Concept Lattices and Median Networks

Uta Priss

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

Abstract. In phylogenetic analysis, median networks have been proposed as an improvement over tree representations. This paper argues that concept lattices represent a further improvement over median networks because FCA provides a detailed formal description and there are a number of existing software solutions for creating lattices. The purpose of this paper is to raise awareness in the FCA community for this interesting application area in bioinformatics.

1 Introduction

The field of phylogenetics tries to establish evolutionary relations among groups of organisms through molecular sequencing, for example, by sampling DNA from organisms and looking at differences. Reconstruction of phylogenetic trees is somewhat hypothetical because evolutionary relationships are established using DNA from currently living organisms. There are established means for inferring such trees using statistical means but in cases where parallel mutations or reversals occur, it is difficult to decide the exact sequence of the mutations. Therefore, instead of deciding which of the possible trees is more likely, one can use a graph which embeds all possible trees. This simplifies the analytic process and leads to more readable diagrams. Bandelt et al. (1995) develop the construction of such graphs into a method using median networks as explained in the next section. Sykes (2001) and Bandelt et al. (1995 and 2000) argue that using median networks is a significant improvement over construction of hypothetical trees using statistical methods.

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. Furthermore, there exist a variety of well-tested software tools for FCA[2] whereas Bandelt et al. (2000) discuss "manual construction" of median networks alongside some algorithms. For FCA researchers this establishes a further application domain in bioinformatics. The following section provides further details about median networks in phylogenetic analyses. 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.

[1] Because this conference is dedicated to FCA, this paper does not provide an introduction to FCA. Information about FCA can be found, for example, on-line (http://www.fcahome.org.uk) and in the main FCA textbook by Ganter & Wille (1999).

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

Local Proceedings in ISBN 978–84–695–5252–0,
Universidad de Málaga (Dept. Matemática Aplicada), Spain.

2 Median networks and phylogenetics

This section provides a very brief introduction to the application area of this paper[3]. Unfortunately, many of the papers in this application area are written for biologists and do not contain mathematically precise definitions of the terms and algorithms. Median graphs are undirected graphs where any three vertices have a unique median which is a vertex that belongs to shortest paths between any two of the three vertices. Examples of median graphs are trees or the Hasse diagrams of distributive lattices if considered as undirected graphs. Median networks are special kinds of median graphs where vertices represent species and parallel edges represent possible genetic changes.

In the field of phylogenetics, evolutionary trees are inferred from observed characteristics of species. Although DNA sequences can be of four values (A, G, C or T), it is unusual for more than one change to occur at the same site in a set of closely related species. Thus, characteristics can be considered binary by only recording whether or not a change occurred. In the case of parallel mutations (or the more rare reversals), it is difficult to know the sequence of the mutations. A median network summarises possible evolutionary trees. In particular, one is interested in "most parsimonious trees" which means that the number of times the endpoints of a tree edge have different values is minimal. Without parallelisms or reversals a median network is a tree. Considering the examples by Bandelt et al. (1995 and 2000), ordinary data sets tend to contain at least some parallelisms. Thus the generated median networks are not usually trees.

A median network is guaranteed to contain all most parsimonious trees (Bandelt et al., 1995). But 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 (where "weight" and "frequency" are defined as follows). In order to construct a median network, one summarises all changes that occur simultaneously with respect to a set of sample species as "weight". Graphically this can be represented by the length of edges. 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.

3 Modelling with FCA

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. Latent species (which are implied by the data but for which no specimen have been found and which are "latent

[3] Based on Bandelt et al. (1995 and 2000), Sykes (2001) and the Wikipedia page on "Median graphs".

vertices" in a median network) correspond to concepts that do not contain objects in their contingent but are the join of object concepts. Each meet-reducible concept in the lattice corresponds to a choice point between different possible trees.

An example is the mitochondrial data from Ward et al. (1991) which was also used by Bandelt et al. (1995). As mentioned above the data can be reduced to a single-valued context by using mitochondrial lineages as objects and sites of changes as attributes. Figure 1 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 are maximal trees. In Figure 1 one clique/tree contains all attributes except 16243 and another clique/tree contains all attributes except 16294 and 16239. These are the only two trees in Figure 1. Bandelt et al. describe a fairly complicated algorithm for deriving a median network using cliques, peripheral elements and torsos where the "torso" data matrix consists of the non-compatible attributes.

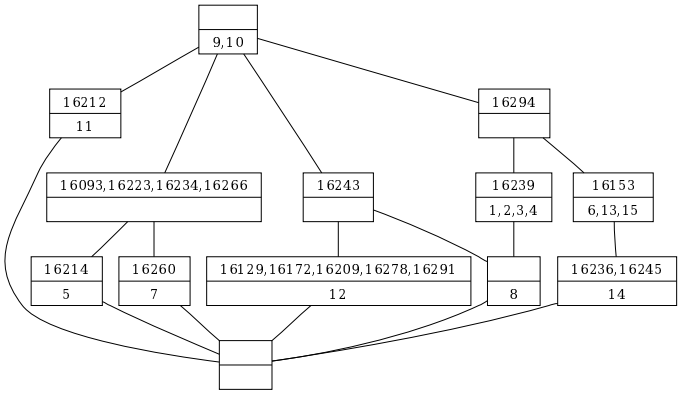

Fig. 1. Concept lattice for HVS I data of Vigilant used by Bandelt et al. (2000)

Figure 2 shows a median network for the data in Figure 1. 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 1 and the median network contain essentially the same information apart from the fact that the lattice contains a bottom node and the median network contains a latent vertex in the torso (to the right of the vertex "8") which is due to the shortest path condition of median networks but not needed for lattices. Although we are not providing a formal proof at this point, based on similar construction algorithms it is to be expected that in general the concept lattice and the median network of a data

table contain the same information apart from some latent vertices and the bottom node in the lattice.

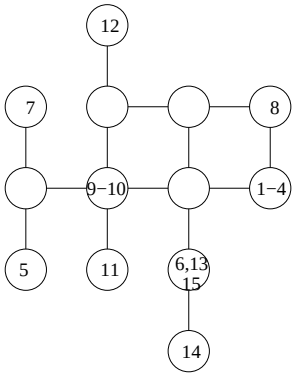

Fig. 2. The median network for Figure 1

4 Conclusion

The aim of this position paper is to stimulate further research into the application of FCA in the bioinformatics domain. It appears that FCA can improve on methods that are currently used in that area and can be used to derive a more consistent and precise terminology. Furthermore, from an FCA view, this application domain raises questions about using frequencies, weights, the construction of latent objects, tree embeddings and attribute splitting which could lead to future FCA research.

References

1. 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.
2. 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.
3. Ganter, Bernhard; & Wille, Rudolf (1999). *Formal Concept Analysis. Mathematical Foundations.* Berlin-Heidelberg-New York: Springer.
4. Sykes, Bryan (2001). *The seven daughters of Eve.* Bantam Press.
5. 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.