Translation of the Genealogical Tree of Queen Victoria into a Bayesian Network

Maximo Carreras, Mike Cincotti, Joanna Jeneralczuk and Toni M. Rath

Abstract

We model the inheritance pattern of the hemophilia disease in a family tree by a Bayesian network. The genetic principles of hemophilia are introduced and mapped into a statistical framework. We give several examples on how to use the *Bayesware Discoverer* system to query the health status of individuals given some evidence. The data used in this project was obtained from [Aronova & Herreid 2001].

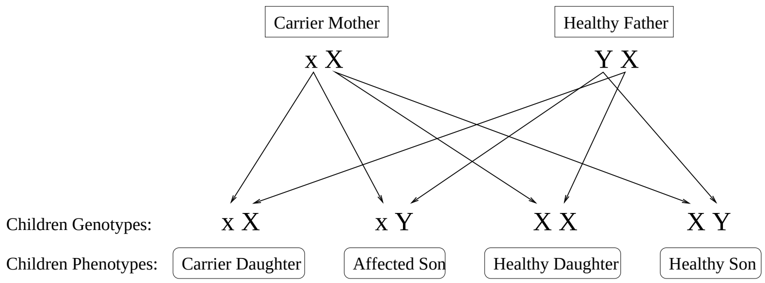
1 Introduction

Hemophilia is a genetically inherited, X-linked bleeding disorder characterized by the inability to properly form blood clots. Due to the genetic pattern of inheritance, the disorder affects males much more frequently (1 in 10,000) than females (1 in 100,000,000).

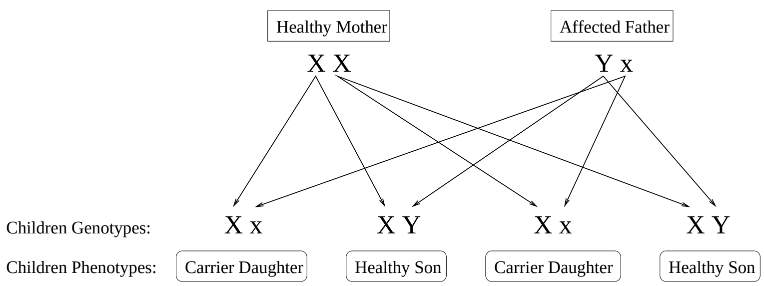
A female has two X chromosomes (XX) and a male has one X chromosome and one Y chromosome (XY). Since the blood clotting gene is carried on the X chromosome, a male with a defective X chromosome will be affected by the disease. A female with one defective X chromosome, on the other hand, will usually have normal blood clotting because the second (healthy) X chromosome will compensate. The woman will be a carrier of the defective gene but she will not show any symptoms. This fact will be discovered if some of her children are hemophiliacs. Hemophilic women are rare because it takes two defective X chromosomes in order for this condition to be seen.

Inheritance of Hemophilia: A male can only inherit the disease from his mother, because his father gives him the Y chromosome. A female can inherit a defective X chromosome from either one of a her parents. If her father is affected, that is a definite indicator that she is a carrier. However, about one third of all cases of hemophilia occur with no previous family history. These cases are assumed to be the result of genetic mutation.

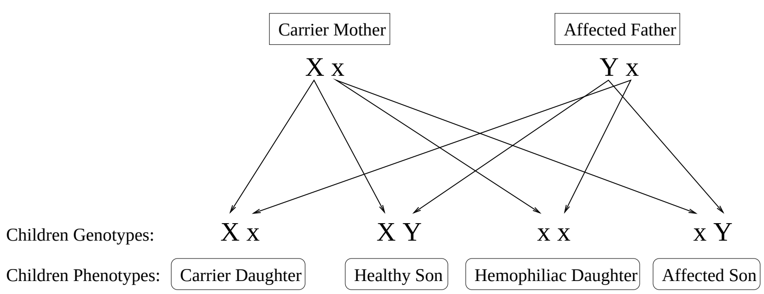
Using the above analysis, children can inherit the following genotypes/ phenotypes (X=healty chromosome, x=defective chromosome):



There is a 1 in 2 chance that the hemophilia gene will be passed on and when this happens a male child will have hemophilia, and a female will be a carrier. In other words, there is a 1 in 4 chance that the child will be male with hemophilia.



