This paper was published as:
P. Szolovits and S. P. Pauker. Pedigree Analysis for Genetic Counseling. In Lun, K. C., et al. (eds.), *MEDINFO 92: Proceedings of the Seventh Conference on Medical Informatics*, pages 679–683. Elsevier (North Holland) 1992.

Pedigree Analysis for Genetic Counseling

Szolovits P [a] and Pauker S P [b]

[a] *MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, Massachusetts 02139, USA*

[b] *Medical Genetics, Harvard Community Health Plan, 10 Brookline Pl. West, Brookline, Massachusetts 02146, USA*

We report on the design and implementation of a prototype program, GENINFER©, to assist genetic counselors in evaluating the risk of recurrence of genetic disorders based on the analysis of family pedigrees. The present version of the program integrates a convenient graphical interface that permits counselors to draw, examine and modify family pedigrees and to enter information relevant to risk analysis. It also includes a general-purpose Bayesian inference mechanism that permits the rapid calculation and display of probabilities of various genotypes, for the consultand and all other pedigree members. This is possible even in the presence of complex pedigrees with multiple consanguineous matings. The ability to support rapid calculation also enables the user to perform sensitivity analyses. Limitations of the prototype include a restriction to single-locus Mendelian disorders and an inability to make use of information from RFLP markers. Planned extensions include remedying these limitations, the incorporation of an algorithm for automated reformatting (layout) of an existing pedigree, improvements in the population genetic models used by the program, and connections to external databases for acquiring data on disease incidence, patterns of inheritance, mutation rates, penetrance, etc.

1. The Need for Computer Support of Genetic Counseling

Although the mathematical principles of pedigree analysis are well known and widely advocated (e.g., [1, 2, 3]), their application in the genetic counseling office is difficult because of lack of easy-to-use tools, and therefore they are only occasionally applied and then with severe simplifications [1]. As our understanding of the mechanisms of genetic disorders grows through knowledge gained in analyzing the human genome, both the opportunities for applying formal analytical techniques to genetic counseling and the difficulties of such applications will increase dramatically.

When counselors are posed the problem of determining the risk to an individual of having a heritable genetic disorder, they should perform a detailed analysis of the risks to that individual implied by all known facts about his family pedigree. In fact, however, that calculation is difficult and tedious. As a result, it is not uncommon for the counselor to estimate the risk solely on the basis of information known about the patient's ancestors, applying classical Mendelian inheritance probabilities. The problem with this approach is that it ignores possibly valuable information available from knowledge about other members of the family. To take an extreme example, consider a couple concerned that the mother may be a carrier for hemophilia, an X-linked recessive disease. For this example, we assume that one male in 3,000 live births is affected, and that the spontaneous mutation rate is small (10^{-5}); the analysis is actually quite insensitive to these assumptions. If she has one brother, who is affected, there is a 50% chance that she is a carrier. Classical Mendelian inheritance predicts that if she is a carrier, her son has a 50% chance of inheriting the disease. The risk to the son, therefore, is 25%. The naive application of this method would not take into account the couple's past experience with other children. However, in the extreme case that the couple have already produced nine older boys, none affected with hemophilia, a correct Bayesian probabilistic account of the situation assures us that the true probability of having an affected child is in fact slightly less than 0.1%. This is because the nine normal boys argue very persuasively that the mother is *not* a carrier. In a more common scenario, consider a family also at risk for hemophilia, but this time where the propositus is one generation further away from the known affected individual. (See Figure 1, below.) Here, the grandmother's brother has the disease, therefore the grandmother has a 50% chance of being a carrier, the mother 25%, and the proposed child, if male, a 12.5% chance of inheriting the disease according to the classical analysis. This is in fact correct, if no other relevant pedigree information is known. However, if the propositus has three unaffected maternal uncles, for example, the true risk is only 2.8%. If, in addition, he has two unaffected brothers, the risk drops to 0.7%.

