Probabilistic Reasoning in the Domain of Genetic Counseling

Nomi L. Harris

October 31, 1989

This work has been supported in part by National Institues of Health grant 1 R01 LM 04493 from the National Library of Medicine.

Keywords: Uncertainty, probabilistic reasoning, belief networks, genetic counseling, cycles, clustering, conditioning

Probabilistic Reasoning in the Domain of Genetic Counseling

Nomi L. Harris

This report is a modified version of a thesis submitted to the Department of Electrical Engineering and Computer Science in May, 1989, in partial fulfillment of the requirements for the degree of master of science.

Abstract

This paper describes a program, GENINFER, which uses belief networks to calculate risks of inheriting genetic disorders. GENINFER is based on Judea Pearl's [17] algorithm for fusion and propagation in probabilistic belief networks. These networks allow the effects of various pieces of information to be propagated and fused in such a way that, when equilibrium is reached, each proposition can be assigned a degree of belief consistent with the axioms of probability theory.

GENINFER takes as input pedigrees of families affected with genetic disorders, as well as supplementary phenotypic information. Other factors that can affect the inheritance of genetic disorders, such as population frequency and mutation, are also taken into acount. GENINFER can handle diseases with incomplete penetrance or age-dependent expressivity. GENINFER'S output consists of genotype probabilities for all family members and estimated genetic risks for prospective children.

Pearl's basic algorithm cannot directly handle multiply-connected networks, which arise in the genetic counseling domain whenever a family pedigree includes consanguinity or more than one child per couple. GENINFER makes use of two cycle breaking methods, clustering and conditioning, to handle these situations.

Research Supervisor: Peter Szolovits

Contents

List of Figures

provide one such mechanism.

A belief network consists of a set of nodes, which represent propositions or variables, connected by directed links, which represent direct relationships between the nodes. Belief networks allow the impacts of various pieces of information to be propagated and fused in such a way that, when equilibrium is reached, each proposition can be assigned a degree of belief consistent with the axioms of probability theory. Judea Pearl's [17] algorithm for fusion and propagation in probabilistic belief networks propagates information through a network by means of messages passed between nodes.

I have implemented a system for genetic counseling, GENINFER, which is based on Pearl's method. This thesis describes how I adapted Pearl's method for use in the genetic counseling domain, and how I supplemented the basic algorithm, which can handle only singly-connected networks, with techniques for handling multiply-connected belief networks.

A description of any family with a single-gene inherited defect (which may be recessive, dominant, or X-linked) can serve as input to GENINFER. The family description is converted to a probabilistic belief network, through which all relevant information can be propagated in order to arrive at a belief distribution for the genotype of each individual. Additional data pertaining to the specific disorder and the possible phenotypes of family members may also be entered; all data is fused in a manner consistent with probability theory. Conditioning, which is a way of dealing with multiply-connected networks, is used for families in which there is consanguinity (marriage between relatives). Clustering is used in order to prevent cycles in families with multiple children. The output of GENINFER is an assessment of the probabilities of each possible genotype for each person in the family, and a risk estimate for future offspring of the consultand (if a consultand is specified).

to a particular gene; it can be described by specifying an unordered pair of alleles. The phenotype is the physical manifestation of the genotype by the organism. For unilocal traits, the possible genotypes are homozygous normal (both alleles normal), heterozygous (one normal allele, one defective), and homozygous affected (both alleles defective). The set of possible phenotypes is usually {*affected, unaffected*}, although for some disorders the affected phenotype may vary in degree of severity.

There are a number of different inheritance patterns by which genes may be passed to descendants. The most common inheritance patterns for unilocal traits are recessive, dominant, and X-linked. A *recessive* trait is not observable in the phenotype unless it is present on both alleles. A person who is heterozygous for a recessive trait will not exhibit the disorder, but will be a *carrier* for that trait. A *dominant* trait is exhibited if it is present at either one or both of the alleles; there are no carriers for dominant traits. An *X-linked* trait is controlled by a gene on the X chromosome. Since males have only one X chromosome, they cannot be heterozygous for an X-linked trait; they either have the defective gene on their single X chromosome (making them *hemizygous* for the trait), or they are unaffected.

1.3 Genetic Counseling

Genetic diseases account for a large proportion of birth defects. People with a family history of a genetic disorder may be concerned about the risk that future children will suffer from the disorder. The role of a genetic counselor is to assess a consultand's risk of passing on a genetic disorder and offer advice on the best course of action. Often, consultands will be relieved to hear that their risk of having an affected child is quite low, and they can proceed with their plans to raise a family. Sometimes the genetic counselor might recommend amniocentesis, which is a technique for collecting a few fetal cells from the uterus of a pregnant woman so that they can be tested for genetic defects.

When implementing an AI program in a particular domain, it is helpful to have the advice of an expert in the domain. The domain expert who advised me was Dr.

1.3.2 How GENINFER can aid genetic counselors

The Bayesian calculations that must be performed in order to advise consultands about their probable risk can be quite complex. However tempting it may be to the genetic counselor to neglect these calculations, it is essential to perform them correctly and completely in order to give consultands an accurate assessment. As Edmond Murphy, a proponent of Bayesian methods in genetic counseling, phrased it,

> There can be no doubt but that an exhaustive analysis of a pedigree, even when the mode of inheritance is simple, may itself be complicated. In the practical situation, the ideal method may not be applied because the counselor either becomes lost in the logic or finds the method tedious... I suggest that if they cannot find the time to do the calculations themselves, they should delegate the job to someone else. ([13], p. 396)

Murphy may not have had a computer in mind when he suggested delegating the arduous calculations to "someone else," but in many respects a computer is the ideal entity for such tasks.

GENINFER is not intended to deprive genetic counselors of their jobs; it does not cover every facet of the genetic counseling process. For example, many of the people who consult a genetic counselor are older women concerned about the risk of having a child with Down's syndrome; pedigree analysis is usually not an important factor when addressing this concern. For cases that fall within GENINFER's capabilities, however, the answers it gives are compatible with those provided by the domain experts. Section 7.2 discusses some extensions that might make GENINFER more useful to genetic counselors.

The extensional approach treats uncertainty as a truth value attached to a formula, and regards the uncertainty of a given formula as a function of the uncertainties of its subformulas. Many rule-based or production systems, such as MYCIN, follow the extensional approach. In the intensional, or *model-based* approach, uncertainty is attached to states of being or subsets of possible worlds. MUNIN [15] is an example of an expert system that uses the intensional approach. In general, extensional systems tend to be computationally efficient but semantically sloppy, while intensional systems are semantically clear but computationally expensive [18]. Much research in uncertainty has focused on attempting to reconcile the tradeoff between semantic clarity and computational efficiency.

Bayesian inference and belief networks, which will be discussed in sections 2.2 and 2.3, are tools that can be used to construct intensional systems. Belief networks clarify the semantics by making causal relationships specific.

2.2 Bayesian Inference

Bayesian inference is a mechanism, based on the use of conditional probabilities, for reasoning under uncertainty. If we want to calculate the probability of an event A, for example, we can take the weighted sum of the probabilities that A occurs, conditional on a set of exhaustive and mutually exclusive events B_i: $P(A) = \sum_i P(A|B_i)P(B_i)$. The conditional belief of a hypothesis H given a piece of evidence E can be calculated with *Bayes' rule*: $P(H|E) = \frac{P(E|H)P(H)}{P(E)}$.

Bayes' rule can be viewed as combining predictive and diagnostic support. Defining the *prior odds* on H as $O(H) = \frac{P(H)}{P(\neg H)} = \frac{P(H)}{1-P(H)}$ and the *likelihood ratio* as $L(E|H) = \frac{P(E|H)}{P(E|\neg H)}$, the *posterior odds* of H given E, $O(H|E) = \frac{P(H|E)}{P(\neg H|E)}$ are given by the product $L(E|H)O(H)$. The prior odds represent the predictive support provided by the background knowledge, while the likelihood ratio represents the diagnostic support given to H by E, the evidence observed.

As an example of how Bayesian revision can be used, consider this hypothetical medical scenario. A 23-year-old woman consults a physician, complaining of fatigue,

2.3 What Are Belief Networks?

Belief networks are a graphical representation that allow probabilistic techniques such as Bayesian updating to be applied to a system of dependent variables. Belief networks (also called Bayesian networks, inference nets, or causal nets) consist of a set of nodes connected by directed links. The nodes represent propositions or variables, and the links represent direct relationships between the nodes. The relationships may be causal, but they are not limited to this interpretation. Belief networks allow the impacts of various pieces of information to be propagated and fused in such a way that, when equilibrium is reached, each proposition can be assigned a probability or degree of belief consistent with the axioms of probability theory [17]. This is possible because of the explicit representation of conditional independence between the variables. The absence of an arc from a node x to a node y implies that y is conditionally independent of x, given the values of the predecessor nodes of x [6].

As an example of how belief networks are constructed, consider this simplified medical scenario. A 45-year-old woman, complaining of abdominal pain and severe diarrhea, consults a physician. These symptoms could be caused by a disease called ulcerative colitis, but there are other possible diagnoses, such as amoebic infection. Amoebic infections are rare in the United States but are more common in certain other countries. When asked whether she has been out of the country recently, the patient replies that she visited Mexico a few weeks ago. This evidence gives support to the hypothesis that the patient's symptoms are due to an amoebic infection. Although ulcerative colitis and amoebic infection are not causally connected to each other, and although the patient could conceivably be suffering from both conditions, increased belief in the amoebic infection hypothesis "explains away" the evidence of severe diarrhea and has the effect of weakening the physician's belief in the ulcerative colitis hypothesis.

This scenario can be represented by the belief network shown in figure 2.1. The four variables—abbreviated as *diarrhea, Mexico, colitis,* and *amoebic*—are represented by nodes in the network. The links, in this case, represent causal relationships

Shachter's method involves removing nodes from the influence diagram by performing value-preserving reductions. For example, "barren" nodes—those with no successors—can be removed from the diagram. Other manipulations allow us to eliminate certain chance and decision nodes or to reverse arcs. Each of these operations changes the conditional probabilities of the nodes without changing the underlying probability distribution of the influence diagram. If an influence diagram is regular, then each step of the algorithm removes at least one node, so the algorithm will always terminate with a single value node remaining.

2.4.2 Pearl

Pearl's method for fusion and propagation in belief networks ([17], [18]) uses local message passing to communicate information between nodes. Messages received by nodes are combined in a manner consistent with Bayesian theory. The probabilistic relationships between the nodes are specified by conditional probability matrices. If the network is singly connected, the parameters reach equilibrium (meaning that all information has been communicated to all nodes in the network) in time proportional to the length of the longest path in the network. Multiply-connected networks must be handled specially to avoid infinite cycling of information around a closed loop. Pearl's method is described in greater detail in chapter 4; methods for dealing with cycles are discussed in chapter 5.

2.4.3 Lauritzen & Spiegelhalter

Lauritzen and Spiegelhalter's [11] method for absorption and propagation of evidence in belief networks is based on topologically manipulating the networks and using a range of local representations for the joint probability distributions. The problem of loops in multiply-connected networks is avoided by clustering the nodes into maximal connected components, or *cliques*. A clique is defined as a set of nodes such that each node in the set has an arc to all other nodes in the set. Clique potentials are conditional probabilities defined on cliques.

collecting terms involving nodes in each clique, removing cliques one at a time. In general, the procedure when i cliques remain is to transform the evidence potential of C_i, $\psi(C_I)$, to $p(R_i|S_i) = \psi(C_i)\sum_{R_i}\psi(C_i)$, and then to multiply the potentials for C_k, a parent clique of C_i, by $\sum_{R_i}\psi(C_i)$. The node probabilities can then be obtained by chaining back through the graph and using the conditional probability tables [11].

Notice that when we transform the potential of C_i, we also change the potential of its parent clique. This is the mechanism by which information is propagated through the graph.

Peeling

The Lauritzen & Spiegelhalter method is related to the peeling method of Cannings et al., which is described in [26]. The peeling process exploits the conditional independence properties expressed by the graph in order to successively "peel" the graph down to the nodes of interest [21]. At each stage in the peeling, there is a "cutset" that divides the graph into two disjoint, independent components: the peeled set (which includes nodes whose information has already been fully incorporated) and the unpeeled set. For each cutset, there is a probability function called an *R function* which encapsulates the information in the peeled set. The peeling method is based on the relationship between R functions on successive cutsets, which derives from the property of conditional independence [26].

There are many similarities between Lauritzen and Spiegelhalter's method and the peeling procedure. The peeled nodes correspond to members of cliques of higher order in the set chain. The cutsets on which R functions are defined are the clique separators through which evidence is propagated. One difference between the two approaches is that the Lauritzen & Spiegelhalter procedure, unlike peeling, chains back through the network to obtain marginal distributions on the individual nodes [21].

Chapter 3

Previous Approaches to the Genetic Counseling Problem

3.1 Bayesian Approaches to Calculation of Genetic Risks

Bayesian techniques can be used to calculate genotype probabilities for individuals in a family at risk for a genetic disorder. Unlike non-Bayesian approaches, which consider only positive information, Bayesian inference allows all of the information in the pedigree, both positive and negative, to be taken into account. This often has the effect of lowering our estimate of the probability that a consultand's future children will be affected with the disorder in question.

Consider the pedigree shown in figure 3.1, in which Betty's two brothers are both affected with hemophilia. Betty is concerned that her next son might be hemophilic. She would like to know the probability that this will occur. A naive calculation of the risk to Betty's next son would yield the incorrect estimate of 0.25 by the following reasoning: There is a 0.5 chance that Betty is a carrier, since her mother has one defective allele and one normal one. If she is a carrier, each of her sons has a 0.5 probability of being affected. The value of 0.25 is obtained by multiplying 0.5 and 0.5. However, this calculation ignores an additional piece of information provided

form of retinitis pigmentosa (which causes blindness). Since Daphne's uncle Clifford is affected, Daphne is worried that she might be a carrier for retinitis pigmentosa. If she were a carrier, each of Daphne's sons would have a 50% percent risk of being affected. A naive calculation of the risk to Daphne's prospective son would give the overly pessimistic probability of 1/8, or 12%. The use of Bayesian probabilities revises our assessment of risk to Daphne's prospective son, lowering it to only 2%.

