B, Wattrang E. Genomic analysis and mRNA expression of equine type I interferon genes. *J Interferon Cytokine Res* 2013, 33:746–759.

49. Wikipedia. http://en.wikipedia.org/wiki/National_team_appearances_in_the_FIFA_World_Cup (accessed July 6 2012).

50. Beranek LL. Subjective rank-orderings and acoustical measurements for fifty-eight concert halls. *Acta Acustica* 2003, 89:494–508.

51. Skålevik M. Available at: http://www.akutek.info/concert_hall_acoustics_files/parameters.htm. (Accessed August 23, 2013).

52. Lapointe F-J, Legendre P. A classification of pure malt scotch whiskies. *Appl Stat* 1994, 43:237–257.

53. Marwick B. A cladistic evaluation of ancient Thai bronze Buddha images: six tests for a phylogenetic signal in the Griswold Collection. In: Bonatz D, Reinecke A, Tjoa-Bonatz ML, eds. *Connecting Empires*. Singapore: National University of Singapore Press; 2012, 159–176.