In general, the "lazy man's risk assessment" that ignores much pedigree data will be considerably less accurate than the correct assessment. In addition, in more complex pedigrees, where disease is manifest along more than one line of inheritance or where consanguinity introduces reinforcing pathways for inheritance of defective genes, the classical methods are inapplicable. Correct probabilistic assessment, taking all known information into account, gives accurate assessments. Clearly, improved accuracy allows patients to make more informed and rational decisions. Several studies show that parents of a child with a hereditary disorder are more likely to plan a future child if they have had the advantage of proper risk assessment [4]. We believe that improving the accuracy of such risk assessment is likely to make couples more comfortable planning a conception. Unless analytical techniques such as those we propose here are used, failure to take into account the presence of unaffected relatives in the pedigree will falsely elevate the perceived risks.

We anticipate that the major research efforts now underway in the Human Genome project will lead to significant changes in the way that genetic counseling is practiced. It will become possible to test people for carrier status for many disorders, and it will be possible to detect *in utero* specific DNA abnormalities associated with a wide variety of diseases. It seems unlikely, however, that all potential parents will be routinely screened or that every fetus will routinely be tested for all possible genetic disorders. Therefore accurate risk assessment from pedigree data will help to choose those patients for whom DNA testing is appropriate as it becomes available. At present we know of over five thousand Mendelian heritable disorders [5], and we suspect the genetic etiology of many other disorders, though their inheritance patterns may not be well understood. As recurring defects in the genome become systematically known, our vocabulary of disorders will consequently expand, because almost any mutation, deletion or substitution in the genome might be associated with some specific disease. The genetic counselor of the future will have to be able to recognize and offer suitable advice to patients with that large variety of disorders—a task almost certain to require assistance from automated technologies. We believe that the tool whose prototype we describe here is an early step toward the kinds of counselor's workstations that must be developed to meet that need.

2. The Mathematics of Pedigree Analysis

Since Mendel, we have understood the heritability of various traits including those that cause or predispose to disease. Probabilistic analysis is applicable to quantifying the risks of disease to individuals, based on a knowledge of their family pedigree and of the genetics of the disease process being considered. It appears that probabilistic analysis was first appreciated in problems of medical diagnosis and therapy planning only in the 1950's or 60's, and the first comprehensive models of genetic analysis did not appear until the early 1970's. Then, as today, the driving force behind the investigations was the concern of the researcher rather than that of the counselor; thus, much of the effort and the framing of the results is in terms of linkage analysis to help determine what the appropriate models of heritability are for some disease of interest. In seeking data for such studies, however, researchers had to develop the appropriate mathematics of pedigree analysis for large and possibly complex families. Many in fact also recognized the applicability of these models to problems of counseling, though the practice of the counselor remained considerably more primitive than what could be supported by such models.

Within the genetics community, there are now many widely-available computer programs for linkage analysis; these programs also include the appropriate mathematical models for counseling. Their genesis as research instruments has, however, left them poorly-equipped to handle the needs for easy and flexible data input and display, and other important aspects of ease of use. At best, others have attempted to provide front-end programs using a graphical interface [6] to collect data to be passed to linkage programs, but such an approach is awkward and fails to provide immediate feedback. For this reason, we have chosen to build an integrated new program, GENINFER©, which combines a convenient graphical interface with an efficient new probabilistic inference algorithm. Our student, Nomi Harris, has produced an earlier version of GENINFER using similar mathematical methods, but lacking a convenient user interface [7].

The mathematical problem of computing the probability of a pedigree is a special case of the more recently-developed domain of Bayesian networks [8]. A Bayes net consists of a number of nodes representing probabilistic variables, interconnected by directed arcs representing probabilistic dependence. Each node depends jointly on those nodes from which arcs impinge on it. Nodes without incoming arcs have an *a priori* probability distribution. Bayes nets may not contain directed cycles, but allow multiple paths from one node to another, though typically at great computational expense. We take each variable (node) to have some small, discrete number of possible values. Though Bayes nets have also been defined using nodes with continuous values, using cumulative distribution functions, here we stay with the discrete case.

We treat a pedigree as a Bayes network wherein the genotype and every relevant aspect of the phenotype of each individual is represented as a probabilistic node [7]. For individuals whose gender is unknown (typically fetuses or possible future pregnancies), their gender is also taken to be such a node. The dependence of the genotype of an individual on the genotypes of his or her parents is represented as a conditional probability table or function, which incorporates the possibility of spontaneous mutations, if significant. For X-linked disorders, the genotype will also depend on gender. The dependence of phenotype on genotype is also expressed via conditional probabilities, taking into account the penetrance of the trait and its dependence on age, if significant. Founders of a pedigree (those without ancestors in the pedigree) are assumed to have population genotype probabilities, which are estimated from disease prevalence and the assumption of Hardy-Weinberg equilibrium. We also permit nodes to represent the outcomes of *observations*, which we take to depend only on other aspects of the phenotype, and *tests*, which we take to depend only on the genotype. Our prototype can handle such nodes in its mathematical treatment, but not in its user interface. Consanguinity is handled by the normal Bayes net machinery, though families with many generations of inbreeding slow the analysis considerably.