In this case each daughter will be a carrier and each son will be not affected.



There is a 1 in 2 chance that a son will have hemophilia, each daughter will be either carrier or hemophiliac.

In the case when both parents do not carry a defective gene, hemophilia can not be passed to their children unless the mutation of the genes can occur. We are going to use these analyses to define the conditional probability distribution of the nodes in the Bayesian network.

2 Data Source

This project focuses on the hemophilia present in the genealogical tree of Queen Victoria's family. Hemophilia entered the royal family through Queen Victoria, who somehow became a carrier for the disease. By the inheritance factors described above, hemophilia then moved down through the generations of the royal family.

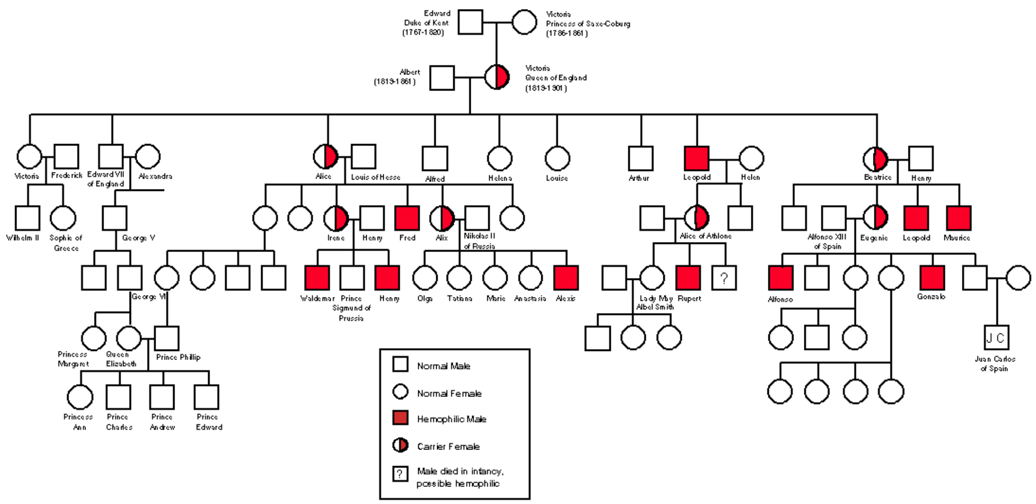


Figure 1: Family tree of Queen Victoria.

A genealogical tree containing all of the known information about which descendants did inherit hemophilia can be found in [Aronova & Herreid 2001]. The tree contained in that document, shown in figure 1, provided the data used in this project.

3 Modeling the Genealogical Tree

Using the *Bayesware Discoverer* program, we modeled the information about hemophilia in the royal family with a Bayesian network. Each member of the family has its own node in the network, with directed links from the parents to their children. The following is an example of the network's representation of this parental relationship.

The inheritance properties of hemophilia are then translated into the conditional probability distribution tables, based on the sex of the child. For men, the table looks like this:

Mother	Father	Affected	Not Affected
Carrier	Affected	0.500	0.500
Not Carrier	Affected	0.000	1.000
Carrier	Not Affected	0.500	0.500
Not Carrier	Not Affected	0.000	1.000

For women, we used this table:

Mother	Father	Carrier	Not Carrier
Carrier	Affected	1.000	0.000
Not Carrier	Affected	1.000	0.000
Carrier	Not Affected	0.500	0.500
Not Carrier	Not Affected	0.000	1.000

People who married into the royal family presented a problem because we knew nothing about their parents. Though the probability that a person taken at random is either affected by hemophilia or a carrier of the disease is quite low (various sources provide different figures), we assume no knowledge of these probabilities. Thus, any node in the tree without its parents shown is given a prior marginal probability distribution: 50% chance of either being a carrier (for women) or affected (for men).