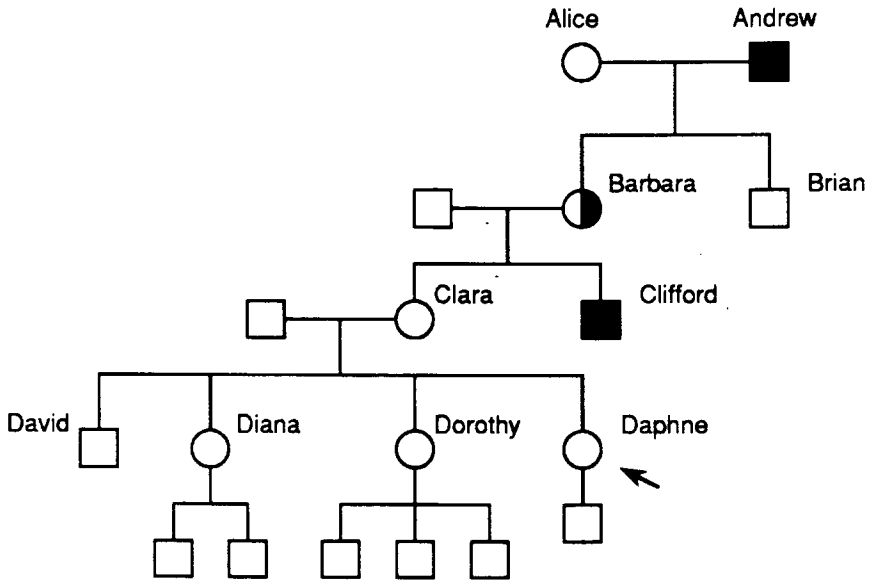


Figure 3.2: Pedigree for family affected with X-linked retinitis pigmentosa

In order to determine Daphne's probability of being a carrier, we must first calculate the probability that her mother, Clara, is a carrier. The table below shows the calculation of Clara's probability of being a carrier [14]. Notice that information

	Betty is a carrier	Betty is not a carrier
Prior probability	0.5	0.5
Conditional probability (one normal son)	0.5	1
Joint probability	0.25	0.5
Posterior probability	.25/(.25 + .5) = .33	.67
Risk to next son	0.17	0

Table 3.1: Calculation of risk to Betty's future children

3.2.1 PEDIG

One early program, PEDIG [9], was written in FORTRAN. It uses only the information in the pedigree; no other data can be considered. PEDIG also suffers from an inability to process families in which there is consanguinity.

3.2.2 GENEX

The GENEX processor, by J. Hilden [10], derives probability formulas, rather than numerical answers, for problems involving inheritance of qualitative traits. In order to use a given pedigree as input, the pedigree must first be broken down by hand into atomic assumptions described in terms of probabilities. The formulas derived by GENEX may, according to Hilden, "provide valuable insight into certain areas, notably in statistical genetics"; however, they are not likely to be of much practical value to genetic counselors.

3.2.3 Prokosch et al.

Prokosch, Seuchter, Thompson, and Skolnick [19] used a commercially available expert system shell (*Intelligence/Compiler*) as the basis of two prototypes of an expert system for human genetics. One approach investigated by Prokosch and his colleagues is object-oriented: family relationships are represented by three frames (KINDRED, INDIVIDUAL, and MARRIAGE). The other approach, fact-based pedigree representation, was favored as being "more readable and easier to program" [19]. In this representation, parent-child relationships are described by Prolog-like statements such as "X is-mother-of Y." Forward-chaining rules must be added to allow the system to deduce other family relationships. For example, to assert the relation "grandfather," the following rule is used:

```
If X is-father-of Z and
   Z is-parent-of Y
then
```

such as age-dependent expressivity. However, Spiegelhalter's method is probably capable of handling the same cases as GENINFER; the decision to use Pearl's algorithm for GENINFER was made before Spiegelhalter's method was published. Section 7.1 discusses the relative merits of Spiegelhalter's method as compared with Pearl's.

Chapter 4

Pearl's Method

4.1 Propagation and Fusion in Singly-Connected Belief Networks

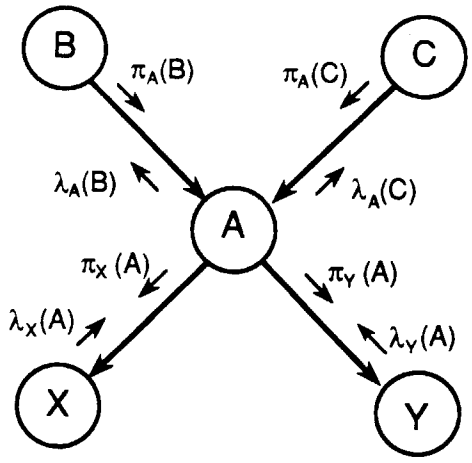


Figure 4.1: Message-passing in a belief network

Each node in a belief network may contain evidence or information, which we wish to propagate to all other nodes in the network in order to calculate a belief distribution for the network. The propagation of information through the belief network is accomplished by means of messages sent between nodes by two parameters, π and λ. π is a vector representing causal support from a node's ancestors, while λ represents diagnostic support from a node's descendants. Each node contains a conditional probability matrix, which characterizes the relationship between the node and its parents.

4.1.1 Calculating beliefs

The belief in a hypothesis depends on three parameters: the strength of the causal support for the hypothesis, the strength of the diagnostic support for the hypothesis, and the conditional probability matrix. Consider the portion of a belief network shown in Figure 4.1 [17]. We are interested in calculating the belief in each possible hypothesis for variable A. Variables B and C are causally related to A, and A is causally related to its children X and Y. Each link is labeled with two dynamic parameters, π and λ, which encode the messages sent between a pair of nodes. $\lambda_A(B)$, for example, represents the message sent from node A to its parent node, B.

After the influence of all data has been propagated through the network, the

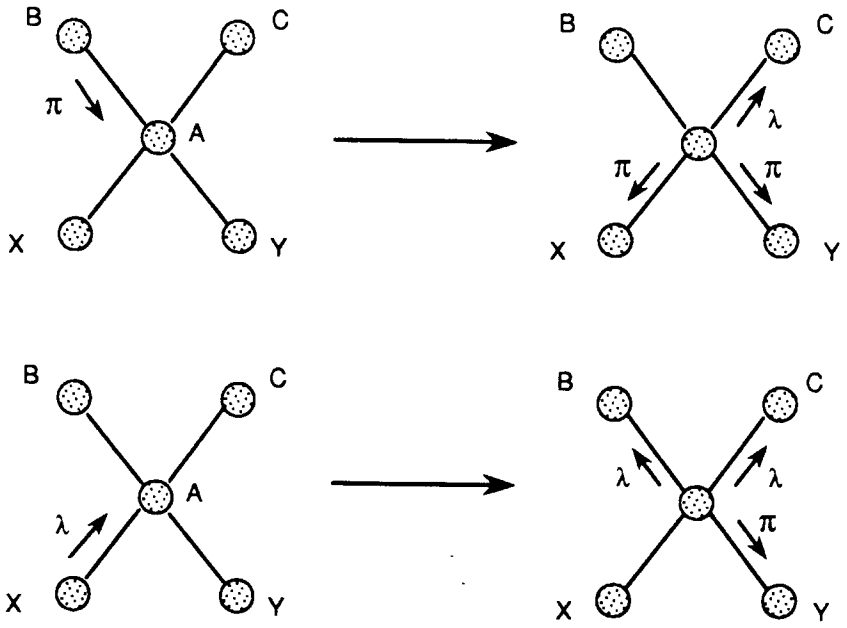


Figure 4.2: Propagation of updates

there are no cycles in the network. Note that Pearl's algorithm is distributed: messages are passed between nodes, not through any central control. When the network has reached equilibrium, the beliefs in the possible hypotheses, conditional on the available evidence, can be obtained by using the fusion equation described in section 4.1.1.

4.2 Applying Pearl's Algorithm to the Genetic Counseling Problem

Pearl's algorithm extends the idea of Bayesian revision to an arbitrary network, such as a pedigree. I have adapted and extended Pearl's general-purpose method for use in the domain of genetic counseling.

man with genotype j will have a child with genotype k. Note that $\forall i, j \sum_k M_{i,j,k} = 1$, where i, j, and k range over the values $\{affected,\ heterozygous,\ normal\}$, since every child must have one of the three genotypes.

Table 4.1 shows a conditional probability matrix for a male individual in a family at risk for an X-linked genetic disorder.

	Mother and father								
Male child	**AA**	**AH**	**AN**	**HA**	**HH**	**HN**	**NA**	**NH**	**NN**
A	1	0	1	0.5	0	0.5	0	0	0
H	0	0	0	0	0	0	0	0	0
N	0	0	0	0.5	0	0.5	1	0	1

Table 4.1: Probability matrix for X-linked disorder

4.2.2 Initializing the parameters

When applying Pearl's method to a specific domain, the initialization of the parameters is the aspect that requires the most modification. Before we begin the propagation process, the π and λ parameters must be initialized to reflect the available evidence. Only links leading to root nodes (i.e., those with no ancestors in the pedigree) are assigned initial πs, and only links leading to leaf nodes are assigned initial λs. The parameters on the other links are calculated during the propagation phase of the algorithm.

Evidence pertaining to individuals' genotypes can often be obtained by considering their phenotypes. This evidence is represented by attaching a dummy leaf to each person node, with the λ on the link set to represent what we know about the person. In effect, the dummy leaf represents the phenotype of its parent, while the parent itself represents the genotype. Figure 4.4 shows the belief network for Betty's family, with dummy leaves added to represent the phenotype of each individual.

Each value λ_i in the initial λ vector represents $P(phenotype|genotype_i)$, or the probability that we would see the observed phenotype if the genotype of the person were i. For example, for a person affected with a dominant disorder, the initial λ

inferences should result." When I experimented with changing the value of p for a particular family, the calculated beliefs varied only slightly.

4.2.3 Propagation

In the intuitive view of the propagation phase of Pearl's algorithm, π and λ messages are passed between nodes. In my program, this message-passing is accomplished by assigning π and λ vectors to the links between nodes, and updating the vectors to reflect the transmission of new information. Once the initial values for some of the parameters have been set, other parameters are put on a queue to await updating. The exact order in which links that are on the queue are updated is not important, as long as we make sure that each parameter is not present in the queue more than once at a given time.

The propagation procedure takes a π or λ parameter off the queue and updates it with the fusion equations. When a parameter is updated, we compare its new values to the values that were present before the update. If the difference between the old values and the new values is small enough to be attributable to roundoff error, we move on to the next item on the queue. Otherwise, the parameters dependent on the newly updated parameter must be put on the queue. If the λ on a link changes, we must update the λs of the links from the parent to the grandparents and the πs for the siblings of the child. When a π vector is updated, we must update the λ on the link from the child to the child's other parent and the πs on the links from the child to the child's children. All person nodes have at least one "child": the dummy leaf node.

One special case that was not mentioned by Pearl applies when the π vector on a link to a root node is being updated. (This will occur when the λ of another child of the root is updated.) Instead of multiplying the summation of the weighted probabilities of the grandparents by the λ vectors of the sibling nodes of the child in order to obtain the new π, we must multiply by the prior probabilities for the parent:

4.3 Advantages of Pearl's Method over Murphy & Chase

The methods described by Murphy and Chase [14] could be directly implemented, but using Pearl's algorithm has a number of advantages over that approach. Unlike the case-specific methods described in Murphy & Chase, Pearl's method is robust and generalizable. Murphy & Chase describe separate procedures for different types of families and different inheritance patterns. With Pearl's method, there is no need to approach different genetic counseling cases differently, nor is it necessary to specify a consultand: *all* available information is propagated to *all* nodes in the belief network. Information outside of the pedigree itself, such as the results of enzyme tests, can be incorporated orthogonally, without disrupting the structure of the underlying family network (see Section 6.4). This supplementary information is automatically fused with the pedigree data to yield correct combined probabilities.

The methods in Murphy & Chase rest on the assumption that no unaffected individual is a carrier unless he or she has affected offspring; however, because this assumption is not explicitly specified, it cannot be adjusted. It is difficult to consider population risks when using the methods in Murphy and Chase. The background risk of a disorder can be specified as input to GENINFER, which allows it to take advantage of increased knowledge about the prevalence of the disease in the population of interest. For example, if a consultand belongs to an ethnic group known to have a higher incidence of the disease in question, this information can be taken into account by the system. Pearl's method also allows penetrance probabilities to be incorporated in a straightforward manner (see section 6.1). The possibility that an instance of a disorder has been caused by a new mutation can be covered by altering the conditional probability matrices (see section 6.3).

A key limitation of many programs or procedures for calculating genetic risk is that they cannot be used on families with consanguinity. I have extended Pearl's basic algorithm to handle such families. The methods I used to handle these multiply-connected family networks are described in the next chapter.

by instantiating a selected group of variables in order to break the communication pathways.

In *stochastic simulation*, each variable is first assigned a fixed value. Each node then examines the current state of its neighbors, computes a belief distribution for its host variable, and randomly selects one value from the computed distribution. Beliefs are computed by calculating the percentage of times that each value is selected by a node [18]. Stochastic simulation is guaranteed to converge eventually on the correct belief assignment, but it generally requires a very long relaxation period before it reaches a steady state [17]. However, Chavez and Cooper [4] have constructed an algorithm that efficiently approximates the solutions to belief networks by means of stochastic simulation. The running time of their algorithm does not increase exponentially with the number of loop-cutset nodes.

Olesen et al. [15] and Lauritzen & Spiegelhalter ([11], [21]) use two forms of clustering to break cycles. For each set of nodes that share a common parent or parents, an extra node is inserted between the children and the parents. All parents in a "family" then point to an intermediate node, which points to each of the children of those parents. This process of introducing intermediate nodes is referred to as "marrying nodes" by Spiegelhalter and as "divorcing multiple parents" by Olesen. In addition, both pairs of researchers use triangulation to form cliques in the multiply-connected networks. The cliques can be treated as clustered nodes, and the hypergraph formed by the cliques and the connections between them is guaranteed to be acyclic [25].

Agosta [1] has derived a closed form solution, based on clustering, for certain multiply-connected belief networks. Unfortunately, the solution is applicable only if the leaf nodes are conditionally independent, which is not the case for family networks.

Although probabilistic inference in singly-connected (acyclic) belief networks can be performed in polynomial time, probabilistic inference in multiply-connected networks has been shown to be NP-hard [7]. Therefore, it may not be possible to find

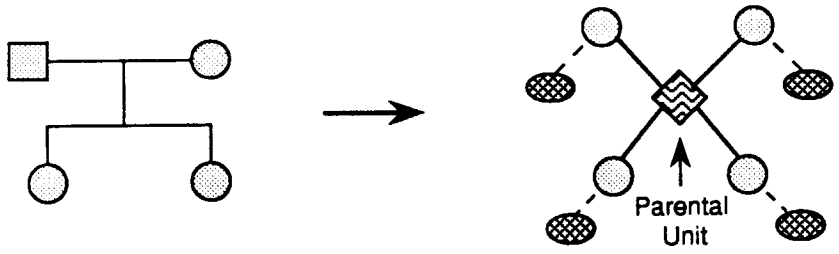


Figure 5.2: Structure of family network with parental unit added

5.4 Clustering: Parental Units

In clustering, instead of connecting each child directly to its parent, an intermediate node is introduced. I call this node a *parental unit*; Spiegelhalter [21] refers to it as a *marriage node*. The parental unit contains no new information, but rather combines the information provided by the parents and passes it on to the children. As Figure 5.2 illustrates, the addition of a parental unit breaks up the figure-eight cycle. Note that each person node must still be assigned a dummy leaf, which is connected directly to it. Because they contain no phenotypic information, parental units are not assigned dummy leaves. The parental unit structure is flexible enough to accommodate families with remarriages and half-siblings, because each person can belong to more than one parental unit.

In unclustered networks, there was only one kind of link between nodes. Links in clustered networks can be of three different types:

1. Links from person nodes down to parental units;
2. Links from parental units down to person nodes;
3. Links from person nodes to dummy leaf nodes.

The heterogeneity of the clustered networks is reflected by the π and λ vectors and the conditional probability matrices. The π and λ messages sent by person nodes will still have three entries. The messages sent by parental units, however, have 9 (3x3) elements, because they represent possible genotypes of a couple. The conditional probability matrices in the parental units have 9x3x3 entries, rather than 3x3x3; each element $N_{i,j,k}$ represents the probability that the parental unit has

[24], where $v_1 ... v_n$ are the possible values that the loop-cutset nodes can take on. $P(A|E, C_1 = v_1, ..., C_n = v_n)$ can be calculated by running Pearl's algorithm on the conditioned network. The calculation of the joint probability of the loop-cutset given evidence E, $P(C_1 = v_1, ..., C_n = v_n|E)$, will be discussed in section 5.5.3.

