Representing Median Networks with Concept Lattices[⋆]

Uta Priss

Ostfalia University of Applied Sciences
Wolfenbüttel, Germany
www.upriss.org.uk

Abstract. Median networks have been proposed as an improvement over trees in phylogenetic analysis. This paper argues that concept lattices represent essentially the same information as median networks but with the advantage that there is a larger FCA research community and a variety of available software tools. Therefore evolutionary analysis is an interesting new application domain for FCA.

1 Introduction

The field of phylogenetics tries to establish evolutionary relations among groups of organisms - usually in form of evolutionary trees. For example, by sampling DNA from organisms and looking at differences evolutional changes can be reconstructed. For obvious reasons most of the DNA is extracted from currently living organisms, thus any reconstruction of phylogenetic trees is somewhat hypothetical. There are established means for inferring such trees (for example, involving "genetic distances", statistical maximum parsimony and maximum likelihood) but in cases where parallel mutations or reversals occur, it is difficult to decide on the exact sequences of the mutations. For example, the left-hand side in Figure 1 shows two possible trees for the changes between 1, 2, 3 and 4. As Sykes (2001, p. 178) explains, in such cases it is often not necessary to ultimately decide which change occurred first, i.e., whether 4 derived from 1 via 2 or via 3. Instead of deciding which of the trees is correct, one can use a graph as shown in the right half of Figure 1 which summarises both possible trees. Not only simplifies this the analytic process, it can also lead to more readable diagrams. Bandelt et al. (1995) have developed the construction of such graphs into a method using median networks as explained in the next section.

Since the graph on the right-hand side of Figure 1 is a lattice and since trees can be embedded into lattices, the question arises as to whether Formal Concept Analysis[1] (FCA) can be used instead of or in addition to median networks. One advantage of using FCA is that FCA has a larger research community than median networks/graphs[2].

[⋆] Published in Pfeiffer et al (eds.), Proc of the 20th ICCS, Springer, LNAI 7735, 2013.

[1] Because FCA has been a topic of this conference for many years, this paper does not provide an introduction to FCA. Information about FCA can be found, for example, on-line (http://www.upriss.org.uk/fca/) and in the main FCA textbook by Ganter & Wille (1999).

[2] As confirmed by retrieving about 10000 hits for a search for "formal concept analysis" on Google Scholar, as opposed to 1200 for "median network" and 900 for "median graph".

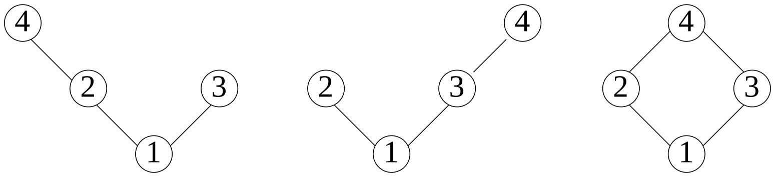

Fig. 1. Two possible trees (on the left) are summarised in one graph (on the right)

Furthermore, there exist more well-tested software tools for FCA[3] compared to median networks and, for example, Bandelt et al. (2000) still discuss "manual construction" of median networks alongside some algorithms.

From the viewpoint of FCA, it is interesting to establish a further application of FCA in the field of genetics or bioinformatics (for median networks this was first suggested by Priss (2012)) and a further connection with a similar or related graphical representation method. This extends previous research showing similarities between FCA and other fields, for example, Priss and Old (2008) show that concept lattices are similar to lattice-based methods developed in information retrieval and computational linguistics. The following section provides further details about median networks. Section 3 discusses how the phylogenetic data can be modelled with FCA and what is different or similar to how the data is modelled with median networks. The paper finishes with a concluding section.

2 Median networks and phylogenetics

This section provides a brief introduction to the application area of this paper[4]. Median graphs are undirected graphs where any three vertices have a unique median. More precisely, an *interval* between two vertices x, y in a graph is defined as $I(x, y) = \{v \mid d(x, y) = d(x, v) + d(v, y)\}$ where $d()$ is the usual distance function in a graph. In other words, an interval consists of the vertices on the shortest paths between two vertices. A graph is called a *median graph* if the following property holds: $\forall_{x,y,z} : |I(x, y) \cap I(x, z) \cap I(y, z)| = 1$. That means that there is a unique vertex (called *median*) that belongs to shortest paths between any two of three vertices.