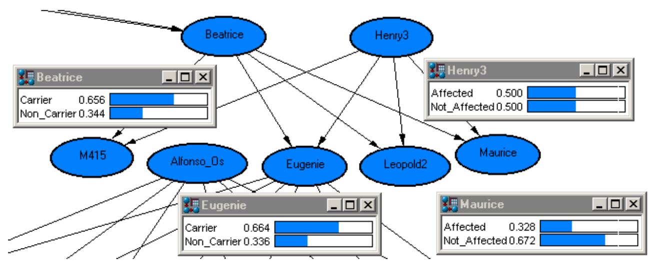
4 Question Answering

The network structure that was set up in section 4 models the way in which the disease is genetically transmitted from one generation to the next. This was achieved by providing the conditional probability distribution of each node given its parents. Nodes with no parents are assigned prior marginal probability distributions. Based on this information, the system can compute any other prior probability values.

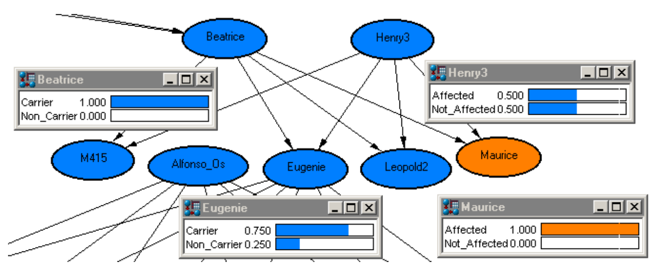
The task is now to compute the posterior probability distribution for a query variable, given exact values of some evidence variables. The flexibility of Bayesian Networks allows any node to serve as either a query or an evidence variable.

Here are some examples:

Example 1: Knowing that Maurice is affected does not change our prior belief (see figure 2(a)) about his father Henry3, who gives his son the Y chromosome. This implies that the mother Beatrice must be carrier of the disease. Also, with Maurice being affected, his siblings have a greater probability of being either carrier or affected (see figure 2(b)).



(a) No evidence supplied,



(b) with evidence.

Figure 2: Simple example (output obtained with *Bayesware Discoverer*).

Example 2: Knowing that Edward_P is affected implies that his mother Elizabeth_Q is a carrier, since Edward_P inherits the Y chromosome from his father Philip_P. Therefore, if we set Mother_QOE and George_VI (Elizabeth_Q's parents) to non-carrier and not affected respectively, we arrive at a contradiction. The system does not produce an error message. Instead, it assigns invalid posterior probability distributions to conflicting nodes (see figure 3).

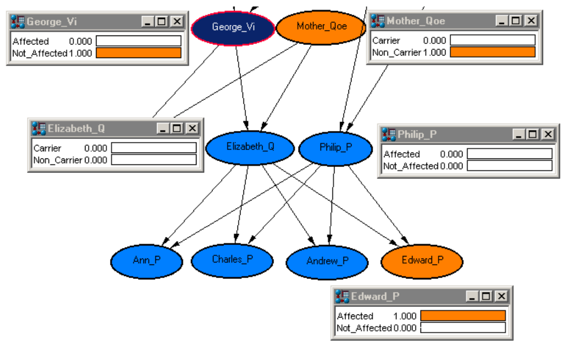
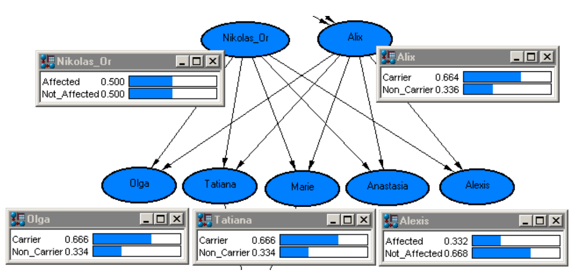


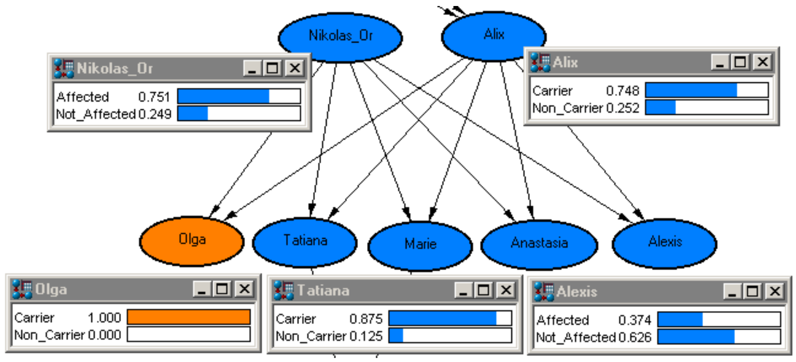
Figure 3: Program output for an example with contradictory evidence.