5.5.1 Choosing a loop-cutset

A loop-cutset must contain at least one node from every cycle in the network, with the additional constraint that a loop-cutset node may not have more than one parent in the same cycle. (If a loop-cutset node is the child to more than one other node in the loop, it will receive top-down information more than once, leading to incorrect updating.)

An ideal loop-cutset contains as few nodes as possible while still satisfying all conditions. Keeping the loop-cutset small helps to minimize the expensive operations that must be performed on it. Finding the optimal loop-cutset for a network is NP-hard, but a reasonably good loop-cutset can be found quickly (in $O(n^2)$ worst-case time complexity) by following this simple heuristic algorithm [24]:

1. Remove (or mark) all nodes that are *not* in any cycle.
2. If there are any nodes remaining, they are in a cycle, so choose a good loop-cutset candidate from the cycle (one that does not have more than one parent in the cycle.) Add this node to the loop-cutset and remove it from the network.
3. Loop back to step 1. If there are no remaining nodes in a cycle, we are done.

In practice, families tend not to have multiple cycles, so the loop-cutset will typically contain only one or two nodes.

5.5.2 Checking for cycles

The algorithm for finding the loop-cutset requires that we check for cycles in the network. In fact, we will not need to condition the network in the first place if it is free of cycles.

We can check for cycles in a belief network by using a version of depth-first search. We start with any node, and follow a link out of it to another node. Each

$$P(C_k) = \prod_{j=1}^{|C|} P(C_j = v_j),$$

where v_j is the value assigned to loop-cutset node C_j in the kth instantiation.

Calculating joint probabilities of loop-cutset instantiations

There is a problem with this formula: how do we know what $P(C_j = v_j)$ is when we can't run the propagation algorithm on the intact network? Suermondt [23] has derived a method for calculating joint probabilities for loop-cutset instantiations. First, the nodes in the network are ordered according to the "is-a-predecessor-of" relationship; this can be accomplished by a topological sort. When the nodes in a network are numbered topologically, any ancestor of a given node has a smaller number than that node. An algorithm that topologically sorts a network can be obtained by modifying the depth-first search algorithm.

The initial beliefs, or priors, are calculated for each node in order of the topological numbering, as follows: If a node has no predecessors, its prior is simply the normalized product of the π and λ vectors on the link to its dummy leaf. If a node has predecessors, we will already have calculated their priors because of the order in which we are processing the nodes. The prior for node A then becomes:

$$Prior(A_i) = \sum_{j,k} P(A_i|Mother_j, Father_k) BEL(Mother_j) BEL(Father_k)$$

The priors are used when calculating the joint probabilities of loop-cutset instantiations [23]. Let c_1 represent the probability that loop-cutset node C_1 takes on the value v_1. For each loop-cutset instantiation $[c_1, ..., c_n]$, we want to calculate $P(c_1, ..., c_n) = P(c_1)P(c_2|c_1)P(c_3|c_1, c_2)...P(c_n|c_1, ..., c_{n-1})$.

The joint probability of the loop-cutset instantiation is calculated as follows [23]:

1. Let C_1 be the first node in the loop-cutset, which has been topologically sorted. Set C_1 to v_1, the first value in the current loop-cutset instantiation.
2. Let x be the prior probability that $C_1 = v_1$.
3. Initialize the joint probability to 1.
4. While there are still loop-cutset members that have not yet been instantiated, do:

Chapter 6

Incorporating Additional Information

The facilities for calculating genetic risk that I have described thus far rely only on simple phenotypic evidence in the pedigree (i.e., affected vs. unaffected) and on the background risk of the disorder. GENINFER is capable of incorporating other sources of information, concerning both the disorder and individual family members. In addition to the population frequency of the disease, the penetrance and the mutation rate may be supplied as input. Some disorders may have age-dependent expressivity; this can be specified so that it is taken into account. Finally, there may be auxiliary phenotypic information, such as enzyme levels, for members of the family; these data are automatically combined with other forms of information to produce combined genotype probabilities.

6.1 Penetrance

In some genetic disorders, there may be individuals who have affected genotypes, yet appear normal. These people can pass on the defective allele to their children. The probability that a person with a defective gene will exhibit the defect is called the *penetrance* of the gene. Incomplete penetrance is different from simple recessivity: in a disorder with incomplete penetrance, there may be two individuals who have

manner similar to penetrance probabilities, since the probability that the disorder is expressed at a given age is equivalent to its penetrance at that age.

6.3 Mutation

A child with a genetic disorder has usually inherited the disorder from his or her parents. Sometimes, however, a genetic defect may be due to a spontaneous mutation that took place in the genes of the affected individual. In the case of certain genetic disorders, e.g., achondroplastic dwarfism, mutation is to blame more often than inheritance. In other disorders, spontaneous mutation may be rare but not unheard of.

The possibility of spontaneous mutation should be taken into consideration both to predict risks for future offspring and to explain the genotypes of ancestors. For example, when a child affected with a dominant disorder is born to two unaffected parents, mutation may be to blame. GENINFER allows the mutation rate of a disorder to be specified in the input. Unlike penetrance, which affects prior probabilities, the mutation rate is taken into account by altering the conditional probability matrices. Table 6.1 shows a conditional probability matrix for an autosomal recessive disorder with mutation rate μ. The exact numbers in the matrix are less important than the fact that some entries that used to be zero have become non-zero. Note that the probabilities in each column still sum to one.

			Mother and father						
Child	**AA**	**AH**	**AN**	**HA**	**HH**	**HN**	**NA**	**NH**	**NN**
A	1	$.5 + \mu$	μ	$.5 + \mu$	$.25 + \mu$	μ	μ	μ	0
H	0	$.5 - \mu$	$1 - \mu$	$.5 - \mu$	$.5 + 2\mu$	$.5 + 2\mu$	$1 - \mu$	$.5 + 2\mu$	μ
N	0	0	0	0	$.25 - 3\mu$	$.5 - 3\mu$	0	$.5 - 3\mu$	$1 - \mu$

Table 6.1: Conditional probability matrix for an autosomal recessive disorder with mutation rate μ

What about the possibility of back mutation? Could a defective allele spontaneously revert to a normal state? While not impossible, this phenomenon is rare

A given piece of information may not definitively reveal the true phenotype; there may be uncertainty associated with any piece of information. The results of a test are therefore weighted by the accuracy of the test.

Supplementary information is included in the network by allowing each person node to have more than one dummy leaf. Each dummy leaf represents some evidence regarding the person's genotype. The information is entered in the form $P(finding|genotype_i)$. If a test is not 100% accurate, the probabilities will not be 0 or 1. This is how uncertainty regarding the significance of the data is encoded. For example, if an individual has tested positive for an abnormally high level of some enzyme, and the probability that a positive result on this test indicates a heterozygous genotype is 0.92, then the vector will contain the element *P(high enzyme level | heterozygous)* = .92. The user does not need to perform a Bayesian revision on the data, because this is done automatically by Pearl's algorithm.

When there was only one dummy leaf per node, the probability that a node had a particular genotype was calculated by multiplying together the final π and λ vectors on the link between the node and its dummy leaf. If supplementary phenotypic data is entered, causing some nodes to have more than one dummy leaf, all leaves must be considered when calculating the belief for a node. The new belief function is:

$$BEL(person_i) = \alpha * \prod_d \lambda_d(person_i) \sum_{g \in GxG} P(person_i|PU_g)\pi_i(PU_g)$$

where PU is the parental unit of person i, $d \in$ dummy leaves of A, and $G = \{$*affected, heterozygous, normal*$\}$.

6.5 Explaining Anomalies

Sometimes the information provided to GENINFER by a user contains apparent inconsistencies. For example, the child of two unaffected parents may be identified as exhibiting a dominant disorder (one with 100% penetrance, let's assume). Situations of this type cause all of the beliefs calculated by GENINFER to come out to zero for one or more individuals. The program checks for this occurrence. When it is detected, the location in the pedigree of the unexpected event is pinpointed, and

Genotype probabilities for BETTY-FAMILY family:

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
hypoth-male	0.16667	0.00000	0.83333
hypoth-female	0.00000	0.16667	0.83333
ARTHUR	0.00000	0.00000	1.00000
ANNE	0.00000	1.00000	0.00000
BENJAMIN	1.00000	0.00000	0.00000
BILL	1.00000	0.00000	0.00000
BETTY	0.00000	0.33333	0.66667
BOB	0.00000	0.00000	1.00000
CLAUDE	0.00000	0.00000	1.00000

Consultands BETTY and BOB are concerned about the risk of passing on HEMOPHILIA, an X-LINKED disorder, to future offspring.
After analyzing all available information, I have assessed the risks as follows:
Female offspring have a 0.0% chance of being affected with HEMOPHILIA and a 17% chance of being carriers.
Male offspring have a 17% chance of being affected and a 83% chance of being normal.

Table 6.2: Output of GENINFER on Betty's family (see Figure 3.1)

Chapter 7

Conclusions

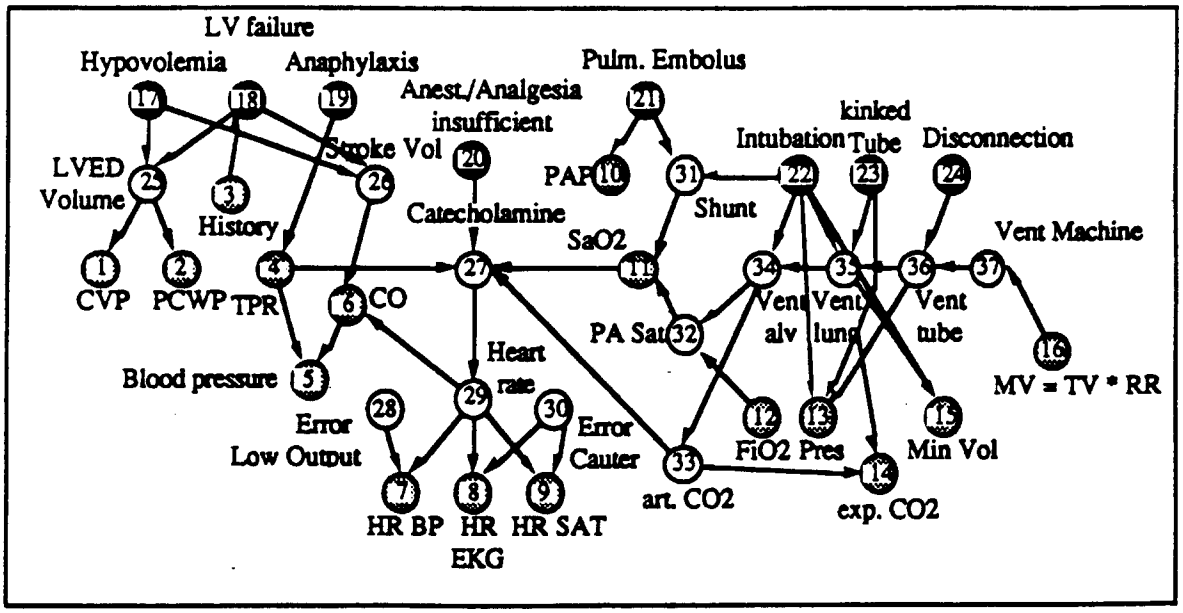
I have shown that Pearl's method for propagation and fusion in probabilistic belief networks can be implemented in a working system in a real-world domain, and that clustering and conditioning can be used together to handle successfully the problem of multiply-connected networks.

Several characteristics of the genetic counseling domain make it well suited to an artificial intelligence approach. The domain encompasses many types of knowledge, both qualitative and quantitative, and a successful approach must be able to combine these diverse sources of information. Cases on which to test a program for genetic counseling are readily available. The problem of uncertainty must be dealt with appropriately. In genetic counseling, unlike some other domains, the uncertainty can generally be expressed numerically, which makes probabilistic reasoning more directly applicable.

In order to adapt Pearl's algorithm for use in the genetic counseling domain, some aspects had to be changed substantially, while others remained relatively untouched. Figuring out how to set the initial parameters was a large part of the battle to implement the algorithm. Moreover, because all evidence is available to the network at the same time, certain boundary conditions had to be adjusted. The procedures that break cycles by clustering and conditioning the networks added substantially to the size of the program.

The choice of Pearl's method for the problem of genetic counseling has a number

Suermondt, Chavez, and Cooper [2] have performed such an experiment in a different domain: they implemented Lauritzen and Spiegelhalter's method and Pearl's method (with conditioning) and compared their performance on a sample network which implements an alarm message system for patient monitoring (*ALARM*). The ALARM network is shown in Figure 7.1 [2].



The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (◉) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

Figure 7.1: The ALARM network

The time complexity of Pearl's conditioning algorithm is proportional to the product of the size of the network, the number of loop-cutset instantiations, and the number of pieces of evidence, whereas the time complexity of Lauritzen and Spiegelhalter's approach is linear in the number of cliques and exponential in the size of the largest clique in the network. Because of the configuration of the ALARM network, Lauritzen and Spiegelhalter's algorithm ran significantly faster on this network than did Pearl's. The ALARM network has five separate loops, which makes the loop-cutset impractically large. With a loop-cutset of this size, Pearl's propagation algorithm must be run 160 ($5 * 2^5$) times.

it could ask the user to enter information relevant to a particular disorder. This capability is already present to a limited degree. For example, GENINFER asks if the disorder being examined is age-dependent; if the user indicates that it is, he or she is prompted to enter the ages of family members.

GENINFER's utility would be increased if it were supplied with more background knowledge about specific genetic disorders. It could keep a database of facts such as the population frequencies and penetrances of various genetic disorders. It could also be stocked with data about disorders with age-dependent expressivity; currently, Huntington's disease is the only disorder for which it has this kind of data.

It has been suggested that I enable the genetic counseling program to run in reverse: given a family affected with a genetic disorder, have the program figure out the inheritance pattern of the disorder. This capability would be useful for cases involving heritable defects that can result from several different inheritance patterns (e.g., retinitis pigmentosa). This problem might be amenable to an approach involving belief networks.

Another possibility for future work is to implement Lauritzen and Spiegelhalter's algorithm in the genetic counseling domain, and empirically compare the running times.

In its current form, GENINFER can provide a genetic counselor with genetic probabilities, but it is not equipped to offer advice on desirable courses of action. Adding a module that employed utility theory and decision analysis would narrow the gap between GENINFER's capabilities and the capabilities of a human genetic counselor. It is not clear, however, that such an addition would be feasible, or that it would be appreciated. Assigning utilities to such variables as the value of having a normal child is a difficult task, and not one that most consultands would feel comfortable with. Moreover, physicians have traditionally displayed a lack of enthusiasm for computer programs that they feel might replace them.

It is clear that no computer program can or should take the place of a human physician. With this caveat in mind, we can continue to explore the ways in which

Bibliography

[1] John M. Agosta. The structure of Bayes nets for vision recognition. In *The Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 1–7. University of Minnesota, 1988.

[2] Ingo A. Beinlich, H.J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. Technical Report KSL-88-84, Knowledge Systems Laboratory, Stanford University, Stanford, California, January 1989. Submitted to AI In Medicine conference, London, 1989.

