13

The Genetics Counselor

G. Hunn and J. Lederberg
Department of Genetics
Stanford University Medical School

Abstract

The Genetics Counselor is a computer program, written in Lisp, designed to handle problems of medical genetics counseling. It is an attempt to apply the methods of artificial intelligence research to medical diagnostic problems. The program attempts to map the data space of a family-tree structure into the hypothesis space of classical Mendelian genetics by use of a heuristic search.

The input data are the family members along with their children (or parents), and phenotype. The program generates a family tree and searches for consanguinity.

The hypothesis space consists of the Mendelian modes of inheritance. Heuristic conditions are used to search the data space to eliminate untenable hypotheses.

If more than one type of inheritance is possible, the program generates the probable genotypes of each member of the family based on the possible hypotheses. It then tests the expected phenotypic expression based on a particular inheritance against observed distribution of the phenotype. After investigating all the hypotheses not expressly invalidated, the program selects the most plausible hypotheses explaining the data, and orders the remaining ones. The likelihood of a parent having an affected offspring is estimated based upon the plausible hypotheses.

INTRODUCTION

The Genetics Counselor is an application of the techniques of artificial intelligence programming to the problem of computer-aided medical diagnosis. The area in which this program operates is medical genetics counseling. The problem characteristically involves the likelihood of a parent having an offspring affected by a particular trait or disease, when it is known that this disease is present somewhere in the family. Much of this counseling is human and interpretive on an emotionally-charged subject;

it is liable to be quite subjective. However, determining the mode of transmission of the trait and a probabilistic estimate of an individual having an affected offspring is an objective component of the counseling.

Medical genetics was selected as the test area for several reasons. It is an area of medicine which has a good theoretical biological basis from which hypotheses may be generated. However, the tasks are not trivial. It takes both physical effort to organize the data and mental energy to reach a definitive conclusion. It also is an area in which the theoretical framework can be codified. Needless to say, the potential usefulness of the program was also a consideration.

The computer program is an initial, very limited emulation of the inductive process in that it does not generate its own hypothesis, but selects from a set based upon genetic theory. We make few claims for its standing as an example of automated induction in its present form. However, the set of hypotheses is capable of expansion and is not limited by the program itself. The hypotheses are an attempt to explain the distribution of an inherited characteristic in a family. They are applied to data, and the most plausible ones are chosen. From the possible hypotheses and the data, predictions are made about the probability of the characteristic appearing in a future generation.

The program is intended to illustrate a difficult but sounder approach to the automation of medical diagnosis. Most prior attempts at automated medical diagnosis have centred about Bayesian statistical methods, and, more recently, decision analysis. The Bayesian approach is based only upon the probabilities of symptoms and diseases and does not consider their logical relationships. Given a set of symptoms, Bayesian diagnosis attempts to determine the probability of a disease, from the probabilities of the symptoms given the disease and the prior probabilities of the disease state themselves. Decision analysis is applied in an attempt to find the optimum path toward the solution of a sequential diagnostic problem. These techniques do not consider the logical relationship between the patient's symptoms and the disease.

Most medical diagnostic processes are inductive. A number of hypotheses are generated from the data which the physician gathers from the patient. Each of these invokes a particular model. The diagnostician then uses this model to explain the data. If he is satisfied that the hypothesis adequately explains the situation, he then invokes his model to prognosticate about possible future events. The genetics counselor attempts to follow this inductive approach.

The computer program is written in LISP and runs on an IBM-360/67 computer at the Stanford Computation Center. It is about 40,000 words. The program uses standard LISP programming techniques, making extensive use of the LISP property list facility for storing connection tables, relationships, and other important parameters. Usual run time for a family of about 40 members is between 1·5 and 3·0 minutes, running under the LISP interpreter.

REVIEW OF GENETICS

(The appendix is a glossary of genetic terminology used in the paper.)

The program is rooted in the classical Mendelian genetics and the principles of chromosomal inheritance. Normally, man has 46 chromosomes — twenty-two autosomal pairs and two sex chromosomes. Males have an X and a Y chromosome while females possess two X chromosomes. Genes are the basic units of heredity. The location of a gene on the chromosome is called a locus. Reproductive cells, sperm and eggs, each carry one-half of the somatic chromosomal complement, i.e., twenty-two autosomes (one chromosome from each pair) and one sex chromosome. Consequently, fifty per cent of each parent's genes are transmitted to an offspring. Sex determination is based on the presence or absence of the Y chromosome and, therefore, depends upon the contribution from the father.