Below is a brief summary of the close relationship between median graphs, distributive lattices and median semilattices (mostly following Bandelt (1984)). In this paper we are using the dual of the usual definition of a median semilattice (which we call a "reverse" median semilattice) because it fits better with the constructions in the next section. A *reverse median semilattice* is a join-semilattice such that every principal filter $\{x \mid x \geq a\}$ is a distributive lattice and any three elements have a lower bound whenever each pair of them does.

- The covering graph of any finite distributive lattice is a median graph.

[3] See http://www.upriss.org.uk/fca/fcasoftware.html

[4] Based on Bandelt et al. (1995 and 2000), Sykes (2001) and Wikipedia pages.

- A finite graph G is the covering graph of a finite distributive lattice $\Longleftrightarrow$ G is a median graph with two vertices 0 and 1 such that every other vertex lies on a shortest path between them.
- In a distributive lattice, Birkhoff's *median operation* can be observed: $m(a, b, c) = (a \wedge b) \vee (a \wedge c) \vee (b \wedge c) = (a \vee b) \wedge (a \vee c) \wedge (b \vee c)$ which also fulfills the axioms of a *median algebra*.
- Every median graph is a covering graph of a reverse median semilattice with largest element a where a is any fixed vertex.
- The covering graph of a reverse median semilattice S is a median graph provided that S is discrete, i.e., all intervals are finite.
- A discrete lattice L with 0 is distributive $\Longleftrightarrow$ the covering graph of L is median.
- Tree graphs are median graphs.

Although the relationship between median graphs and lattices is mathematically well-understood, there are still open questions left with respect to how FCA can be used to generate meaningful concept lattices from the data.

As mentioned in the introduction, in the field of phylogenetics, it is attempted to infer evolutionary trees from observed characteristics of species. Trees are considered best if they are *most parsimonious* which means that the number of presumed evolutionary changes is minimal. For example, in the right-hand side of Figure 1 it would be more parsimonious to assume that 2 evolved directly from 1 instead of evolving from 1 via 4 and 3. A goal of phylogenetic analysis is to compute all "most parsimonious trees" for a given data set, thus out of all possible trees the ones with minimal number of changes. Unfortunately, this is a computationally complex task. *Median networks* (or *Buneman graphs*) are median graphs where each vertex represents a species and each edge represents a genetic change. Bandelt et al. (1995) argue that since a median network is guaranteed to contain all most parsimonious trees, it is a preferred representation of evolutionary change and a significant improvement over other methods which artificially construct a tree from the data using statistical methods (see also Sykes (2001) and Bandelt et al. (1995 and 2000)).

Figure 2 shows an example of a median network on the left-hand side. The example is hypothetical and not based on real data. On the left side are white mice versus brown mice on the right. The top two vertices represent large mice, the other ones small mice. The bottom two vertices represent tailless as opposed to tailed mice. The vertex on the left in the middle is empty (*latent*) because no species in the data displays these characteristics. This vertex is generated from the data because without it, it would not be a median graph and not contain all most parsimonious trees. Without assuming that small tailed white mice are latent, the difference between small and large white mice would have coincided with loss of tail whereas in brown mice first the size changed, then the tail was lost. Not all possible combinations are latent. For example, the existence of large tailless mice is not implied by the data. The right-hand side of Figure 2 is explained in the next section.

The median network in Figure 2 summarises all possible evolutionary trees. If one assumes that the root of the trees is the top left vertex, four trees are possible. For example, large white mice could have first become small and then brown or first become brown and then small. While the sequence between the changes in colour and size is

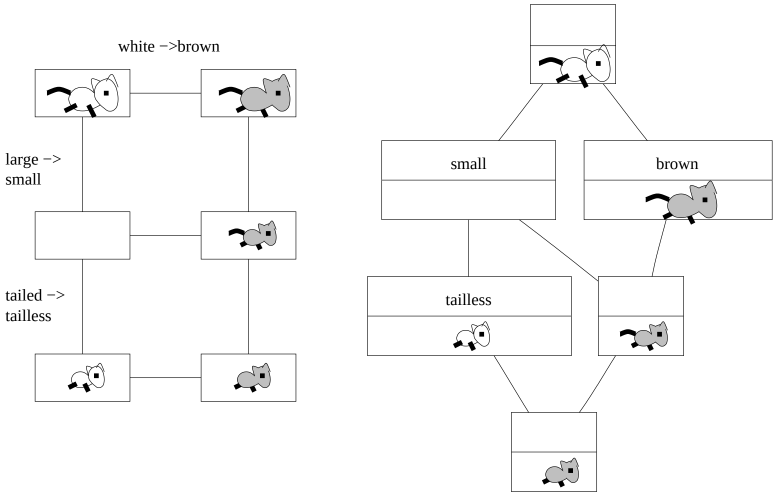

Fig. 2. Median network with latent vertex (left) and concept lattice (right)