[3] C. Cannings and E.A. Thompson. *Geneological and Genetic Structure.* Cambridge University Press, 1981.

[4] R. Martin Chavez and Gregory F. Cooper. A fully polynomial randomized approximation scheme for the Bayesian inferencing problem. Memo KSL 88-72, Knowledge Systems Laboratory, Stanford University, Stanford, California, October 1988. Accepted for publication in *Networks.*

[5] Peter Cheeseman. In defense of probability. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 1002–1007, Los Angeles, California, August 1985.

[6] Gregory F. Cooper. Expert systems based on belief networks–current research directions. Memo KSL 87-51, Knowledge Systems Laboratory, Stanford University, Stanford, California, August 1987.

[7] Gregory F. Cooper. The computational complexity of probabilistic inference using belief networks. Memo KSL-87-27, Knowledge Systems Laboratory, Stanford University, May 1988.

[8] Persi Diaconis and Sandy Zabell. Some alternatives to Bayes's rule. In B. Grofman and G. Owen, editors, *Information and Group Decision Making*, Proceedings of the Second Conference on Political Economy, pages 25–38. University of California, Irvine, 1986.

[9] Ivar Heuch and Francis H.F. Li. PEDIG–a computer program for calculation of genotype probabilities using phenotype information. *Clinical Genetics*, 3:501–504, 1972.

[24] H.J. Suermondt and Gregory F. Cooper. Updating probabilities in multiply-connected belief networks. In *The Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 335–343. University of Minnesota, 1988.

[25] R.E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.*, 13:566–579, 1984.

[26] A. Thomas. Approximate computation of probability functions for pedigree analysis. *IMA Journal of Mathematics Applied in Medicine and Biology*, 3:157–166, 1986.

[27] Elizabeth A. Thompson. *Pedigree Analysis in Human Genetics.* The Johns Hopkins University Press, 1986.

[28] M. Yannakakis. Computing the minimum fill-in is NP-complete. *SIAM J. Algebraic Discrete Methods*, 2:77–79, 1981.

```
Mother (if known):
Father (if known):
Additional phenotypic information:

(|person56| IS #<Name: ANNE; Gender: FEMALE; Parents: UNKNOWN
 and UNKNOWN; Pheno: UNAFFECTED)

More people to enter?  (y or n, default y):   y

Person's name (must be unique):   Arthur
Gender:   male
Phenotype (affected, unaffected, or unknown):   unaffected
Mother (if known):
Father (if known):
Additional phenotypic information:

(|person57| IS #<Name: ARTHUR; Gender: MALE; Parents: UNKNOWN
 and UNKNOWN; Pheno: UNAFFECTED)

More people to enter?  (y or n, default y):  y

Person's name (must be unique):   Benjamin
Gender:   male
Phenotype (affected, unaffected, or unknown):   affected
Mother (if known):   Anne
Father (if known):   Arthur
Additional phenotypic information:

(|person58| IS #<Name: BENJAMIN; Gender: MALE; Parents: ANNE
 and ARTHUR; Pheno: AFFECTED)


;  Input other people...


More people to enter?  (y or n, default y):   n

If there is a specific consultand, please enter her name, and then
 her husband's name (if known).
Consultand:   Betty
Husband:   Bob

(|family55| = #<Family: BROWN.  Disorder: HEMOPHILIA (X-LINKED).
 Consultands: BETTY, BOB
 Background risk: 0.01; Penetrance: 1; Mutation rate: 0.>

BETTY-FAMILY family before propagating information:
#<Family: BETTY-FAMILY.  Disorder: SPASTIC-PARAPLEGIA (X-LINKED).
 Consultands: BETTY, BOB.
 Background risk: 0.01; Penetrance: 1; Mutation rate: 0.>
```

SMITH family before propagating information:
#<Family: SMITH. Disorder: RETINITIS-PIGMENTOSA (X-LINKED).
Consultands: DAPHNE, NONE.
Background risk: 1.0e-4; Penetrance: 1; Mutation rate: 0.>

PERSON	GENDER	PHENOTYPE	PARENTS
ALICE	FEMALE	UNAFFECTED	UNKNOWN, UNKNOWN
ANDREW	MALE	AFFECTED	UNKNOWN, UNKNOWN
BARBARA	FEMALE	UNAFFECTED	ALICE, ANDREW
BRIAN	MALE	UNAFFECTED	AMALIA, UNKNOWN
CLARA	FEMALE	UNAFFECTED	BARBARA, UNKNOWN
CLIFFORD	MALE	AFFECTED	BARBARA, UNKNOWN
DAVID	MALE	UNAFFECTED	CLARA, UNKNOWN
DIANA	FEMALE	UNAFFECTED	CLARA, UNKNOWN
DOROTHY	FEMALE	UNAFFECTED	CLARA, UNKNOWN
DAPHNE	FEMALE	UNAFFECTED	CLARA, UNKNOWN
DIANASON1	MALE	UNAFFECTED	DIANA, UNKNOWN
DIANASON2	MALE	UNAFFECTED	DIANA, UNKNOWN
DOROTHYSON1	MALE	UNAFFECTED	DOROTHY, UNKNOWN
DOROTHYSON2	MALE	UNAFFECTED	DOROTHY, UNKNOWN
DOROTHYSON3	MALE	UNAFFECTED	DOROTHY, UNKNOWN
DAPHNESON1	MALE	UNAFFECTED	DAPHNE, UNKNOWN

Genotype probabilities for SMITH family:

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
hypoth-male47	1.94157e-2	0.00000	0.98058
hypoth-female46	1.94157e-6	1.95118e-2	0.98049
ALICE	0.00000	1.00000e-4	0.99990
ANDREW	1.00000	0.00000	0.00000
BARBARA	0.00000	1.00000	0.00000
BRIAN	0.00000	0.00000	1.00000
CLARA	0.00000	0.11649	0.88351
CLIFFORD	1.00000	0.00000	0.00000
DAVID	0.00000	0.00000	1.00000
DIANA	0.00000	2.32994e-2	0.97670
DOROTHY	0.00000	1.29448e-2	0.98706
DAPHNE	0.00000	3.88314e-2	0.96117
DIANASON1	0.00000	0.00000	1.00000
DIANASON2	0.00000	0.00000	1.00000
DOROTHYSON1	0.00000	0.00000	1.00000
DOROTHYSON2	0.00000	0.00000	1.00000
DOROTHYSON3	0.00000	0.00000	1.00000
DAPHNESON1	0.00000	0.00000	1.00000

Consultand DAPHNE is concerned about the risk of passing on
RETINITIS-PIGMENTOSA, an X-LINKED disorder, to future offspring.
After analyzing all available information, I have assessed the risks as

A.3 Age-dependent expressivity

The consultands in the following two examples, Betty and Bob, are concerned about the risk of Huntington's disease, since Betty's father and brother are affected with Huntington's. In the first example, Betty and Bob are fairly old, so the probability that they are carrying the Huntington's gene but have not yet expressed it is low. The couple in the second example is young, so there is a higher probability that they might pass on the Huntington's allele to their offspring, without yet having manisfested the disease themselves.

A.3.1 Old parents

```
BROWN family before propagating information:
#<Family: BROWN.  Disorder: HUNTINGTON (AUTOSOMAL-DOMINANT).
 Consultands: BETTY, BOB.
 Background risk: 5.0e-5; Penetrance: 1; Mutation rate: 0.>
```

PERSON	GENDER	AGE	PHENOTYPE	PARENTS
ARTHUR	MALE	65	AFFECTED	UNKNOWN, UNKNOWN
ANNE	FEMALE	64	UNAFFECTED	UNKNOWN, UNKNOWN
BENJAMIN	MALE	45	AFFECTED	ANNE, ARTHUR
BETTY	FEMALE	42	UNAFFECTED	ANNE, ARTHUR
BOB	MALE	40	UNAFFECTED	UNKNOWN, UNKNOWN

```
Genotype probabilities for BROWN family:
```

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
ARTHUR	1.52547e-5	0.99998	0.00000
ANNE	0.00000	2.59332e-6	1.00000
BENJAMIN	8.64445e-7	1.00000	0.00000
BETTY	0.00000	0.15256	0.84744
BOB	0.00000	1.80006e-5	0.99998

```
Consultands BETTY and BOB are concerned about the risk of passing on
HUNTINGTON, an AUTOSOMAL-DOMINANT disorder, to future offspring.
After analyzing all available information, I have assessed the risks
as follows:
Future offspring have a 8% risk of being affected with HUNTINGTON.
```

A.3.2 Young parents

```
BROWN family before propagating information:
```

```
   PERSON       HOMOZYGOUS AFFECTED   HETEROZYGOUS   HOMOZYGOUS NORMAL
-------------------------------------------------------------------------
  JUDY                 0.00000           0.27836          0.72164
  CHARLIE              0.00000           1.74409e-2       0.98256
  NOMI                 0.00000           0.14550          0.85450
  ELAINE               0.00000           0.14550          0.85450

Consultands JUDY and CHARLIE are concerned about the risk of passing on
TAY-SACHS, an AUTOSOMAL-RECESSIVE disorder, to future offspring.
After analyzing all available information, I have assessed the risks as
follows:
Future offspring have a 0.077% risk of being affected with TAY-SACHS
and a 14% chance of being carriers.
```

A.5 Anomalous situation

In this example, a child with a dominant disorder is born to two unaffected parents. This anomalous situation results in all-zero belief functions for some of the family members. This outcome is detected by GENINFER, which proposes possible explanations for the anomaly.

```
HARRIS family before propagating information:
#<Family: HARRIS.  Disorder: ACHONDROPLASTIC-DWARFISM (AUTOSOMAL-DOMINANT).
 Consultands: JUDY, CHARLIE.
 Background risk: 0.002; Penetrance: 1; Mutation rate: 0.>
   PERSON       GENDER    PHENOTYPE       PARENTS
-------------------------------------------------
  JUDY          FEMALE    UNAFFECTED    UNKNOWN, UNKNOWN
  CHARLIE       MALE      UNAFFECTED    UNKNOWN, UNKNOWN
  NOMI          FEMALE    UNAFFECTED    JUDY, CHARLIE
  ELAINE        FEMALE    AFFECTED      JUDY, CHARLIE

Genotype probabilities for HARRIS family:

   PERSON       HOMOZYGOUS AFFECTED   HETEROZYGOUS   HOMOZYGOUS NORMAL
-------------------------------------------------------------------------
  JUDY                 0.00000           0.00000          0.00000
  CHARLIE              0.00000           0.00000          0.00000
  NOMI                 0.00000           0.00000          0.00000
  ELAINE               0.00000           0.00000          0.00000

There is an apparently anomalous situation in the HARRIS family.
ELAINE, who would be expected to be unaffected, is listed as affected.
There are several possible explanations for this.
1. The penetrance of ACHONDROPLASTIC-DWARFISM is not really 100%
   (so JUDY or CHARLIE might actually have the affected genotype,
    despite appearing unaffected).
```

A.7.1 One loop

Output of GENINFER on pedigree from [21]. This pedigree has one cycle, caused by the marriage between Charles and his niece Florence. Charles and Florence are concerned about having a child with APKD, since Florence's brother George has an affected son. Charles is selected as the node for the loop-cutset, and the propagation algorithm is run once for every possible genotype that George could have.

```
SPIEGELHALTER-FAMILY before propagating information:
#<Family: SPIEGELHALTER-FAMILY.  Disorder: APKD (AUTOSOMAL-RECESSIVE).
 Consultands: FLORENCE, CHARLES.
 Background risk: 0.002; Penetrance: 1; Mutation rate: 0.>
   PERSON          GENDER    PHENOTYPE       PARENTS
-------------------------------------------------------
  hypothetical275 FEMALE    UNKNOWN       FLORENCE, CHARLES
  ANNETTE         FEMALE    UNAFFECTED    UNKNOWN, UNKNOWN
  BARTLEBY        MALE      UNAFFECTED    UNKNOWN, UNKNOWN
  CHARLES         MALE      UNAFFECTED    ANNETTE, BARTLEBY
  DONNA           FEMALE    UNAFFECTED    ANNETTE, BARTLEBY
  FLORENCE        FEMALE    UNAFFECTED    DONNA, UNKNOWN
  GEORGE          MALE      UNAFFECTED    DONNA, UNKNOWN
  HILDA           FEMALE    UNAFFECTED    UNKNOWN, UNKNOWN
  JOHN            MALE      AFFECTED      HILDA, GEORGE

(PEDIGREE HAS CYCLE--FORMING CUTSET)
(CUTSET IS
 (#<Name: CHARLES; Gender: MALE; Parents: ANNETTE and BARTLEBY;
 Pheno: UNAFFECTED; PU: #<Parental unit: parents are ANNETTE, BARTLEBY>>))

(CONDITIONING NETWORK...)

(CONFIG (CHARLES = UNAFFECTED) RESULTED IN JOINT CUTSET PROB 0.6)
(SAVED BELIEF OF (0.0 0.25187972 0.74812037) FOR |hypothetical273|)
(SAVED BELIEF OF (0.0 0.25687972 0.74312025) FOR ANNETTE)
(SAVED BELIEF OF (0.0 0.25687972 0.74312025) FOR BARTLEBY)
(SAVED BELIEF OF (0.0 0.0 1.0) FOR CHARLES)
(SAVED BELIEF OF (0.0 0.49874687 0.5012531) FOR DONNA)
(SAVED BELIEF OF (0.0 0.50375944 0.49624062) FOR FLORENCE)
(SAVED BELIEF OF (0.0 1.0 0.0) FOR GEORGE)
(SAVED BELIEF OF (0.0 1.0 0.0) FOR HILDA)
(SAVED BELIEF OF (1.0 0.0 0.0) FOR JOHN)

(CONFIG (CHARLES = HETEROZYGOUS) RESULTED IN JOINT CUTSET PROB 0.4)
(SAVED BELIEF OF (0.12593986 0.5 0.37406015) FOR |hypothetical273|)
(SAVED BELIEF OF (0.0 0.2568797 0.74312025) FOR ANNETTE)
(SAVED BELIEF OF (0.0 0.2568797 0.74312025) FOR BARTLEBY)
(SAVED BELIEF OF (0.0 1.0 0.0) FOR CHARLES)
(SAVED BELIEF OF (0.0 0.4987469 0.5012532) FOR DONNA)
(SAVED BELIEF OF (0.0 0.50375944 0.49624062) FOR FLORENCE)
(SAVED BELIEF OF (0.0 1.0 0.0) FOR GEORGE)
```

PERSON	GENDER	PHENOTYPE	PARENTS
JEANETTE	FEMALE	UNAFFECTED	UNKNOWN, UNKNOWN
KATE	FEMALE	UNAFFECTED	JEANETTE, UNKNOWN
KYLE	MALE	UNAFFECTED	JEANETTE, UNKNOWN
LAURA	FEMALE	UNAFFECTED	KATE, KYLE
LANCE	MALE	UNAFFECTED	KATE, KYLE
MARK	MALE	AFFECTED	LAURA, LANCE

```
(PEDIGREE HAS CYCLE--FORMING CUTSET)
(CUTSET IS
 (#<Name: KYLE; Gender: MALE; Parents: JEANETTE and UNKNOWN;
  Pheno: UNAFFECTED; PU: #<Parental unit: parents are JEANETTE, NIL>>
  #<Name: LANCE; Gender: MALE; Parents: KATE and KYLE;
  Pheno: UNAFFECTED; PU: #<Parental unit: parents are KATE, KYLE>>))
```