At a given locus many different forms of a gene may be present. These multiple forms are called alleles. If the same allele is present on each chromosome of a given pair, the individual is considered homozygous. If different alleles are present, the individual is heterozygous. The genotype is the actual complement of genes present. The phenotype is the mapping of the genotype onto an observable characteristic (e.g., blond hair, blue eyes, etc.).

'Transmission of a trait', 'mode of inheritance', and 'genetic transmission', as used in this paper, refer to the manner in which the phenotype is related to genotype. It will depend upon the locus of the gene (i.e., whether it is on an autosomal or sex chromosome) and on the gene dose required for expression (i.e., whether a single gene of a pair is required for expression of a phenotype, or whether both alleles at a given locus are required). A dominant trait requires that only a single causative gene be present on one of a pair of chromosomes. Recessive traits depend upon the presence of the gene on both chromosomes. Sex-linked traits are carried on the X chromosome (transmission of traits on the Y chromosome have rarely, if ever, been demonstrated). Penetrance refers to the expression of the phenotype when the gene causing the trait is present. For example, in a dominant trait, if a phenotype is expressed every time the gene is present, it has a 100 per cent penetrance; however, if it is expressed less than 100 per cent of the time, it is considered incompletely penetrant. The same concepts can be applied to the recessive traits. Sex-limited traits are autosomally transmitted traits which, for some reason, are preferentially expressed in one sex over another (e.g., baldness).

GENETICS COUNSELING

The physician as a genetics counselor first collects data from the patient. This relates to family members and information which can be gathered about these family members extending as far back as the patient's memory will allow. The information usually contains the sex, family relationships, and phenotypes. Added information may be obtained from laboratory studies. The counselor then sets about the task of determining the mode of inheritance of the trait. He first organizes the data into a family-tree structure which specially indicates affected individuals (figure 1). Then applying rules of genetics, plus insight and intuition from past medical experience, he attempts to solve the problem. First, he determines if the trait appears to have a genetic basis. Then he solves the question of mode of transmission. Lastly, he attempts to determine the likelihood of a child being affected.

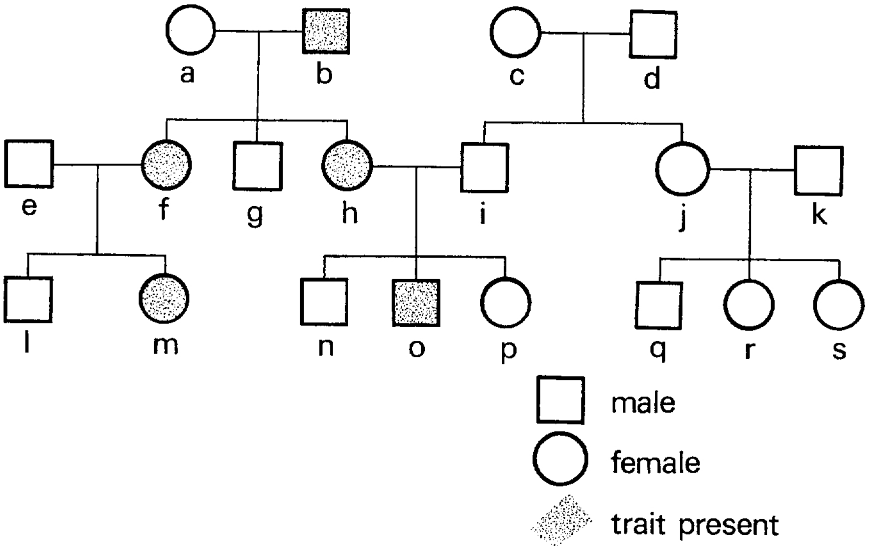

Figure 1. The family tree as is usually depicted. The distribution of affected individuals is characteristic of an autosomal dominant trait.

The genetics counselor is usually consulted by parents who have had an affected offspring, and want advice about having more children. Less frequently, he will be consulted by prospective parents who are either affected themselves or have a disease in their family and are concerned about the likelihood of offspring being affected. Along with their questions, the parents may bring many anxieties. These can relate to the guilt feelings of producing an abnormal child, or the feeling that there is something morally wrong with carrying a defective gene, or other personal problems which may be focused onto the question of childbearing. Most parents have basic anxieties of giving birth to abnormal children.