not known, the change in size definitely preceded the loss of tail. If one assumes that the change in size for white mice occurred before the change in colour, then the change in colour is an instance of *parallel mutation* because large white mice became brown independently of small white mice becoming brown. If the change in colour occurred first, then the change in size would be parallel mutation. If no parallel mutations or reversals were to occur in some data, then its median network would automatically be a tree. But considering the examples by Bandelt et al. (1995 and 2000), most data sets tend to contain at least some parallel mutations.

If the sample size is large, an unmodified median network may be too complex to be graphically represented. Bandelt et al. (1995) suggest a method for reducing median networks based on weight and frequency. In order to construct a median network, one summarises all changes that occur simultaneously with respect to the sample species as "weight". For example, if colour changes in mice always correspond to changes in ear size (hypothetically), then one would not draw separate edges for colour and ear changes. Instead one would record one change but with a higher weight. Graphically this can be represented by drawing a longer edge.

In the same manner, if several species have the exact same characteristics, one creates only one vertex for this group of species but records a higher frequency for this vertex. This can be graphically represented by a larger node for the vertex. Using frequencies and weights one can reduce the network by eliminating some of the edges which are less likely to have occurred. Bandelt et al. (1995) state that in all examples they considered so far even reduced networks still contained all most parsimonious trees, but there is no guarantee that that is always the case.

Characteristics in phylogenetic analysis are often binary, i.e., having two possible values. In the example in Figure 2, the characteristics are naturally binary (such as large or not large). Other characteristics can be made binary. For example, although DNA sequences can be of four values (A, G, C or T), Bandelt et al. (1995) argue that it is unusual for more than one change to occur at the same site in a set of closely related species. Thus it is sufficient to record for each site whether a change occurred or not, ignoring the value of the change.

A median network contains all most parsimonious trees independently of where the root of the tree is. There are methods for determining the root or evolutionary ancestor of a set of species although it might not be easy and the root might be latent. One method is to compare a set of species with an *outgroup* or *reference group* which is more distantly related to all the other species than they are too each other.

3 Modelling with FCA

It is straightforward to represent the example on the left-hand side in Figure 2 as a concept lattice as presented on the right-hand side. One advantage of using FCA is the availability of established mathematical vocabulary for describing the phylogenetic phenomena. Important phylogenetic notions can be directly translated into FCA terminology. Series of evolutionary changes that are unambiguous correspond to attribute implications in the lattice. For example, the implication from "tailless" to "small" in the lattice in Figure 2 corresponds to the evolutionary loss of tail occurring after the change in size. Latent species correspond to concepts that do not contain objects in their contingent. Each meet-reducible concept in the lattice corresponds to a choice point between different possible trees.

Table 1 shows a more complex example using mitochondrial data from Ward et al. (1991) which was also used by Bandelt et al. (1995). In FCA terms it represents a many-valued context. The second row from the top shows the default values for each column. A dot in the matrix means that the default value occurs. A letter indicates a change. As can be seen in the table, only one type of change occurs in each column. For example, in the first data column the default value is "T" which is changed to "C" in three rows. No changes to "A" or "G" occur in the first data column. As discussed by Bandelt et al. (1995) this is usually the case. Therefore such tables can be interpreted as binary matrices or single-valued contexts by only considering whether the default value or a change occur and ignoring the type of change.

In FCA terminology, the formal objects in Table 1 are 28 mitochondrial lineages. The right-hand column indicates the frequency of the lineages. For example, lineage 1 occurred in 3 individuals. A total number of 63 individuals was involved in the study. The formal attributes encode the positions where the DNA sequences occur in the human reference sequence. If one encodes the attributes so that each cross represents the positions where an object differs from the reference group then the top of the lattice will correspond to the root of the possible evolutionary trees. This is because, as discussed in the previous section, comparison with a reference group can be used to determine the root. Using FCA the preprocessing of summarising objects with identical row values and attributes with identical column values is not really necessary because such objects (or equivalently attributes) would be grouped into the contingent of a single concept automatically in the concept lattice.