For any Bayes network, and for networks representing pedigrees in particular, the joint probability of any complete assignment of values to all the variables is given by

$$P(x_1, x_2, \ldots, x_n) = \prod_{1 \le i \le n} P\left(x_i \middle| \pi_{x_i}\right) \tag{1}$$

where π_{x_i} are the predecessor nodes on which x_i depends. In any case of clinical interest, of course, we will not know the actual value of every variable. When some subset $\{x_1, \ldots, x_k\}$ of variables have no known value, then the joint probability of the others may be obtained by summing many terms of the form (1), one term for each possible combination of values of the x_i:

$$(x_{k+1}, \ldots, x_n) = \sum_{x_1, \ldots, x_k} P(x_1, x_2, \ldots, x_n) \tag{2}$$

The number of terms in the sum is exponential in the number of variables with unknown values, and can be impractical to compute in such a direct manner. However, in actual pedigrees it is often the case that once some particular variable has been instantiated (e.g., once we assume some particular value for an unknown genotype), descendant portions of the pedigree become independent. This is the insight behind the "peeling" methods in common use in genetics, and it is generalized by various factoring methods in the Bayes net literature. Our program, for example, uses a heuristic method due to Cooper [9], which is equivalent to factoring the right-hand side of (2). A discussion of this technique in our application, and various extensions to speed up calculations, may be found in [10]. For simple Mendelian genetic models, such techniques are capable of giving virtually instant updating of probability estimates even for pedigrees of a few dozen individuals, on high-end personal computers.

3. The GENINFER Program

Figure 1 shows the graphical interface of the GENINFER program. As a Macintosh program, it shares the common conventions of all such programs, including facilities for creating, opening, closing, and saving pedigrees and their corresponding display windows; moving and changing the size of the window displaying a pedigree; and selecting and rearranging graphical elements representing the individuals in the pedigree. The tool palette at the top left allows the selection of input modes. The arrow is for general selection and movement of pedigree elements. The horizontal connector is for drawing spousal relations. The vertical connector is for connecting parents and their children. The other tools permit the user to draw common genotype/phenotype combinations for unaffected and affected males and for affected, carrier and genotypically-normal females. These symbols change depending on the nature of the disease model being analyzed. Other information about any individual may be entered by selecting the individual and opening a detailed dialog. There, one may specify other combinations of genotype and phenotype, the person's name, date of birth, age at onset of disease, whether the person has been examined by the consultant, etc. The same dialog also shows a current estimate of that person's probabilities for their possible genotypes. A separate dialog allows specification of the disease model under consideration, including whether it is autosomal or X-linked, dominant or recessive, prevalence of the disorder, penetrance, and spontaneous mutation rate.