Throughout the counseling experience these subjective considerations should be paramount. The counselor must be able to relate the facts about inheritance and probabilities of having affected offspring to emotional difficulties the parents experience. The counseling process revolves around this interaction.

In general, genetic counseling produces a desirable effect in those seeking advice. It helps allay many fears and anxieties, and presents the parents with facts upon which a decision can be based. Counseling tends to dissuade high-risk parents from having children, while it reassures low-risk parents.

THE PROGRAM

The computer program attempts to solve the problem in much the same manner as the human problem-solver. It follows the general approach of organizing the data into a family tree, applying the rules of genetics to the information, and attempting to reach some logical conclusion about the mode of transmission of the trait. The mode of transmission is the hypothesis the program attempts to apply to the data.

The program deals only with single factor inheritance. This means that the disease or trait under consideration is produced by the action of genes present at a single locus. Though the program handles only this type of trait, it can handle different traits. It does this sequentially. If you want to determine the mode of transmission of several traits within the same family, it will do this, and can separate an individual mode of transmission for each trait. At present, it cannot handle the question of genetic linkage between these traits, but this would be an easy extension.

Conceptually the program is divided into three parts: (1) data input and organization; (2) hypothesis testing; and (3) decision making. There is some overlap between the second and third portions. The body of the program is the hypothesis testing.

1. Data input and organization

The input consists of the names of the traits which are to be considered along with the frequencies of the associated genes in the general population. If the gene frequency is not known, then the estimate of whether the trait is common or rare is entered. The individuals, along with their sex, children (or parents), and phenotype under consideration, form the primary data. Usually, only the children of a given person are entered. The only restriction on the input is that the names of the individuals and the traits are unique identifiers within the population under consideration and can be interpreted as LISP atoms.

The parents (or children), siblings, uncles, aunts, cousins, half relationships, and grandparents and grandchildren of each person are determined by the program. No special algorithm is used for this. The program merely knows the meaning of the terms as commonly used to identify the family relationships. For example, it knows that siblings have the same parents, and uncles and aunts are the sibs of one's parents, etc.

At this point, the population is characterized by a direct graph in which the nodes are the individuals and the links are the relationships of parents and offspring. (*See* figure 2.) Each individual is connected only to the nodes representing his parents and offspring. Multiple families within a population may be considered, with each family being considered separately. The nodes within a family are ordered according to a hierarchy of generations. This ordered graph is the family tree. Each individual's connections and generation are stored on its property list.

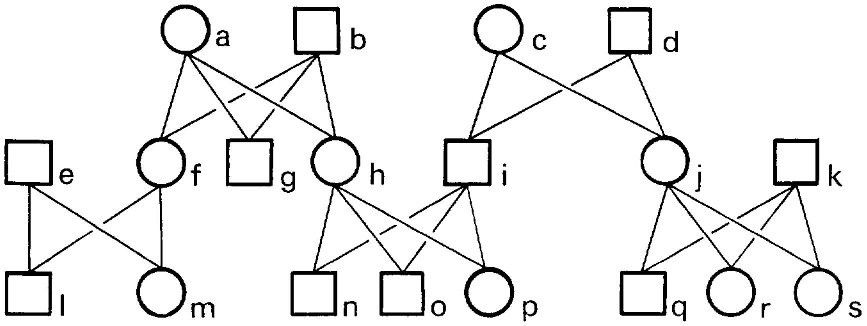

Figure 2. This graph represents the same family as figure 1 and shows the connections as represented in the program.