Table 1. Nuu-Chah-Nulth mitochondrial lineages (Ward et al., 1991) as a formal context

	69	88	91	106	124	149	162	166	190	194	200	219	233	247	251	255	267	271	275	296	301	302	304	319	339	344	
	T	C	C	G	C	T	C	T	G	T	C	C	C	C	G	C	C	C	T	G	T	T	C	T	T	A	
1	.	.	.	.	.	.	.	C	A	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	3
2	.	.	.	.	.	.	.	.	A	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	2
3	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	1
4	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	1
5	.	T	.	A	.	.	T	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.	2
6	.	T	.	A	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	A	.	.	.	.	C	.	2
7	C	T	.	A	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	A	.	.	.	.	C	.	1
8	.	T	.	A	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.	2
9	C	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.	2
10	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	G	1
11	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.	5
12	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.	9
13	.	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	T	.	.	A	.	.	.	.	C	.	1
14	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	T	T	.	.	A	.	.	.	.	C	.	1
15	.	T	.	.	.	.	.	.	.	.	T	.	.	.	.	T	T	.	.	A	C	.	.	.	C	.	2
16	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	.	.	.	.	T	.	C	.	1
17	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	.	C	.	1
18	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	2
19	.	.	T	.	.	.	.	.	.	.	T	.	.	.	.	.	.	T	.	.	.	C	.	.	C	.	1
20	.	.	.	.	.	C	.	.	.	.	T	.	.	.	A	.	.	.	.	.	.	C	.	.	C	.	3
21	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	C	.	3
22	C	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	3
23	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	C	.	.	C	.	.	.	.	1
24	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	C	.	.	C	T	.	.	.	7
25	.	.	.	.	.	.	.	.	.	.	T	T	.	.	.	.	.	.	C	.	.	C	T	C	.	.	3
26	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	.	.	C	T	C	.	.	1
27	.	.	.	.	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
28	.	.	.	.	.	.	.	C	.	C	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	1

Because the median network and concept lattice for Table 1 are fairly complex, we will first discuss a network and lattice derived for a simpler context of the same type before discussing the one in Table 1. Figure 3 shows a concept lattice for a data table discussed by Bandelt et al. 2000 (using HVS I data by Vigilant et al.). Two attributes are called *compatible* in Bandelt's terminology if they are lattice-theoretically comparable or their meet is the bottom node. Bandelt calls a set of attributes a *clique* if the attributes are pairwise compatible and the set is maximal with respect to inclusion. In other words, cliques represent maximal trees. In Figure 3 one clique contains all attributes except 16243 and another clique contains all attributes except 16294 and 16239. These are the only two cliques in Figure 3. Bandelt et al. describe a fairly complicated algorithm for deriving the median network using cliques, peripheral elements and torsos (where the *torso* data matrix consists of the non-compatible attributes).

Figure 4 shows a median network for the data in Figure 3. In contrast to Bandelt et al. (2000), the attributes, frequencies and weights are omitted in the figure. This means that all nodes are of the same size and the length of the edges does not carry meaning. The lattice in Figure 3 is not distributive and thus not a covering graph of a median graph. Nevertheless if one omits the bottom node from the lattice then its covering graph and the median network in Figure 4 differ only by one vertex: the vertex next to the one labelled with "8" in Figure 4. Using the statements about reverse median semilattices from the last section, an algorithm for converting a concept lattice as in Figure 3 into a median network as in Figure 4 consists of omitting the bottom node and

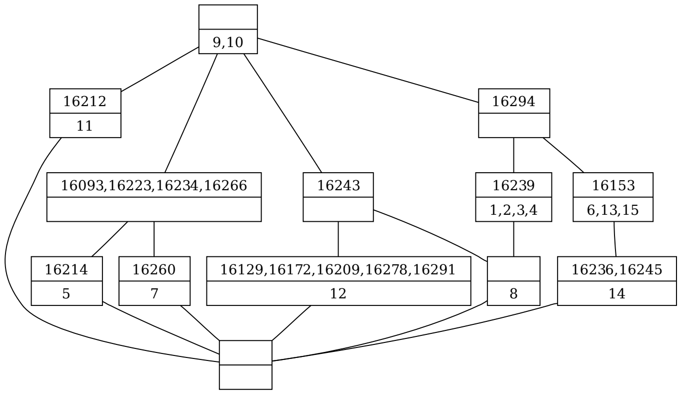

Fig. 3. Concept lattice for HVS I data of Vigilant used by Bandelt et al. (2000)

then checking every principal filter for distributivity and turning it into a distributive lattice if it is not already one.