Genotype probabilities for DOUBLE-CYCLE family:

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
JEANETTE	0.00000	0.35719	0.64281
KATE	0.00000	0.71430	0.28570
KYLE	0.00000	0.50008	0.49992
LAURA	0.00000	1.00000	0.00000
LANCE	0.00000	1.00000	0.00000
MARK	1.00000	0.00000	0.00000

This empty page was substituted for a blank page in the original document.

```
(SAVED BELIEF OF (0.0 1.0 0.0) FOR HILDA)
(SAVED BELIEF OF (1.0 0.0 0.0) FOR JOHN)

(CONFIG (CHARLES = AFFECTED) RESULTED IN JOINT CUTSET PROB 0.0)
(SAVED BELIEF OF (0.25187972 0.7481203 0.0) FOR |hypothetical273|)
(SAVED BELIEF OF (0.0 0.25687972 0.7431203) FOR ANNETTE)
(SAVED BELIEF OF (0.0 0.25687972 0.7431203) FOR BARTLEBY)
(SAVED BELIEF OF (1.0 0.0 0.0) FOR CHARLES)
(SAVED BELIEF OF (0.0 0.49874687 0.5012531) FOR DONNA)
(SAVED BELIEF OF (0.0 0.5037594 0.4962406) FOR FLORENCE)
(SAVED BELIEF OF (0.0 1.0 0.0) FOR GEORGE)
(SAVED BELIEF OF (0.0 1.0 0.0) FOR HILDA)
(SAVED BELIEF OF (1.0 0.0 0.0) FOR JOHN)

Genotype probabilities for SPIEGELHALTER-FAMILY:
```

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
hypothetical75	5.00750e-2	0.35023	0.59970
ANNETTE	0.00000	0.25138	0.74862
BARTLEBY	0.00000	0.25138	0.74862
CHARLES	0.00000	0.40000	0.60000
DONNA	0.00000	0.49975	0.50025
FLORENCE	0.00000	0.50075	0.49925
GEORGE	0.00000	1.00000	0.00000
HILDA	0.00000	1.00000	0.00000
JOHN	1.00000	0.00000	0.00000

```
Consultands FLORENCE and CHARLES are concerned about the risk of passing
on APKD, an AUTOSOMAL-RECESSIVE disorder, to future offspring.
After analyzing all available information, I have assessed the risks as
follows:
Future offspring have a 25% risk of being affected with APKD and a 75%
chance of being carriers.
```

A.7.2 Multiple loops

The pedigree for this rather unusual family has two loops caused by two generations of brother-sister inbreeding. The loop-cutset therefore contains two nodes, one from each loop.

```
DOUBLE-CYCLE family before propagating information:
#<Family: DOUBLE-CYCLE.  Disorder: THALESSEMIA-A (AUTOSOMAL-RECESSIVE).
 Consultands: NONE, NONE.
 Background risk: 1.0e-4; Penetrance: 1; Mutation rate: 0.>
```

```
2. ACHONDROPLASTIC-DWARFISM has variable expressivity.
3. JUDY and CHARLIE are not really ELAINE's parents.
4. A spontaneous mutation caused ELAINE to be affected with
   ACHONDROPLASTIC-DWARFISM.
5. There was user error in entering the pedigree data.
```

A.6 Disorder caused by new mutation

This pedigree is the same as the one in the previous example, but this time the mutation rate of the disorder is non-zero, which allows the possibility that a new mutation was responsible for the affected child.

```
HARRIS family before propagating information:
#<Family: HARRIS.  Disorder: ACHONDROPLASTIC-DWARFISM (AUTOSOMAL-DOMINANT).
 Consultands: JUDY, CHARLIE.
 Background risk: 0.002; Penetrance: 1; Mutation rate: 0.005.>
   PERSON       GENDER    PHENOTYPE       PARENTS
-------------------------------------------------
  JUDY          FEMALE    UNAFFECTED    UNKNOWN, UNKNOWN
  CHARLIE       MALE      UNAFFECTED    UNKNOWN, UNKNOWN
  NOMI          FEMALE    UNAFFECTED    JUDY, CHARLIE
  ELAINE        FEMALE    AFFECTED      JUDY, CHARLIE

Genotype probabilities for HARRIS family:

    PERSON       HOMOZYGOUS AFFECTED   HETEROZYGOUS   HOMOZYGOUS NORMAL
-------------------------------------------------------------------------
  JUDY                 0.00000           0.00000           1.00000
  CHARLIE              0.00000           0.00000           1.00000
  NOMI                 0.00000           0.00000           1.00000
  ELAINE               0.00000           1.00000           0.00000

Consultands JUDY and CHARLIE are concerned about the risk of passing on
ACHONDROPLASTIC-DWARFISM, an AUTOSOMAL-DOMINANT disorder, to future
offspring.
After analyzing all available information, I have assessed the risks as
follows:
Future offspring have a 0.5% risk of being affected with
ACHONDROPLASTIC-DWARFISM.
```

A.7 Consanguinity

Consanguinity, or inbreeding, in the family pedigree causes the belief network for the family to have one or more loops. These loops are broken by conditioning the network.

```
#<Family: BROWN.  Disorder: HUNTINGTON (AUTOSOMAL-DOMINANT).
 Consultands: BETTY, BOB.
 Background risk: 5.0e-5; Penetrance: 1; Mutation rate: 0.>
   PERSON        GENDER    AGE    PHENOTYPE       PARENTS
-------------------------------------------------------------
  ARTHUR         MALE      45     AFFECTED      UNKNOWN, UNKNOWN
  ANNE           FEMALE    44     UNAFFECTED    UNKNOWN, UNKNOWN
  BENJAMIN       MALE      23     AFFECTED      ANNE, ARTHUR
  BETTY          FEMALE    22     UNAFFECTED    ANNE, ARTHUR
  BOB            MALE      25     UNAFFECTED    UNKNOWN, UNKNOWN

Genotype probabilities for BROWN family:

    PERSON       HOMOZYGOUS AFFECTED   HETEROZYGOUS   HOMOZYGOUS NORMAL
-----------------------------------------------------------------------
  ARTHUR              4.56519e-5         0.99995           0.00000
  ANNE                0.00000            1.96632e-5        0.99998
  BENJAMIN            6.55445e-6         0.99999           0.00000
  BETTY               0.00000            0.45655           0.54345
  BOB                 0.00000            7.49981e-5        0.99992

Consultands BETTY and BOB are concerned about the risk of passing on
HUNTINGTON, an AUTOSOMAL-DOMINANT disorder, to future offspring.
After analyzing all available information, I have assessed the risks
as follows:
Future offspring have a 23% risk of being affected with HUNTINGTON.
```

A.4 Additional phenotypic information

The consultands are concerned about bearing a child with an autosomal recessive disorder. A carrier test performed on Judy yields a positive result, which implies with 95% certainty that she is, in fact, a carrier.

```
HARRIS family before propagating information:
#<Family: HARRIS.  Disorder: TAY-SACHS (AUTOSOMAL-RECESSIVE).
 Consultands: JUDY, CHARLIE.
 Background risk: 0.01; Penetrance: 1; Mutation rate: 0.>
 PERSON      GENDER    PHENOTYPE      PARENTS          ADDITIONAL INFO
-----------------------------------------------------------------------
 JUDY        FEMALE    UNAFFECTED   UNKNOWN, UNKNOWN     (0 .95 .05)
 CHARLIE     MALE      UNAFFECTED   UNKNOWN, UNKNOWN
 NOMI        FEMALE    UNAFFECTED   JUDY, CHARLIE
 ELAINE      FEMALE    UNAFFECTED   JUDY, CHARLIE

Genotype probabilities for HARRIS family:
```

```
follows:
Female offspring have a 0.000019% chance of being affected with
RETINITIS-PIGMENTOSA and a 1.951% chance of being carriers.
Male offspring have a 1.942% chance of being affected and a 98% chance
of being normal.
```

A.2.2 High background risk

Output of GENINFER on pedigree shown in Figure 3.2, with population risk set to 0.01 (1000 times as high as in the previous example).

```
SMITH family before propagating information:
#<Family: SMITH.  Disorder: RETINITIS-PIGMENTOSA (X-LINKED).
 Consultands: DAPHNE, NONE.
 Background risk: 0.01; Penetrance: 1; Mutation rate: 0.>
```

Genotype probabilities for SMITH family:

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
hypoth-male67	1.96575e-2	0.00000	0.98034
hypoth-female66	1.96575e-4	2.92643e-2	0.97054
ALICE	0.00000	1.000000e-2	0.99000
ANDREW	1.00000	0.00000	0.00000
BARBARA	0.00000	1.00000	0.00000
BRIAN	0.00000	0.00000	1.00000
CLARA	0.00000	0.11751	0.88249
CLIFFORD	1.00000	0.00000	0.00000
DAVID	0.00000	0.00000	1.00000
DIANA	0.00000	2.36482e-2	0.97635
DOROTHY	0.00000	1.32037e-2	0.98680
DAPHNE	0.00000	3.93149e-2	0.96069
DIANASON1	0.00000	0.00000	1.00000
DIANASON2	0.00000	0.00000	1.00000
DOROTHYSON1	0.00000	0.00000	1.00000
DOROTHYSON2	0.00000	0.00000	1.00000
DOROTHYSON3	0.00000	0.00000	1.00000
DAPHNESON1	0.00000	0.00000	1.00000

```
Consultand DAPHNE is concerned about the risk of passing on
RETINITIS-PIGMENTOSA, an X-LINKED disorder, to future offspring.
After analyzing all available information, I have assessed the risks as
follows:
Female offspring have a 0.002% chance of being affected with
RETINITIS-PIGMENTOSA and a 3% chance of being carriers.
Male offspring have a 1.966% chance of being affected and a 98% chance
of being normal.
```

PERSON	GENDER	PHENOTYPE	PARENTS
ARTHUR	MALE	UNAFFECTED	UNKNOWN, UNKNOWN
ANNE	FEMALE	UNAFFECTED	UNKNOWN, UNKNOWN
BENJAMIN	MALE	AFFECTED	ANNE, ARTHUR
BILL	MALE	AFFECTED	ANNE, ARTHUR
BETTY	FEMALE	UNAFFECTED	ANNE, ARTHUR
BOB	MALE	UNAFFECTED	UNKNOWN, UNKNOWN
CLAUDE	MALE	UNAFFECTED	BETTY, BOB

Genotype probabilities for BETTY-FAMILY family:

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
hypoth-male	0.16667	0.00000	0.83333
hypoth-female	0.00000	0.16667	0.83333
ARTHUR	0.00000	0.00000	1.00000
ANNE	0.00000	1.00000	0.00000
BENJAMIN	1.00000	0.00000	0.00000
BILL	1.00000	0.00000	0.00000
BETTY	0.00000	0.33333	0.66667
BOB	0.00000	0.00000	1.00000
CLAUDE	0.00000	0.00000	1.00000

```
Consultands BETTY and BOB are concerned about the risk of passing
on HEMOPHILIA, an X-LINKED disorder, to future offspring.
After analyzing all available information, I have assessed the
risks as follows:
Female offspring have a 0.0% chance of being affected with
HEMOPHILIA and a 17% chance of being carriers.
Male offspring have a 17% chance of being affected and a 83%
chance of being normal.
```

A.2 Big pedigree with different prior risks

The output of GENINFER on the big pedigree in Figure 3.2 is shown here; input is not shown. The prior or background risk of the disorder in question, retinitis pigmentosa, is set to two different values so that the results may be compared. As was mentioned, the genotype probabilities are only slightly changed, even when the background risk is changed 1000-fold.

A.2.1 Low background risk

Output of GENINFER on pedigree shown in Figure 3.2, with population risk set to 0.0001.

Chapter A

Appendix

This appendix contains several examples of GENINFER running. The first example shows both input and output; the other examples show only output.

A.1 Betty's family

GENINFER running on Betty's family (figure 3.1). Both input and output are shown; text typed in by the user is shown in italics.

```
(geninfer)

Welcome to GenInfer.  This program evaluates genotype probabilities
in a family with some genetic disorder.  You will be asked to enter
information about the family and then about the individuals in the
family.

Relationships between family members are specified by listing each
person's parents, so you should type in individuals from the top of
the pedigree down.

First I will ask you for some information about the family being
counseled.

Family name:  Brown
Name of disorder:  hemophilia
Inheritance-type (autosomal-recessive, autosomal-dominant, or
 X-linked):  X-linked
Population frequency of disease allele (default 0.01):
Penetrance of disorder (between 0 and 1, default 1):  0.99
Does this disorder exhibit age-dependent expressivity?  (default no):

Please enter individuals in family, starting with the oldest generation.
Person's name (must be unique):  Anne
Gender:  female
Phenotype (affected, unaffected, or unknown):  unaffected
```

[10] J. Hilden. Computerized derivations of Mendelian probability formulae: the GENEX processor. In A. Hoskuldsson et al., editor, *Nordic Symposium in Applied Statistics and Data Processing*, pages 395–410. NEUCC–Technical University of Denmark, Lyngby, 1982.

[11] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, B50, 1988.

[12] R.S. Ledley and L.B. Lusted. Reasoning foundations of medical diagnosis. *Science*, 130:9–21, 1959.

[13] Edmond A. Murphy. How much difference does the use of Bayesian probability make? In Herbert A. Lubs and Felix de la Cruz, editors, *Genetic Counseling.* Raven Press, 1977.

[14] Edmond A. Murphy and Gary A. Chase. *Principles of Genetic Counseling.* Year Book Medical Publishers, Chicago, 1975.

[15] Kristian G. Olesen and et al. A MUNIN network for the median nerve–a case study on loops. *Applied Artificial Intelligence*, 3(2-3), 1989.

[16] S. G. Pauker and S. P. Pauker. Prescriptive models to support decision making in genetics. In G. Evers-Kiebooms, J. J. Cassiman, H. Vandenbeghe, and G. d'Ydewalle, editors, *Genetic Risk, Risk Perception, and Decision Making*, pages 279–296. Alan R. Liss, New York, 1987.

[17] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[18] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[19] H.U. Prokosch, S.A. Seuchter, E.A. Thompson, and M.H. Skolnick. Applying expert system techniques to human genetics. *Submitted to Computers and Biomedical Research*, July 1988.

[20] Ross D. Shachter. Evaluating influence diagrams. *Operations Research*, 34:871–882, 1986.

[21] David J. Spiegelhalter. Fast algorithms for probabilistic reasoning in influence diagrams, with applications in genetics and expert systems. In *Conference on Influence Diagrams.* University of California, Berkeley, May 1988.

[22] David J. Spiegelhalter and Steffen L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. Technical Report R 89-10, Aalborg Universitetscenter, 1989.

[23] H.J. Suermondt and G.F. Cooper. Initialization for the method of conditioning. Memo KSL 89-29, Knowledge Systems Laboratory, Stanford University, Stanford, California, April 1989.

[illegible] This capability is already present to a limited deg[illegible] if the disorder being examined is age-dependent; if the user indicates that it is, he or she is prompted to enter the ages of family members.

GENINFER's utility would be increased if it were supplied with more background knowledge about specific genetic disorders. It could keep a database of facts such as the population frequencies and penetrances of various genetic disorders. It could also be stocked with data about disorders with age-dependent expressivity; currently, Huntington's disease is the only disorder for which it has this kind of data.