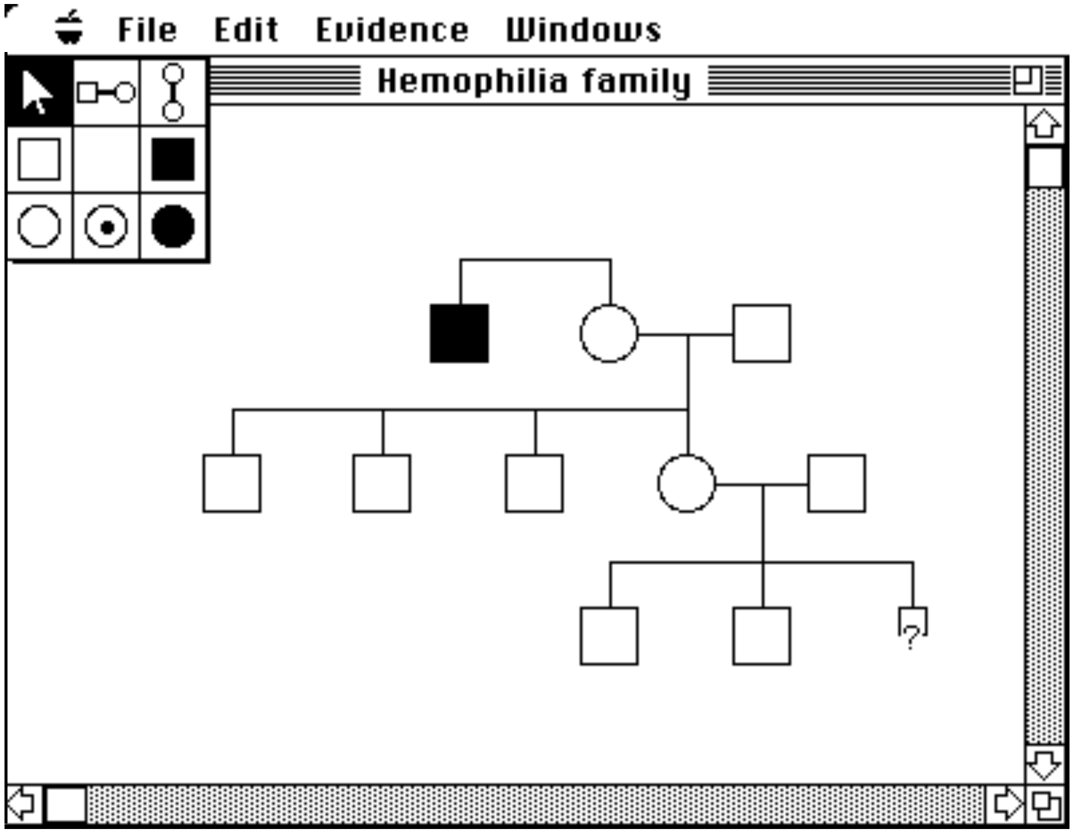

FIGURE 1. The graphical interface of GENINFER. The tools at the upper left permit placement of new individuals and the creation of spousal and parent-child links. The family shown represents the hemophilia A family discussed in the introduction, where a couple with two normal boys is concerned about the risk to a contemplated third male child, given that the maternal grandmother has an affected brother, but four unaffected children. Squares represent males, circles females. The small square at the bottom right represents the propositus, a contemplated male fetus, whose phenotype is marked with "?" because it is, of course, unknown. All others except the great-uncle are phenotypically unaffected. An analysis ignoring the five unaffected males would suggest a risk of 12.5% to the fetus. Given a population prevalence of 1/3000 males and a mutation rate of 10^{-5}, the correct risk is now about 0.7%. Even without brothers, the unaffected uncles would drop the risk to only 2.8%.

Selecting different tools during construction of the pedigree makes it very easy for the first-time user to understand how to proceed, but (like many modal interfaces) slows the more experienced user. Therefore, we also provide combinations of the Macintosh's Command, Option, Control and Shift keys that allow almost all operations to be performed without explicitly selecting modes from the palette. Like playing chords on a keyboard, this requires more practice but significantly speeds the input of complex pedigrees. For example, the pedigree in Figure 1 can be drawn in under two minutes, including typing all the required numeric information.

The program's analysis can be viewed by opening dialogs about each individual of interest. In addition, the program can display a small bar-graph next to each individual, showing the relative likelihood of that person's various possible genotypes. Though that display is not refined enough for making conclusions, it gives an excellent overview of the pedigree, showing these probabilities for each individual simultaneously.

4. Limitations of the Prototype Program and Possible Enhancements

Although we believe that the present program supports our approach to providing tools for genetic counselors, it contains many serious limitations that will need to be lifted in order to make it an effectively usable tool. The first is the current limitation to simple single-locus, two-allele Mendelian disorders. Extension to multi-allele systems (where there are three or more possible variant genes that may appear at a single locus) is simple, and requires only the installation of already-designed tools for specifying more complex conditional probability tables. Extension to multi-locus disorders is much more complex, but essential if we are to be able to handle not only true multi-local diseases but also the influence of genetic markers (e.g., RFLP's) on risk assessment. Both cases require consideration of the inheritance of linked loci, and thus of the recombination fractions (or LOD scores) between them.