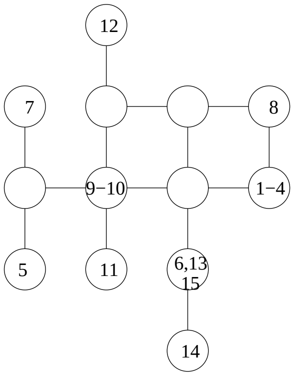

Fig. 4. The median network for Figure 3

The principal filter in Figure 3 that is not distributive is shown on the left in Figure 5 alongside the median network of the torso of Figure 4. In contrast to Figure 2 where both the lattice and the graph produce a latent vertex, in this case the lattice does not have one. The reason is because in Figure 2 the attribute "small" is shared by small tailless white mice and small brown tailed mice whereas in Figure 5 object "12" does not share any attributes with the objects "1, 2, 3, 4". The median network in Figure 5 generates a latent species because the difference between objects "12" and "8" consists only of one characteristic whereas the difference between “5,7,9-11” and “1-4” consists of two characteristics. The lattice in Figure 5 does not contain all most parsimonious trees but the median network on the right side could be generated from it. This is an issue that would need to be discussed with evolutionary biologists. After years of working with FCA, the author’s intuition is that the lattice on the left is a more appropriate representation of the data because it makes fewer assumptions about information that is missing (i.e., latent species). But, presumably, evolutionary biologists have different intuitions about the data than mathematicians. Thus although there is a clear algorithm for converting concept lattices into median networks, the question is whether it is really necessary to do so or whether a concept lattice would be a sufficiently informative representation of the data without containing all most parsimonious trees.

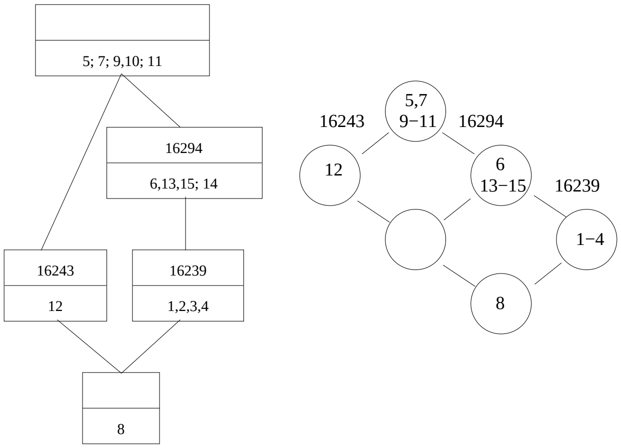

Fig. 5. Concept lattice (left) of the only non-distributive principal filter in Figure 3 and median network (right) which is the “torso” of the median network in Figure 4

Coming back to the data presented in Table 1, Figure 6 shows the reduced median network from Bandelt et al. (1995) for the data. Again, the frequencies and weights are not represented in the diagram. The root of the tree is the node labelled “X”. The network contains 10 latent vertices. Ward et al. (1991) identified four clusters among the lineages ($\{1, 2\}$, $\{5, 6, 7, 8, ..., 15\}$, $\{23, 24, 25, 26\}$, and $\{27, 28\}$) by deriving a phylogenetic tree using statistical methods. Bandelt et al. (1995) criticise the tree presented by Ward et al. because they believe that one cluster is missing and several other clusters could be modified. The large boxes in Figure 6 are meant to indicate the clusters according to Bandelt et al. who observe that the cluster consisting of 18, 19, 20 and 21 (and possibly also 16, 17 and 22) is missing from Ward’s tree and that maybe 3 and 4 should also belong to the cluster of 1 and 2. They argue that the information about the clusters is very clear in the median network but might not be visible in a tree. They further state that these problems are not restricted to Ward's paper but can be observed in other papers as well.