It has been suggested that I enable the genetic counseling program to run in reverse: given a family affected with a genetic disorder, have the program figure out the inheritance pattern of the disorder. This capability would be useful for cases involving heritable defects that can result from several different inheritance patterns (e.g., retinitis pigmentosa). This problem might be amenable to an approach involving belief networks.

Another possibility for future work is to implement Lauritzen and Spiegelhalter's algorithm in the genetic counseling domain, and empirically compare the running times.

In its current form, GENINFER can provide a genetic counselor with genetic probabilities, but it is not equipped to offer advice on desirable courses of action. Adding a module that employed utility theory and decision analysis would narrow the gap between GENINFER's capabilities and the capabilities of a human genetic counselor. It is not clear, however, that such an addition would be feasible, or that it would be appreciated. Assigning utilities to such variables as the value of having a normal child is a difficult task, and not one that most consultands would feel comfortable with. Moreover, physicians have traditionally displayed a lack of enthusiasm for computer programs that they feel might replace them.

It is clear that no computer program can or should take the place of a human physician. With this caveat in mind, we can continue to explore the ways in which

Other factors which handicap Pearl's performance on this example are the large sets of data that must be propagated sequentially, and the peripheral locations of most of the measurement nodes, which increase the number of possible loop-cutset instantiations. The Lauritzen-Spiegelhalter procedure, in contrast, runs faster when there is a large set of evidence, because evidence simplifies the clique trees [2]. In the ALARM network, no node has more than three parents, so the maximum clique size stays relatively small.

Although Lauritzen & Spiegelhatler's algorithm outperformed Pearl's algorithm on the ALARM network, there is reason to believe that the difference in performance might be minimal in the genetic counseling domain. One factor is that the Lauritzen-Spiegelhalter algorithm requires overhead time to moralize and triangulate a network. This makes it more suitable for applications, such as ALARM, in which a single large network is going to be used repeatedly. The time required to configure networks might be more of a drawback for GENINFER, since a new belief network is constructed for each pedigree.

Another point to consider is that preliminary results have suggested that the Lauritzen-Spiegelhalter algorithm is efficient for networks with many small cycles, but less good for networks with one or two large cycles, because of the work involved in triangulating such networks. If a pedigree has any cycles (disregarding the artifactual cycles caused by multiple-child families, which are eliminated by clustering), they are likely to be fairly large: matings between siblings are less common than consanguinity involving more distantly related individuals.

7.2 Possible Extensions

In order to make GENINFER more accessible to genetic counselors, the current user interface probably should be replaced with a graphical interface. One possibility would be to let the user draw pedigrees with the mouse, and have the program ask questions about family members in order to acquire the needed information.

It might be possible to endow the system with greater "intelligence" so that

of advantages. Probabilistic reasoning is a suitable approach for a field in which risks and likelihoods play such a central role and uncertainty is readily expressed numerically. Pedigrees fit naturally into the belief network approach, and this approach has the advantage of being able to fuse all available data pertaining to the pedigree. Supplementary data of various types can be incorporated orthogonally, without disturbing the underlying structure of the family network. It is easy to assess risks for prospective offspring by adding hypothetical children to the belief networks. Clustering works well with the family networks, since each child has only two parents. Since prior probabilities of genotypes are available, the system can make intelligent "guesses" about the genotypes of people for whom no phenotypic information is available.

The main disadvantage of Pearl's method is its slowness, particularly for families whose family networks contain cycles. It might be possible to improve the running time of GENINFER by using more heuristics and taking advantage of special cases. For example, if a child with a recessive disorder is born to a couple with normal phenotypes, it is clear to a human expert that either the parents are both carriers or the appearance of the disorder in the child was caused by a new mutation. The program takes a while to reach this conclusion, because it must propagate all information through the entire pedigree. It might be worthwhile to have the program scan the pedigree before beginning the propagation algorithm in order to check for genotype assessments that follow immediately from the structure of the pedigree. Another possibility would be to use Lauritzen and Spiegelhalter's method, which shares most of the advantages of Pearl's method and may run faster.

7.1 Pearl vs. Lauritzen and Spiegelhalter

Implementing Spiegelhalter's method for calculating genotype probabilities would be an interesting experiment: the running time of that program could then be compared with the running time of the program based on Pearl's method. Beinlich,

family (Figure 3.1) is shown in table 6.6. Several sample runs of GENINFER, showing both input and output, are included in the Appendix.

Genotype probabilities for BETTY-FAMILY family:

PERSON	HOMOZYGOUS AFFECTED	HETEROZYGOUS	HOMOZYGOUS NORMAL
hypoth-male	0.16567	0.00000	0.83433
hypoth-female	0.00000	0.16567	0.83433
ARTHUR	0.00000	0.00000	1.00000
ANNE	0.00000	1.00000	0.00000
BENJAMIN	1.00000	0.00000	0.00000
BILL	1.00000	0.00000	0.00000
BETTY	0.00000	0.33333	0.66667
BOB	0.00000	0.00000	1.00000
CLAUDE	0.00000	0.00000	1.00000

Consultands BETTY and BOB are concerned about the risk of passing on HEMOPHILIA, an X-LINKED disorder, to future offspring. After analyzing all available information, I have assessed the risks as follows:

Female offspring have a 0.0% chance of being affected with HEMOPHILIA and a 17% chance of being carriers.

Male offspring have a 17% chance of being affected and a 83% chance of being normal.

Table 6.2: Output of GENINFER on Betty's family (see Figure 3.1)

possible explanations for the apparent anomaly are proposed. In the situation just described, the following explanations would be proposed:

- The penetrance of the disease is not really 100%.
- The disorder has variable expressivity, and the parents of the affected child are actually affected with mild cases of the disorder.
- The putative parents of the affected child are not the actual biological parents.
- The mutation rate of the disorder is non-zero; a spontaneous mutation occurred in the affected child.
- The user made one or more errors when entering the data.

6.6 Input and Output of GENINFER

The current version of GENINFER has an interface that prompts users to enter information about the genetic disorder being investigated, and then lets them enter data for individuals in the pedigree. The user is asked to enter the family name, disorder, inheritance type, background risk, penetrance, etc. For some fields, such as penetrance, a default value is supplied, which the user can accept or modify. Family members are entered in topological order, starting with the oldest ancestors; family relationships can then be completely specified by simply identifying each person's mother and father. For each family member, the user is asked to enter the individual's gender, phenotype (which may be "unknown"), parents, and any supplementary phenotypic evidence that is available, such as the results of enzyme tests. The user can also specify a particular consultand and, optionally, the consultand's spouse or partner.

The output of GENINFER is a list of the probabilities that each member of the family is homozygous affected, heterozygous, or homozygous normal. If the pedigree appears to contain anomalous or contradictory information, possible explanations are proposed, as was discussed in the previous section. If a consultand has been specified, GENINFER calculates the consultand's risk of bearing an affected child. (For X-linked disorders, separate risks are calculated for male and female offspring.) For example, the table of genotype probabilities that GENINFER outputs for Betty's

enough that it can be disregarded. There are many ways for a normal gene to become defective, but very few ways for a defective gene to become normal. In general, the possibility of back mutation can be safely ignored [14].

Considering spontaneous mutation as a possible cause of genetic disease is more useful for explaining unusual pedigree configurations than for predicting genetic risk to future offspring. I found that non-zero mutation rates cause problems with one aspect of the conditioning algorithm. If a particular loop-cutset configuration results in an "impossible" assignment of genotypes to individuals, we want to be sure the joint probability for this instantiation comes out to zero. If the mutation rate of the disorder is non-zero, however, these incorrect configurations will not be caught, and the final beliefs will be incorrect. In order to avoid this problem, the program sets the mutation rate to zero if it is necessary to condition the network. Setting the mutation rate to zero has little effect on the calculated beliefs, and it prevents gross errors from occurring. However, this problem probably should be addressed in future versions of GENINFER.

6.4 Combining Multiple Sources of Information

Phenotypic information can take more than one form. In the simplest cases, it is clear from straightforward observation that an individual either has the genetic disease or does not have it. Sometimes, however, there may be other sources of information, such as enzyme levels, that suggest the presence of a defective allele. One of the advantages of using Pearl's algorithm is that it allows all available data to be supplied to the network and combined appropriately.

Types of data that might be relevant to a genetic consultation include:

- Results of carrier tests
- Enzyme levels
- Blood groups
- Restriction fragment length polymorphisms

the same genotype, and yet different phenotypes. For example, an individual could have the allele for a dominant disorder and yet not exhibit the disorder. Neurofibromatosis is an example of a genetic disorder with relatively low penetrance.

The user can specify the penetrance of a disorder as a number between 0 and 1. If the penetrance probability is not specified by the user, the program assumes 100% penetrance. If a disease has 100% penetrance, then all individuals with the allele will exhibit a phenotype consistent with their genotype. Penetrance information is incorporated into the belief network by entering it into the initial π and λ parameters, changing the prior beliefs. The prior probability that a person with a normal phenotype has a defective genotype becomes 1 - *penetrance*. If *penetrance* is 1, this probability will be 0, just as it was before we considered penetrance probabilities. The prior probabilities are the only quantities affected by penetrance; the conditional probability matrices, for example, are unchanged.

6.2 Age-dependent Expressivity

Some genetic disorders do not reveal their presence until the affected individual reaches adulthood. The most familiar example of this kind of late-onset genetic defect is Huntington's disease, which is caused by an autosomal dominant allele. People with the Huntington's allele seem normal until some time in middle age, when the devastating symptoms begin to appear. By this time, they may already have had children, each of whom has a 50% chance of inheriting the disorder. Because the presentation of symptoms of Huntington's disease is time-dependent, determination of phenotype is not clear-cut. A person who has reached the age of 65 without symptoms probably does not carry the defective gene, but we can reach no such conclusion about an asymptomatic 25-year-old.

In order to handle disorders with age-dependent expressivity, GENINFER must be supplied with data about the percentage of people who express the disorder at each age range. For testing purposes, I have provided it with data for Huntington's disease. The age-dependent probabilities of presentation can be handled in a

5. Multiply the joint by x.

6. Propagate the influence of the "evidence" $C_1 = v_1$ through the network, using the normal propagation and fusion equations. After propagation, the other nodes will have parameters consistent with the "observed" value for C_1.

7. Restore node C_1 to its original state (i.e., its value is no longer fixed.)

8. Consider the next loop-cutset node, C_2, and the next value, v_2. Set x to $\mathrm{BEL}(C_2 = v_2)$. Because we already propagated the instantiation of C_1 to v_1, $\mathrm{BEL}(C_2 = v_2) = P(C_2 = v_2 | C_1 = v_1)$.

9. Loop back to step 4.

5.5.4 Speeding up conditioning

The exponential running time of the conditioning algorithm can be a problem for belief networks with multiple cycles. One way to minimize the time required by conditioning is to choose a minimal (or close to minimal) loop-cutset. Another possibility is to process the loop-cutset instantiations in parallel [24]. Each time the propagation algorithm is run during conditioning, the order of updates will be the same; the difference lies in the initializations of certain parameters. Therefore, it should be feasible to maintain, on each link, π and λ vectors for each loop-cutset instantiation, and update all of them at once.

If a particular loop-cutset instantiation results in an impossible assignment of genotypes to individuals (for example, an affected person being labeled as "homozygous normal"), the joint probability for the instantiation will come out to zero, and the beliefs calculated during the instantiation will be irrelevant. This case can be taken advantage of in order to heuristically speed up conditioning: if a loop-cutset instantiation is known to be impossible, it is not necessary to propagate the influence of that particular instantiation. I have not implemented this heuristic, but doing so would probably be straightforward.

time we follow a link to a node, we mark both the link and the node as being visited. When searching for the next node to visit, we follow only unvisited links out of the current node. If we ever follow an unvisited link and reach a node that has already been visited, there must be a cycle, because there is more than one path between two nodes. Moreover, the already-visited node that alerted us to the presence of a cycle must be in the cycle; we can use it as a starting point to search for a loop-cutset node.

5.5.3 Conditioning the network

Once the loop-cutset has been found, the network must be physically disconnected at the nodes in the loop-cutset in order to break the cycles. A copy of the loop-cutset node is included on both sides of a break, as shown in Figure 5.3.

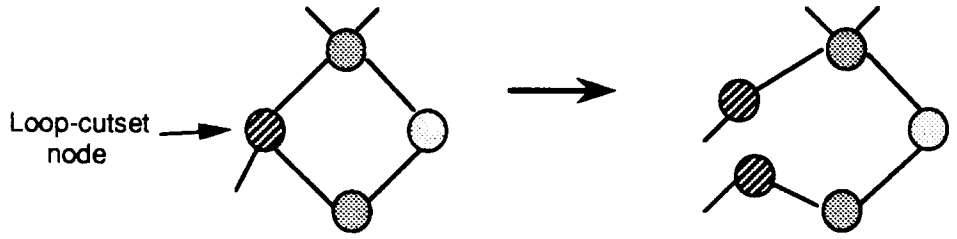


Figure 5.3: Disconnecting the network at a loop-cutset node

The next step is to instantiate all the nodes in the loop-cutset and run the propagation algorithm once for all such instantiations, of which there will be an exponential number: $|genotypes|^{|cutset|}$. (In practice, the size of the loop-cutset is usually small, so this exponential complexity is not a major problem in this domain.) In order to find the conditioned beliefs for each node, we sum the products of the values found for each loop-cutset instantiation and the weights of the loop-cutset instantiations:

$$BEL(A_i) = \sum_{k=1}^{|G|^{|C|}} BEL_k(A_i)P(C_k)$$

where C_k represents the kth instantiation of the loop-cutset, and

type k, given that the parents are of types i and j. The only information conveyed by these matrices is how the genotype probabilities of the parents are fused in the corresponding parental unit. The equations for propagating the influence of π and λ vectors must be altered somewhat in order to handle the heterogeneous vectors and matrices, but the basic mechanism of message-passing is unchanged.

The use of clustering eliminates looping due to artifactual cycles. However, as all possible combinations of propositions from the individual nodes in a cluster must be represented in the "supernode" that comprises the cluster, this method is not practical for large cycles such as those that result from matings between related individuals. These cycles can be broken by conditioning the network.

5.5 Conditioning

A multiply-connected belief network can be *conditioned* by selecting a *loop-cutset* from the network and considering all possible combinations of values that nodes in the loop-cutset can take on [24]. Each possible combination is treated as a separate case. Conditioning is sometimes referred to as reasoning by assumptions, because for each configuration of the loop-cutset, we are assuming that the nodes in the loop-cutset have those values, and reasoning about the rest of the network based on those assumptions. Pearl argues that the use of conditioning is not foreign to human reasoning: when we find it difficult to estimate the likelihood of a given outcome, we may make hypothetical assumptions to simply the process [17]. By considering each possible case separately, conditioning prevents infinite cycling without loss of information.