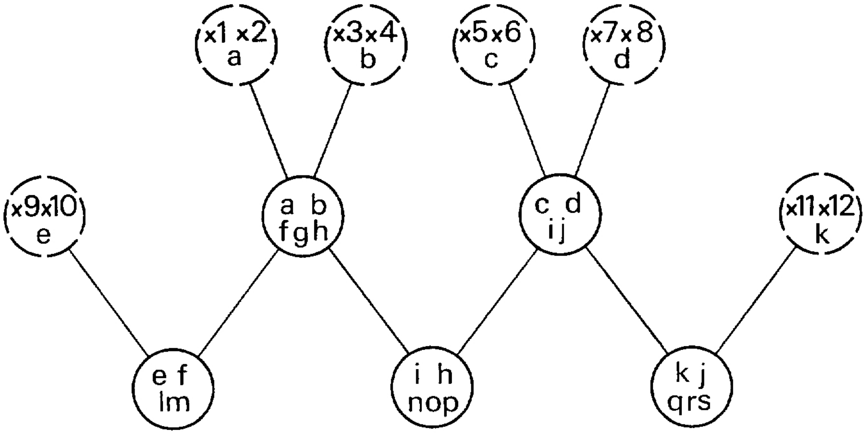

Figure 3. Family-tree graph in figure 2 transforms into this graph in searching for consanguinity (not present in this family). The individuals whose first letter is '*x*' represent the family members generated by the LISP atom generator (GENYSM1).

An important part of the program is the search for consanguinity within a family. Consanguinity refers to the marriage of persons with a common ancestor. This is done by a ring-finding algorithm. In the graph of the family tree shown in figure 2, many rings are present which do not represent consanguinous relationships, and, consequently, a transformation of this family-tree graph into another graph form is desirable (as in figure 3). Each node of this transformed graph consists of an individual family unit, that is, mother, father, and children. An edge leading from a node comes from a child and extends to the successor node. If the child has no offspring, then no link is directed away from that node for that child. The edges entering the node come from a preceding node in which that parent is a child. If a parent or parents of an individual are not known, the LISP atom generator (GENSYM1) generates a unique atom which will act as the unknown parent in the analysis, and the node is formed consisting of the two new atoms plus the individual.

This graph is then searched for rings. The presence of a ring indicates consanguinity. If any are present the ring-finding algorithm determines all the ring structures in the graph. This algorithm first prunes terminal chains from the graph. It then looks at the nodes which have two input lines, and searches for another path between those preceding nodes. If any two rings contain identical segments the program will combine them into a third larger ring. Rings which do not represent consanguinity (e.g., siblings married to another set of siblings will appear as a ring) are unusual. However, if they do appear, they are not considered for further analysis. All the consanguinous marriages are determined from these rings. The non-marital relationship (i.e., first cousin, second cousins, etc.), the offspring of these marriages, and the coefficient of inbreeding of these marriages are also found from the ring structures. The coefficient of inbreeding measures the probability that a single gene in an individual comes from an identical ancestor. It is determined from the number of edges on the path between the consanguinous parents. An average coefficient of inbreeding for the entire population is also calculated.

2. Hypothesis testing

The hypotheses considered are the more common modes of transmission of a trait in a family. These are dominant with 100 per cent penetrance, dominant with incomplete penetrance, recessive, sex-linked recessive, sex-linked dominant, and sex-limited traits. Other hypotheses could be considered, but these would rarely be true or the data necessary to validate them would not, in general, be present in a single family. Though this limited number has been selected, there are no limitations within the program to the number of hypotheses that may be considered. More hypotheses may be added or deleted at one's discretion.

Each hypothesis has a set of conditions (table 1) associated with it. Each condition is encoded as a separate LISP predicate function. These conditions either are based upon the theory of chromosomal inheritance, or are heuristics gleaned from textbooks or empirical inspection of family-tree data considered representative of a particular hypothesis. Most of the conditions are standard. However, some may be open to question. The number of conditions associated with each hypothesis is variable, and ranges from three, for a completely dominant trait, to eight for a sex-linked recessive trait. Some of the hypotheses (e.g., sex-linked recessive) have a set of associated conditions which are firmly based in theory. Others (e.g., dominance with incomplete penetrance) are associated with a more heuristic set.

Table 1. Conditions associated with each mode of genetic transmission.