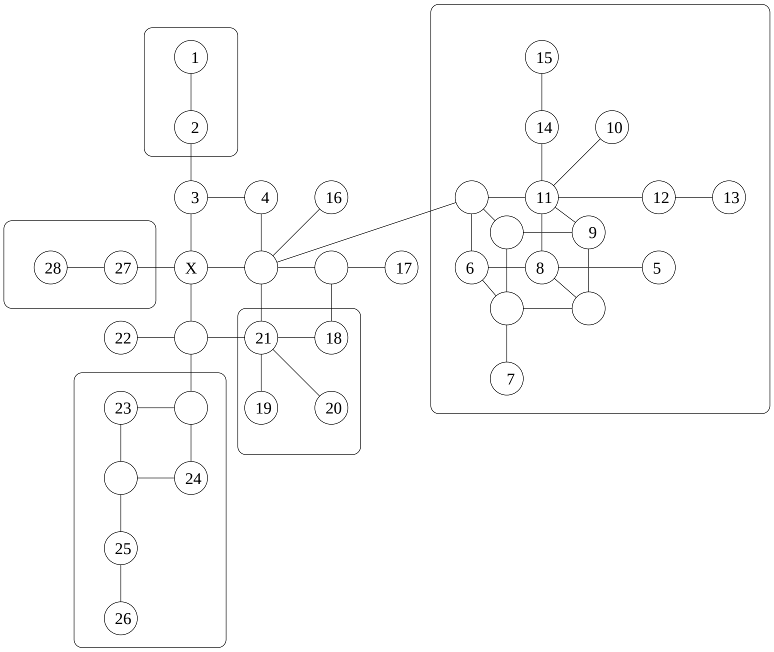

Fig. 6. Reduced median network for Table 1 (following Bandelt et al. (1995))

The median network in Figure 6 is reduced. The reduction algorithm is described by Bandelt et al. in great detail. Effectively the reduction algorithm splits some attributes into versions a and b so that objects in one cluster have version a and the objects in other clusters have b. For example, attribute 166 applies to lineages in two different clusters. If the attribute is split into 166a for lineage 1 and 166b for lineages 27 and 28, then the structure of the network is simplified. The reasoning behind this is that if the same change occurs for lineages that are in very different clusters, it is quite likely that the change does not represent a single event but instead happened several times independently. The basis for these decisions are frequencies and weights. We do not have an exact list of which attributes were split in Figure 6. Therefore the attributes that were split in Figure 7 are not necessarily the same as in Figure 6. We chose to split attributes 69, 166 and 190. Furthermore we completely omitted attribute 200 because it applies to almost all objects. The resulting lattice is shown in Figure 7. Structurally, the graphs in Figure 6 and Figure 7 are quite similar although in Figure 7 attribute 16 is closer to the cluster involving 23 to 26 and there is a connection between 4 and 14. We do not know whether either representation is more plausible from a phylogenetic viewpoint.

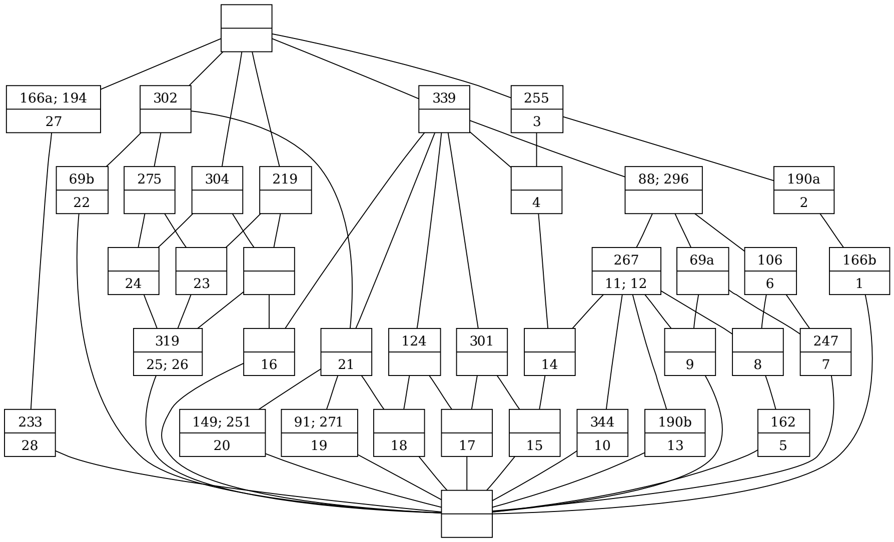

Fig. 7. Reduced lattice for Table 1 (splitting 69, 166, 190 and omitting 200)

In order to decide which attributes to split, one needs to first determine which objects form clusters. Figure 8 shows the object ordering (implications) of Table 1. Apart from the already mentioned connection between 4 and 14, the clusters emerging from the object ordering are the same as the ones discovered by Ward et al. and Bandelt et al. Thus we propose an algorithm for reducing concept lattices as follows: determine clusters of objects by considering the object ordering. Then investigate attributes that apply to objects belonging to different clusters. If these attributes are high up in the lattice, consider splitting the attribute. The resulting lattice will have fewer line crossings and be more "tree like". We are not necessarily proposing that the attributes are completely automatically selected, but that instead expert advice is considered in the selection process.