Because conditioning breaks the cycles in a multiply-connected network, evidence can be propagated in the conditioned network in the normal manner using Pearl's algorithm. The resulting beliefs are then weighed by the joint probability of the instantiated nodes in the loop-cutset. Given a piece of evidence E and a loop-cutset consisting of nodes $C_1, ..., C_n$, then for any node A,

$$P(A|E) = \sum_{C_1 ... C_n} P(A|E, C_1 = v_1, ..., C_n = v_n) P(C_1 = v_1, ..., C_n = v_n|E)$$

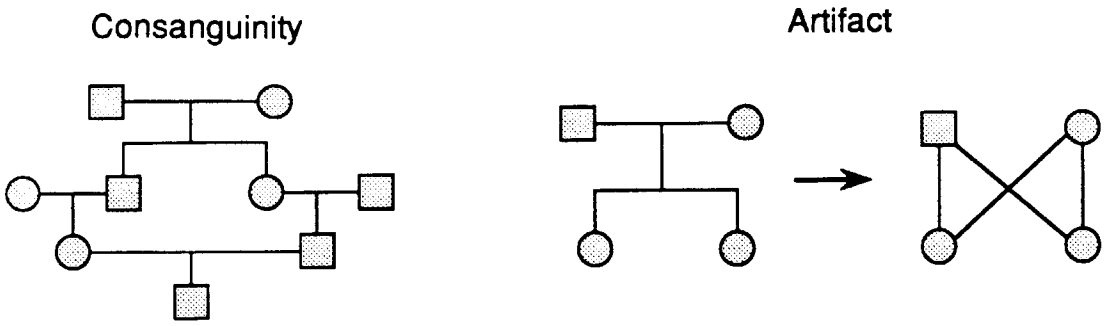


Figure 5.1: Two types of cycles in family networks

a general, exact solution to probabilistic reasoning in multiply-connected belief networks. Instead, fruitful approaches may involve special-cased algorithms or heuristic techniques that minimize the combinatorial complexity of the calculations.

5.3 Multiply-connected Family Networks

Belief networks for families may include two types of cycles. Some small proportion of families have cycles caused by consanguinity (for example, if two cousins marry). The number of nodes in these cycles will depend on the degree of consanguinity. Another type of cycle is more ubiquitous: it appears every time two parents have two or more children in common. These cycles are an artifact of a representation that connects each child with both of its parents. If there are two children, this will lead to an undirected figure-eight cycle (see Figure 5.1).

GENINFER uses a combination of clustering and conditioning to deal with the two types of cycles that can appear in family networks. Although conditioning can be used to break any cycle, its exponential time complexity makes it computationally undesirable. If conditioning were used to handle the artifactual cycles in all families with multiple children, the program would be unacceptably slow. Therefore, another technique must be used to break up these small cycles. I chose the clustering approach.

Chapter 5

Multiply-Connected Belief Networks

5.1 Why Cycles Are a Problem

Pearl's propagation method is restricted to singly connected graphs, i.e., graphs with at most one path between any two nodes. Because propagation of information is not under central control, information could cycle indefinitely if there were loops in the network. Even if the parameters converged to a stable equilibrium, the posterior probabilities that were calculated would not be correct, because the propagation equations are based on conditional independence assumptions that may be violated in multiply-connected networks [18]. For example, when we calculate a new π message, we assume that the parents of the child node have no common ancestors in the network.

5.2 Coping with Cycles

Pearl mentions three ways to handle graphs with cycles: clustering, conditioning, and stochastic simulation [17]. In *clustering*, groups of nodes are made into "supernodes" or clusters, so that the network formed by the clusters and the interconnections between them is acyclic. *Conditioning* prevents messages from cycling indefinitely

$$\pi_A(B_i) = Prior(B_i) \sum_{j,k \in \{A,H,N\}} \pi_B(C_j)\pi_B(D_k)P(B_i|C_j, D_k)$$

where A is the child node, B is the parent, and C and D are the parents of B.

The propagation phase is complete when the π and λ messages have reached stable values and are no longer changed by updates. If the network has no cycles, propagation will take time proportional to the longest path in the network. However, networks with cycles may never reach equilibrium. In the genetic counseling domain, cycles can be caused by consanguinity or by families with multiple children. Chapter 5 explains how this problem is handled.

4.2.4 Calculating genotype probabilities

When propagation is completed, the final parameters can be used to calculate the belief that each person is affected, heterozygous, or normal. The belief that a person has genotype k is the normalized product of λ_k and π_k on the link to that person's dummy leaf:

$$\text{BEL}(person_i) = \alpha\pi_{dummy}(person_i)\lambda_{dummy}(person_i)$$

The genotype beliefs thus obtained can be used to calculate the risk to future children of each person in the pedigree. GENINFER also allows a specific consultand or couple to be specified; the genotype probabilities for future children of the consultand(s) are then calculated. This is accomplished by having the program assign a "hypothetical" child to the consultand and calculate genotype probabilities for this child. Unlike the dummy leaves, this hypothetical child is treated like a regular person node; it has its own dummy leaf. If the disorder under consideration is X-linked, the risks to male and female offspring may be different, so two hypothetical children are created, one of each gender.

vector on the link to the dummy leaf would be (1, 1, 0), because P(*affected phenotype* | *homozygous affected*) = 1, P(*affected phenotype* | *heterozygous*) = 1, and P(*affected phenotype* | *homozygous normal*) = 0. (Note that this does not take into account the possibility of incomplete penetrance; this issue will be discussed in Section 6.1.) If the node being initialized is a root node for which we have phenotypic information, the π on the dummy link is initialized to match the λ on that link.

Often, the pedigree contains members whose phenotypes are not known. GENINFER permits their phenotypes to be specified as "unknown," and sets the λ vector on the dummy link to (1, 1, 1). If an individual of unknown phenotype is a root node, the π vector is set to reflect the background level of the disease in the population. I assumed that the genotype distribution of the population follows the Hardy-Weinberg equilibrium, i.e., if the frequency of the defective allele is p, and the frequency of the normal allele is q (where $q = 1 - p$), then p^2 of the population is homozygous affected, $2pq$ is heterozygous, and q^2 is homozygous normal. The value of p differs for different diseases and different populations. GENINFER allows the user to specify p for each disorder.

Thompson [27] points out that although the basis for allele frequencies and the applicability of the Hardy-Weinberg equilibrium may be difficult to justify, using these assumptions seldom presents a practical problem. As Thompson says, "The ability to assign a prior probability to a genotype is crucial, but the exact numerical value assigned seldom matters. Provided sensible assumptions are made, reliable

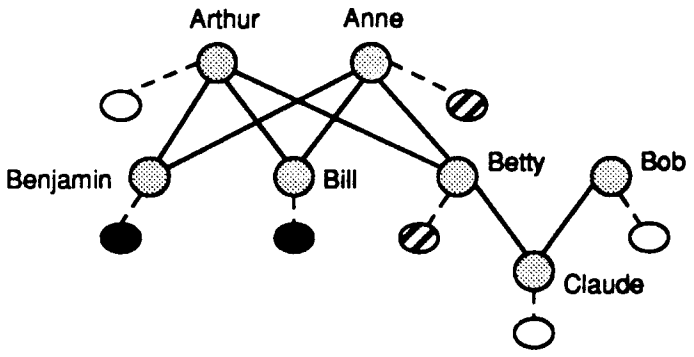


Figure 4.4: Dummy leaves represent phenotypes

4.2.1 Modeling the domain

In order to use Pearl's algorithm for pedigree analysis, the pedigree must be converted to a belief network in which the nodes represent people in the family and the links between nodes represent parent-child relationships. Other family relationships, such as "sibling," do not need to be specified explicitly; they are implicit in the network structure. Figure 4.3 shows the belief network that would be constructed for Betty's family.

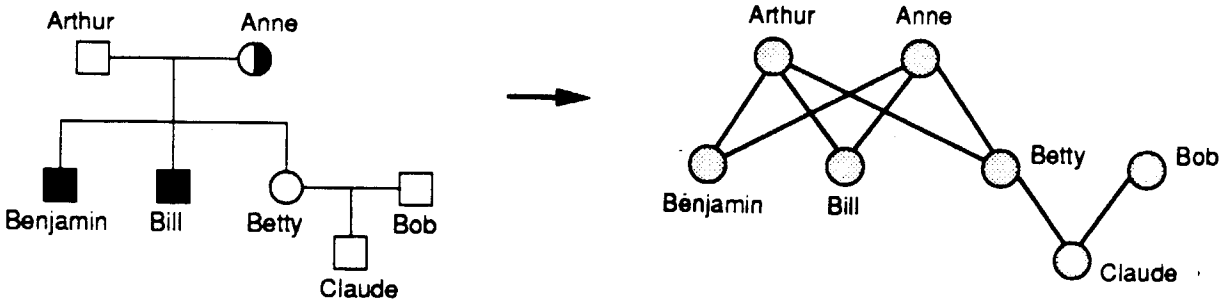


Figure 4.3: Belief network for Betty's family

In the genetic diseases that I have considered, there are three possible genotypes: homozygous affected, heterozygous (which may mean an affected or unaffected phenotype, depending on the inheritance pattern of the disorder in question), or homozygous normal. (For the case of males with X-linked disorders, I assume that they can be affected or normal, but not heterozygous. This is not strictly true, but it creates no inconsistencies.) Each of the three genotypes is considered as a "hypothesis" for the genotype of a node. Running Pearl's algorithm on a family network allows us to calculate the beliefs in each genotype for each family member.

The inheritance pattern of a disorder is encoded by the conditional probability matrix assigned to each node. The contents of these matrices depend on the inheritance pattern and, in the case of X-linked disorders, the gender of the individual. Since each person has two parents, and there are three possible genotypes for each person, the conditional probability matrices are three-dimensional matrices of size 3x3x3. Entry $M_{i,j,k}$ represents the probability that a woman with genotype i and a

belief in hypothesis i for variable A can be calculated as follows:

$$\mathrm{BEL}(A_i) = \alpha\lambda_X(A_i)\lambda_Y(A_i)\sum_{j,k} P(A_i|B_j, C_k)\pi_A(B_j)\pi_A(C_k)$$

where α is a normalization constant. [17]. $\lambda_X(A_i)$ refers to the diagnostic support from node X toward hypothesis i for variable A.

Alternatively, from the π and λ on the link to a single child node, we can calculate the belief distribution for the parent node A in a straightforward fashion:

$$\mathrm{BEL}(A_i) = \alpha\pi_X(A_i)\lambda_X(A_i).$$

4.1.2 Propagating information

Once initial values for π and λ have been assigned, the information represented by these vectors can be propagated throughout the network [17]. The new π on a link depends on the λs sent by the child node's sibling nodes and the πs sent by the parents of the parent node:

$$\pi_X(A_i) = \lambda_Y(A_i)[\sum_{j,k}\pi_A(B_j)\pi_A(C_k)\ P(A_i|B_j, C_k)],$$

where j and k range over all possible values for B and C. λ depends on the π of the spouse (i.e., the other parent of the child) and the λs of the children:

$$\lambda_A(B_i) = \sum_j[\pi_A(C_j)\sum_k \lambda_X(A_k)\lambda_Y(A_k)\ P(A_k|B_i, C_j)]$$

Each time a parameter is updated, all of the parameters that are causally related to it must be updated as well. In this way, information represented by the parameters is propagated in all directions through the network. When a π on a link is updated, the πs of the child nodes and the λ of the spouse node must be revised. When a link's λ parameters are updated, the λs of the parents and the πs of the child's siblings need to be recalculated. Figure 4.2 illustrates the propagation of updates of the parameters.

Whenever we update a parameter, we can put all of the parameters that depend on it on a queue to await updating. If the value of a parameter does not change when it is updated, the parameters that depend on it are not put on the queue. Propagation is complete when the queue of parameters to be updated is empty. The network will reach this stable state in time proportional to its diameter, assuming

Chapter 4

Pearl's Method

4.1 Propagation and Fusion in Singly-Connected Belief Networks

Pearl's method for fusion and propagation in probabilistic belief networks [17] allows all information relevant to a set of hypotheses to be combined in a manner consistent with Bayesian theory. It can be used in any domain for which uncertainty can be expressed numerically and conditional relationships between variables can be specified. This chapter describes the basic method and then explains how it was adapted to the genetic counseling domain.

Pearl's method calculates joint probabilities by making use of the chain-rule representation. The joint probability of all the nodes in the network can be expressed as a product of conditional probabilities, with each factor containing only one variable on the left side of the conditioning bar [17]. If the variables in the network are $x_1, ...x_n$, then

$$P(x_1, x_2, ..., x_n) = P(x_n|x_{n-1}, ..., x_1)...P(x_3|x_2, x_1)P(x_2|x_1)P(x1).$$

This means that the joint probability of any instantiation of all the variables in an n-node belief network can be calculated as a product of only n probabilities rather than all 2^n [6]. Quantifying the dependencies among nodes in this way also ensures the consistency and completeness of the belief network [17].

```
X is-grandfather-of Y;
```

It is not clear whether rules must be provided for every possible relationship: is there a rule, for example, that represents the relationship "is-second-cousin-once-removed-of"?

Genealogical analysis by Prokosch's system begins with an "intelligent" data acquisition phase in which heuristics help to guide requests for pedigree information. Next, forward-chaining inference evaluates all facts in the database and asserts all obtained relationships as facts. Subsequent steps depend on the question asked of the expert system. For example, the prototype system was used to surmise the ancestral source of a recessive allele. For this application, an expert system called GENEX (not the same GENEX written by Hilden) was called upon. The ultimate goal of Prokosch and his colleagues is to create an expert system shell for human genetics which will let experts in the domain create their own knowledge base and add modules for specific applications [19].

It is not clear what mechanism Prokosch's system uses to calculate genotype probabilities. Issues such as incorporating supplementary phenotypic data and handling families with consanguinity are not discussed. The most interesting contribution of this paper is the idea of heuristically deciding what pedigree information to request from the user and which ancestors to calculate probabilities for. This heuristic approach could potentially speed up genetic risk calculations.

3.2.4 Spiegelhalter

Of researchers who have considered the application of computer techniques to genetic counseling, Spiegelhalter was the only one to advocate the use of probabilistic belief networks. A recent paper by Spiegelhalter [21] explores, from a theoretical standpoint, the application of Lauritzen and Spiegelhalter's method [11] (see Section 2.4.3) to the problem of genetic inheritance.

Spiegelhalter has not yet implemented his proposal in a working system, nor does he address all of the aspects of the genetic counseling problem covered by GENINFER,

	Clara is a carrier	Clara is not a carrier
Prior	0.5	0.5
Conditional	$1/2 * 5/8 * 9/16 = 45/256$	1
Joint	45/512	256/512
Provisional posterior	$45/301 = 0.15$	$256/301 = 0.85$

Table 3.2: Calculation of Clara's risk of being a carrier

must be propagated both up and down in the tree in order to calculate the probabilities. To calculate the probability that Clara is a carrier, we use information provided by her son and her grandchildren. We then use Clara's risk when calculating probabilities for her daughters: her daughters' prior probability of being carriers is half their mother's risk.

3.1.1 Applying Bayesian methods to medicine

Although Bayesian methods provide more accurate estimates of risk than the "classical" Mendelian formulas, they are not in common use by most genetic counselors. This is due not only to historical ignorance of Bayesian techniques—a weakness that is only recently beginning to be corrected—but also to the fact that these methods are very complicated to use, especially for large families. Bayes' rule was invented in 1763, but it was not until the 1950s that Bayesian methods became widely used. In 1959, Ledley and Lusted's seminal paper [12] introduced formal probabilistic reasoning methods, including a simplified form of Bayes' rule, to the medical community. Murphy and Chase [14] were among the first to advocate the Bayesian approach to genetic counseling. They describe a collection of techniques that can be used to assess genetic risks in various types of families.