Example 3: Assuming that Olga is a carrier increases Alexis's probability of being affected, because we know that the disease runs in the family (see figure 4(b)). Notice how the probability of Alexis's father Nikolas of Russia of being affected increases considerably but the one of his mother Alix of being a carrier only slightly. If we further assume that Tatiana is a carrier then Alexis's probability of being affected decreases (figure 4(c)). The explanation for this strange behavior lies in the fact that it is more likely for a female child to inherit hemophilia from her father. Therefore, the more female children who are carrier the greater the chance that the father is the transmitter. That implies that the probabilities of Alix and Alexis of being a carrier and affected respectively decrease. Finally, setting Anastasia to non-carrier makes it impossible for the father to be the transmitter and therefore the mother is a carrier (figure 4(d)). As a consequence, Alexis has a 50% chance of being affected.

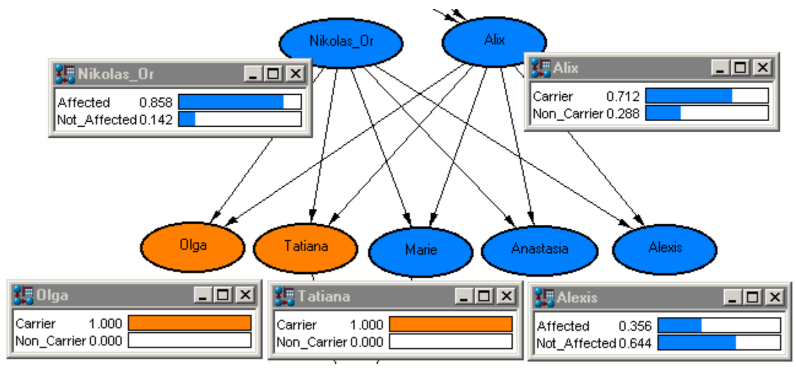


(a) No evidence supplied,

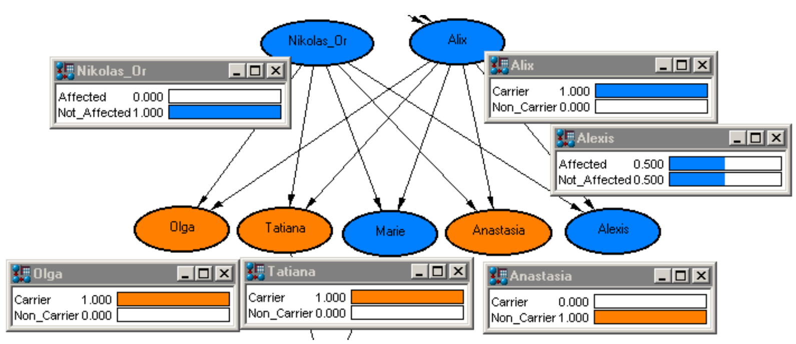
Figure 4: Example involving complex reasoning (continued on next page).



(b) evidence *Olga is carrier*,



(c) further evidence *Tatiana is carrier*,



(d) further evidence *Anastasia is non-carrier*.

Figure 4: Example involving complex reasoning (continued).

5 Conclusion

It is quite simple to map a genealogical tree into a Bayesian network using the *Bayesware Discoverer* system. The program has a friendly graphical user interface that allows simple querying and supplying of evidence. It does not require the use of a programming environment. Also, Bayesian network data files are stored in human–readable form, which allows bulk editing or automated processing.

Bayesian networks only require the user to provide knowledge about the dependencies between directly related nodes in order to encode complex situations as a whole. Evidence propagates through the network in any direction from where it was supplied. This makes it possible to answer questions about ancestors given descendants and vice–versa.

References

[Aronova & Herreid 2001] Y. Aronova–Tiuntseva and C. F. Herreid: *Hemophilia: "The Royal Disease"*. Case study; available online from the National Center for Case Study Teaching at Buffalo, NY at http://ublib.buffalo.edu/libraries/projects/cases/hemo.htm.

[Russell & Norvig 1995] S. Russell and P. Norvig: *Artificial Intelligence - A Modern Approach.* Prentice Hall, Upper Saddle River, NJ, 1995.