4 Conclusion

This paper discusses the representation of phylogenetic data as concept lattices instead of or in addition to median networks. Both concept lattices and median networks contain essentially the same information but FCA has a larger research community. The paper sketches an algorithm for converting a concept lattice into a median network and for reducing a lattice based on clustering of objects. Further discussion with phylogenetics researchers will need to establish in how far they would be willing to accept concept lattices that do not contain all most parsimonious trees as a representation of their data.

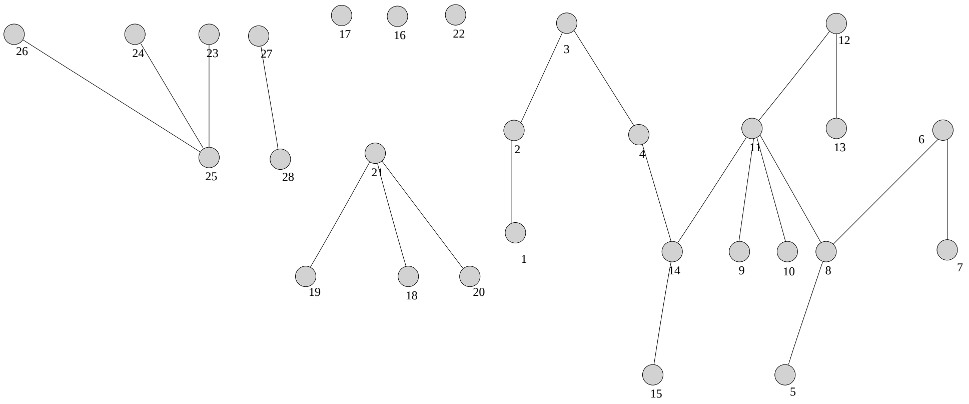

Fig. 8. Object ordering for Table 1

More experiments with larger data sets are needed to determine the practical feasibility of the suggested algorithms and to compare more examples of median networks and concept lattices with respect to the readability of the diagrams. One aim of the paper is to alert the wider FCA community to this application area. Because median graphs have many interesting properties and applications themselves, establishing a connection between them and FCA could lead to further interesting research (for example, social network analysis or other graph and networking applications).

References

1. Bandelt, H. J. (1984). *Discrete Ordered sets whose covering graphs are median.* Proceedings of the American Mathematical Society, 91, 1.
2. Bandelt, H. J.; Forster, P.; Sykes, B. C.; Richards, M. B. (1995). *Mitochondrial portraits of human populations using median networks.* Genetics, Oct, 141, 2, p. 743-753.
3. Bandelt, H.-J.; Macaulay, V.; Richards, M. (2000). *Median networks: Speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA*. Molecular Phylogenetics and Evolution, 16, 1, p. 8-28.
4. Ganter, Bernhard; Wille, Rudolf (1999). *Formal Concept Analysis. Mathematical Foundations.* Berlin-Heidelberg-New York: Springer.
5. Priss, Uta; Old, L. John (2008). Lattice-based Modelling of Thesauri In: Lattice-Based Modeling Workshop, Olomouc, Czech Republic, 2008. Available at: http://researchrepository.napier.ac.uk/3477/.
6. Priss, Uta (2012). *Concept Lattices and Median Networks.* In: Szathmary; Priss (eds.), Proceedings of the Ninth International Conference on Concept Lattices and Their Applications, Universidad de Malaga, p. 351-354.
7. Sykes, Bryan (2001). *The seven daughters of Eve.* Bantam Press.
8. Ward, R. H.; Frazier, B. L.; Dew-Jacer, K.; Pääbo, S. (1991). *Extensive mitochondrial diversity within a single Amerindian tribe.* Proc. Natl. Acad. Sci., USA, 88, p. 8720-8724.