3.2 Previous Programs Dealing with Genetic Risk

Several previous programs have addressed the genetic counseling problem, but none so far have made use of Pearl's method. Spiegelhalter's approach [21] is the most promising, although it has not yet been implemented in a working system.

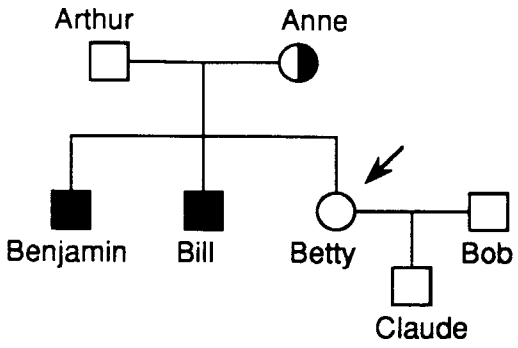


Figure 3.1: Pedigree for Betty's family

by the pedigree: the fact that Betty already has a normal son weakens the belief that she is a carrier.

The calculation described above neglects to take into account some of the information provided by the pedigree: we must consider Betty's descendents as well as her ancestors, and negative as well as positive information. Because Anne is an obligate carrier, Betty's prior probability of being a carrier for hemophilia is 0.5, as we calculated above. However, we have more information about Betty: the fact that she already has a normal son makes it less likely that she is a carrier. The Bayesian approach allows negative information, such as Betty's normal son, to be taken into account when calculating the probability that Betty is a carrier.

Table 3.1 shows the Bayesian derivation of Betty's risk of having a hemophilic son [14]. The row labeled "Conditional probability" lists the probability that Betty would have had a normal son if she were (or were not) a carrier. The joint is the product of the prior probability and the conditional probability. The posterior probability is obtained by normalizing the joint probabilities. The risk that Betty's next son will be hemophilic is half her risk of being a carrier.

Because simple Mendelian calculations fail to back-propagate information provided by unaffected offspring, they often overestimate risk. In Betty's case, for example, the Bayesian calculation leads to a risk estimate of 0.17 for Betty's next son, rather than 0.25. The larger the pedigree, the larger the disparity between simple Mendelian calculations and Bayesian revisions tends to be.

Consider the pedigree shown in Figure 3.2 for a family affected with an X-linked

2.5 Advantages and Disadvantages of a Probabilistic Approach to Uncertainty

One reason to favor probabilistic inference over extensional approaches is that it may be more compatible with the way people think. As Pearl argues, "The notions of dependence and conditional dependence are more basic to human reasoning than are the numerical values attached to probability judgements" [17]. Pearl feels that much of human knowledge can be represented by dependency graphs, and that we mentally trace links in these graphs in order to query or update that knowledge.

Cooper [6] points out that since probability is a widely used language for expressing uncertainty, expert systems that are probability-based have a better chance of being compatible with other systems. Another advantage is that in a probabilistic system, statistical data can be used directly as a form of knowledge.

The use of probabilistic inference in expert systems also has several drawbacks. One major obstacle to implementing systems based on probabilistic inference is that it is often difficult for a domain expert to state explicitly all of the variables and quantitative dependencies that are present within the domain. A partial solution to this problem has been proposed by Spiegelhalter and Lauritzen [22], who suggest sequentially updating probabilities as a database of cases accumulates. Another possible approach is the use of prototypical probability functions, which express a joint conditional probability by using many fewer than 2^{n-1} probabilities [6].

Even with techniques such as these, however, probabilistic inference is not appropriate for all domains. For example, if our subjective probabilities change as a result of introspection, without any change in the empirical data, Bayes' rule will not be capable of modeling the consequent changes in belief [8]. Moreover, some researchers feel that Bayesian methods are a poor model for human thought processes. Bayesian inference is clearly not applicable to all problems involving uncertainty, but it is nonetheless a useful paradigm.

Graph transformations

Starting with a directed graph, the first step is to transform it into a "moral" graph by dropping the directions from links and connecting nodes with common children. In order for node probabilities to be calculable by clique potentials, the graph must be *triangulated* (i.e., all cycles of length four or more must have a chord or "shortcut"). We would like to find a triangulation so that the cliques thus formed have the minimum number of total states, since computational efficiency depends on the number of possible states of a clique. [21]. The problem of finding the minimal fill-in that completely triangulates the graph is NP-complete [28], but a fairly good fill-in can be found in $O(N+E)$ time (where N is the number of nodes in the graph and E is the number of edges) by using an algorithm such as maximum cardinality search [25].

Once the graph has been triangulated, the maximal cliques can be regarded as clusters and treated as single nodes during propagation, since the hypergraph formed by the set of cliques of a triangulated graph is guaranteed to be acyclic [25]. The cliques also satisfy the *running intersection property*: there is an ordering of the cliques $C_1, C_2, ..., C_N$ such that $\forall i > 1, C_i \cap (C_k \cup ... \cup C_{i-1}) \subseteq C_k$ for some $k < 1$. In other words, the nodes of a clique also contained in previous cliques are all members of *one* previous clique, known as the parent clique [11]. We can therefore create a *junction-tree* in which each clique is joined to its unique parent. The junction-tree has the property that if any two cliques C_i and C_j have common nodes, then these nodes are contained in all cliques along the unique path between C_i and C_j [21]. The running intersection property allows the joint probability of a configuration of the network to be expressed as a product of functions on cliques [11].

Initialization and absorption

When the cliques are ordered in a junction-tree, then for each clique C_i there is a parent clique C_k, $k < i$. The *separator*, S_i, is defined as $C_i \cup C_k$ (i.e., the nodes in C_i "inherited" from the parent clique), and the residual R_i is $C_i \setminus S_i$ (the "new" nodes in C_i). The procedure for obtaining the joint probability of the cliques involves

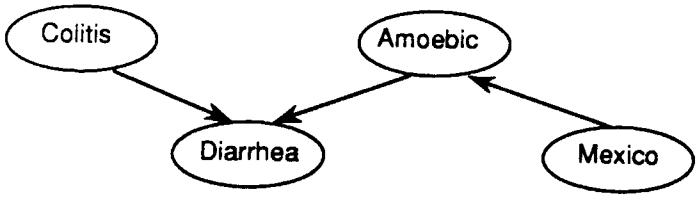


Figure 2.1: Belief network for amoebic infection/ulcerative colitis example

between variables.

2.4 Probabilistic Reasoning Techniques for Belief Networks

A number of researchers have derived methods for calculating beliefs or joint probabilities using belief networks. I will describe briefly the methods of Shachter, Pearl, and Lauritzen and Spiegelhalter. Pearl's method and Lauritzen and Spiegelhalter's method are capable of handling the same problems; their performance on various examples differs only in terms of running time. Shachter's method, unlike the others I describe, is useful for problems that involve decision-making under uncertainty.

2.4.1 Shachter

Shachter's method [20] operates on a type of network called an *influence diagram*. Influence diagrams may have three types of nodes: chance nodes, decision nodes, and value nodes. Directed arcs leading to random variable nodes (chance nodes or value nodes) indicate probabilistic dependence, while directed arcs to decision nodes indicate which information is available at the time of the decision. Shachter assumes that there is a single random variable associated with a unique value node, which represents the expected utility of the outcome. Influence diagrams may not have cycles: a cycle would violate the decision maker's free will or the assumption of time precedence. An influence diagram is *regular* if: (1) it has no cycles; (2) the value node (if present) has no successors; and (3) there is a directed path that contains all of the decision nodes (i.e., there is a total ordering of all the decisions).

headaches, and loss of appetite. The physician suspects that the patient may be suffering from a certain rare viral infection. Although this virus is found in only 2% of patients with these symptoms, it is important to check for this possibility, because if the virus goes untreated it could be fatal. However, the treatment for the virus has possibly detrimental side effects, so the physician does not want to administer it unnecessarily. The patient's blood is therefore tested for the presence of the virus.

The lab that tests for the virus has a good track record, but not a perfect one. If the patient is infected, there is a 99% chance that the virus will be detected. If, however, the patient is not infected, there is a 4% chance that the test will show a false positive result. If the patient described above tests positive for the presence of the virus, what should be the physician's belief that she is actually infected with the virus?

We can calculate $P(Infected|Positive)$, where *Positive* represents a "positive" result on the blood test, and *Infected* represents the event that the patient really has the viral infection, by using Bayes' formula:

$$\begin{aligned} P(Infected|Positive) &= \frac{P(Positive|Infected)P(Infected)}{P(Positive|Infected)P(Infected)+P(Positive|Uninfected)P(Uninfected)} \\ &= \frac{.99*.02}{.99*.02+.04*.98} \\ &= .34 \end{aligned}$$

The support accorded to the hypothesis that the patient is infected with the virus has been increased by the evidence of the positive test result, but the belief in this hypothesis is still only 1/3, even though the test is quite accurate by most standards.

Bayes' formula presupposes some simplifying conditions that may not always hold. Note that the likelihood ratio involves the term $P(E|\neg H)$, which is assumed to be a constant. However, since $\neg H$ can stand for any disease other than H, this conditional probability may vary, depending on which $\neg H$ we are considering. Belief networks remedy this limitation by allowing the likelihood ratio to change if new evidence arrives.

Chapter 2

Uncertainty

Most systems, whether they are natural or synthetic, can be represented as a set of interdependent elements. However, for many real-world domains we may not have a complete picture of all of the variables and the relationships between them. The variables may not be limited to a simple true/false dichotomy, and the implications between variables may be fuzzy. In order to model a real-world domain accurately, it is therefore desirable to have some way of representing and handling uncertainty. Although there are a number of alternatives, many researchers favor belief networks because they provide a natural and efficient way to handle uncertainty.

2.1 Approaches to Handling Uncertainty

Pearl [18] classifies approaches to handling uncertainty into three schools: logicist, neo-calculist, and neo-probabilist. *Logicists* attempt to handle uncertainty with nonnumerical techniques such as nonmonotonic logic. The *neo-calculist* school uses numerical representations of uncertainty, but rejects probability calculus and uses alternatives such as the Dempster-Shafer calculus, fuzzy logic, and certainty factors. The *neo-probabilists,* which include Pearl, hold to traditional probability theory, supplementing it with additional computational facilities to make it suitable for AI problems.

Approaches to uncertainty may also be categorized as *extensional* or *intensional.*

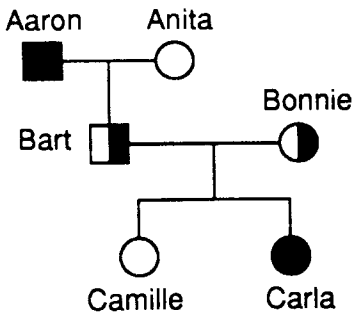


Figure 1.1: Pedigree for a family affected with albinism

Susan Pauker, a genetic counselor at the Harvard Community Health Plan. I also referred to textbooks on genetic counseling, particularly Murphy and Chase [14], for examples with which to compare the results obtained by my program.

1.3.1 Pedigrees

The most important source of information for a genetic counselor is the family history of the consultand. Pedigrees are family tree diagrams showing the incidence of a particular genetic disorder in a family. In pedigree diagrams, men are represented by squares, women by circles. (Individuals of unknown gender, such as unborn fetuses, may be indicated by diamonds.) The offspring of a couple are shown hanging from a line drawn between the two members of the couple. A filled-in circle or square represents a person who exhibits the trait in question; a half-filled circle or square indicates a definite carrier. In Chapter 4, I will show how pedigrees can be transformed into belief networks.

Figure 1.1 shows a pedigree for a family affected with albinism, an autosomal recessive disorder. Since Bart and Bonnie have an affected child, they must both be carriers for albinism. Camille may or may not be a carrier for the disorder.

1.1 Overview

The next two sections of this chapter present some of the basic principles of human genetics and genetic counseling. Chapter 2 gives an overview of approaches to uncertainty in artificial intelligence. Chapter 3 discusses previous approaches to the genetic counseling problem, both human- and computer-based. Chapter 4 first describes Pearl's basic algorithm and then explains how I adapted it for use in the domain of genetic counseling. In Chapter 5, methods for handling multiply-connected belief networks are discussed. Chapter 6 describes how supplementary data pertaining to the family and disorder of interest are incorporated by GENINFER. Finally, Chapter 7 reviews the insights gained by this project and discusses opportunities for future work.

1.2 Principles of Human Genetics

Humans have 22 pairs of autosomal chromosomes plus one pair of sex chromosomes, which are two X's for females and an X plus a Y for males. Thus, all genes occur in pairs (called *alleles*), with the exception of genes on the X chromosome, which are found in pairs only in females.

Genetic disorders can be classified into three basic categories: aneuploid, unilocal, and multilocal [14]. *Aneuploid* disorders, of which Down's syndrome is the most common example, are caused by an abnormal number of chromosomes. *Unilocal* conditions are attributable to a single base pair substitution at one point in a chromosome—in other words, they are caused by a single defective allele or pair of alleles. Many genetic disorders, such as cystic fibrosis and sickle cell anemia, are unilocal. *Multilocal* disorders are caused by defects at several different genetic loci (i.e., more than one gene is responsible). Although multilocal disorders, like unilocal disorders, are inherited, it is difficult to predict their occurrence or trace their progress through the generations. My system deals only with unilocal disorders.

Two concepts that are central to the study of genetics are *genotype* and *phenotype*. Genotype refers to the genetic composition of an individual with regard

Chapter 1

Introduction

In the early days of artificial intelligence, many researchers attempted to produce general-purpose programs capable of solving a range of problems. As we learn more about how difficult it is to solve seemingly simple AI problems, it becomes clear that trying to tackle such large issues may be less productive than selecting a specific domain, such as some aspect of medicine, in which to test ideas. Once a domain has been selected, designing programs in the domain may give rise to new ideas that have the potential to be extended and generalized. Thus, research in medical artificial intelligence has a dual purpose: to advance basic AI research, and to produce programs that are useful to medical professionals. The domain of genetic counseling provides opportunities for progress toward both goals. In particular, it is a good springboard for research in probabilistic belief networks.

As artificial intelligence techniques are applied to an ever-widening field of domains, the problem of how to handle uncertainty emerges repeatedly. The domain of genetic inheritance is no exception: many of the questions we might ask in this field involve uncertainty. For example, is a particular individual heterozygous or homozygous for an allele of interest? Are the children of a given couple likely to be affected with a particular genetic disorder? Was the gene that caused a baby to be born blind transmitted by her mother, her father, or both? In order to provide coherent answers to questions such as these, an expert, human or otherwise, must have some mechanism for handling uncertainties and probabilities. Belief networks

Acknowledgements

I would like to thank all those who made this thesis possible, and in particular, the following:

- The National Science Foundation for funding my graduate work;
- Stephen Pauker and Peter Szolovits for suggesting the genetic counseling project;
- Peter Szolovits for advising me and offering helpful comments on this thesis;
- Susan Pauker for domain expertise and enthusiasm;
- Michael Wellman for invaluable help with Pearl's algorithm and conditioning;
- Ramesh Patil for encouragement and helpful advice;
- Jaap Suermondt for providing me with a pre-publication draft of his paper on conditioning;
- Greg Cooper for sending me several useful papers on belief networks and conditioning;
- David Spiegelhalter for sending me copies of his papers and providing me with useful information about other related work;
- Judith Harris for suggesting the name GENINFER and for providing helpful comments on a draft of my thesis;
- Ed Yampratoon for working on a user interface for GENINFER.