A simple way to handle multi-local disorders is to treat the genotype as a multidimensional variable, one per locus. Recombinations can then be encoded in the conditional probabilities linking parents to offspring. This is not efficient, however. Even two loci with two alleles per locus permit 2^4 genotypes for one individual. Therefore, the table linking parents to offspring needs to contain $2^4 \times 2^4 \times 2^4 = 2^{12} = 4096$ entries. Many of these are zeros or redundant. A more sophisticated representation, such as phased haplotypes and elimination of impossible genotypes, can do better, as is done in many linkage analysis programs. However, non-zero mutation rates can compromise such methods. The relationship between genotype and phenotype is, of course, also considerably more complex in the case of multiple loci.

The graphical interface will also need to be able to show several traits and/or genes at several loci for each individual, and to be able to supply meaningful and customizable legends for this larger variety of symbols. We plan to take an approach similar to that used in Pedigree/Draw [11]. In addition, we have explored the automatic re-drawing of user-entered pedigrees to make them neater without significantly altering the layout chosen by the user. The current prototype can constrain symbols to be placed on a grid, which is of some help. Ed Yampratoom has implemented an interesting heuristic layout algorithm that appears to do a good job of making the minimal changes needed in a drawn pedigree to reduce crossed lines [12]. We plan to incorporate this in our next version. More mundane improvements will be adding the ability to print multi-page pedigrees and to produce reports in a summary tabular format.

In the longer term, we must also seek on-line (network or telephone) connections to data bases describing models of genetic disorders. As the models become more and more complex, it will be unreasonable for the user to have to enter myriad model parameters for each case. The coming availability of systematically summarized data from sources such as Online Mendelian Inheritance in Man should make such parameters easily available.

Although GENINFER is an incomplete, evolving program, it has already received favorable attention from clinical geneticists and counselors to whom we have shown it. The ease with which one can accomplish accurate risk analyses and the ability to make rapid changes in the pedigree and model offer the sort of attractive use that has made spreadsheets popular for many applications. We hope that further development of our prototype and a planned series of rigorous tests of its usability and usefulness will help to increase the timeliness and accuracy of genetic counseling.

Acknowledgments. This research has been supported (in part) by grant R01 LM 04493 from the National Library of Medicine. We thank Ms. Nomi Harris for her contributions to an earlier version of the present program.

5. References

[1] E. A. Murphy and G. A. Chase. *Principles of Genetic Counseling*. Yearbook Medical Publishers, Chicago, 1975.

[2] J. Ott. *Analysis of Human Genetic Linkage*. The Johns Hopkins University Press, Baltimore, Maryland, 1985.

[3] E. A. Thompson. *Pedigree Analysis in Human Genetics*. The Johns Hopkins University Press, Baltimore, 1986.

[4] J. R. Sorenson, J. P. Swazey, and N. A. Scotch, editors. *Reproductive Pasts, Reproductive Futures: Genetic Counseling and its Effectiveness*. Birth Defects Original Article Series XVII:4. National Foundation–March of Dimes, Alan R. Liss, New York, 1981.

[5] V. A. McKusick. Mendelian Inheritance in Man. The Johns Hopkins University Press, Baltimore, Maryland, ninth edition, 1990.

[6] C. J. Chapman. A visual interface to computer programs for Linkage Analysis. *Am. J. Med. Gen.* 36:155–160, 1990.

[7] N. Harris. Probabilistic belief networks for genetic counseling. *Computer Methods and Programs in Biomedicine*, 32:37–44, 1990.

[8] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[9] G. F. Cooper. Bayesian belief-network inference using nested dissection. Technical Report KSL-90-05, Stanford University, Knowledge Systems Laboratory, Stanford, California, Jan. 1990.

[10] P. Szolovits. Compilation for Fast Calculation over Pedigrees. In MacCluer, J. W. *et al.* (eds), Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes. *Cytogenetics and Cell Genetics* (in press). Basel: S. Karger

[11] Mamelka, P. Pedigree/Draw 4.3 Release Notes. Department of Genetics. Southwest Foundation for Biomedical Research, P.O. Box 28147, San Antonio, Texas, 1991.

[12] Yampratoom, E. Automatic Layout for Pedigree Diagram. Bachelor's Thesis, MIT Dept. of Electrical Engineering and Computer Science. 1991.