1. Dominant trait – 100% penetrance
   - *a.* No skipped generations
   - *b.* All affected individuals have affected parents
   - *c.* 50% of offspring of affected individuals have trait
2. Dominant trait – incomplete penetrance
   - *a.* At least one affected individual has an ancestor who is affected
   - *b.* No abnormal sex distribution
   - *c.* At least one affected individual has unaffected parents
   - *d.* Trait appears in more than one generation
3. Recessive trait. (If trait is common only *c*, *d*, *e* are considered, otherwise all five conditions used.)
   - *a.* (1) Trait present in fewer than 25% of the families in the family tree, *or* (2) present in only one generation, *or* (3) consanguinity present in more than one generation
   - *b.* Not present in parents of affected individual unless the parents are products of consanguinous marriage
   - *c.* In sibships with the trait present, 25% of the total are affected
   - *d.* If one parent is affected, 50% offspring have trait
   - *e.* If both parents affected then all offspring affected
4. Sex-linked recessive trait
   - *a.* Unusual sex distribution
   - *b.* Male predominantly-affected sex
   - *c.* No male carriers
   - *d.* No affected male has an affected son
   - *e.* All affected females have affected fathers
5. Sex-linked dominant trait
   - *a.* Dominant trait
   - *b.* No affected male has an affected son
   - *c.* Some females affected
   - *d.* All daughters of affected males are affected
6. Sex-limited trait (XSEX – sex predominantly affected, LSEX – other sex)
   - *a.* Abnormal sex distribution
   - *b.* All XSEX children of LSEX affected individuals have trait
   - *c.* All XSEX parents of an LSEX affected individual have trait
   - *d.* 50% XSEX offspring of XSEX parents are affected

These conditions may be changed arbitrarily, either by addition, substitution, or deletion. Thus, with a small amount of programming, either the hypothesis set or the related set of associated conditions may be altered. At present this must be done by a human operator.

To test each hypothesis, the set of associated conditions is applied to the family tree in question. If for any given hypothesis a condition is not true for that family tree, that particular mode of inheritance is eliminated. The remaining hypotheses are then placed on a global list for further evaluation.

If all the modes of transmission are eliminated, the program weakens the criteria by eliminating some of the heuristic conditions from each hypothesis. It then applies this less stringent set of conditions to the family data. If all the hypotheses are again eliminated, the program concludes that the trait does not fit into any of the modes of inheritance with which it is familiar. This means that either the program does not know enough genetics, or the trait is not inherited.

The plausible hypotheses are checked against the family data. A genotype is determined for each individual based upon the hypothesis under consideration and the available phenotypic information. From these genotypes an expected distribution of phenotypes is calculated and then compared with the actual distribution. This is done for each remaining hypothesis.

The generation of genotypes from the phenotypes may be straightforward, particularly if the population gene frequency is known. However, if this is not known, and if there is missing information concerning the presence or absence of the trait, the program tries to apply some degree of reasoning, as explained below. If there is just too little information, the program will guess, using as much information as it can.

Some strategy is used in determining the genotypes. First, the genotypes which are obligatorily determined by the particular mode of inheritance under consideration are calculated. For example, if the hypothesis is that the trait is recessive, any individuals who exhibit the phenotype will be considered homozygous for that particular gene. Next, the probable carriers, that is, those who possess the gene, but who do not manifest the trait, are found. Family members whose phenotype is known, but whose genotype is not determined with certainty, on the basis of the hypothesis under consideration, are considered next. This list of people is first ordered by generation, and the genotypes are determined in this ordered sequence. The genotype for an individual in this latter group is estimated on the basis of his parental genotype (if already known) or an estimate of the probable parental genotype. This latter estimate is determined by the population gene frequency, if known, and is determined in conjunction with phenotypic and genotypic information about siblings, grandparents, etc. This parental-based probable genotype is then conditioned upon the phenotypes of his offspring. The final genotype determined is actually a list of probabilities that the individual is homozygous for either of the two alleles, or is heterozygous.

The problem of determining genotypes of those people whose phenotype is unknown is tackled next. This is done in much the same manner as is described in the preceding paragraph. However, this may be particularly troublesome if the population gene frequency is not known or parental or sib information is unavailable. It is at this point that the program will resort to the educated guess. An example of the 'guess logic' is: the trait is rare, there is no evidence of near relatives possessing the trait (or no information about near relatives), so let us assume he does not possess the gene which produces the trait.

From the genotypes, expected phenotypes of offspring are determined. For each mating, the expected proportion of genotypes of offspring is found. These are summed over the entire family. The genotypes are then mapped into the phenotypes and the expected number of persons with the trait present and the expected number with the trait absent are calculated. These are compared to the observed values. As an example, a family of 20 may be expected to have 8 homozygous dominant (DD), 6 heterozygotes (Dd), and 6 homozygous recessive (dd). With the recessive hypothesis, 6 individuals with the trait present and 14 persons with the trait absent may be expected. This is then compared with the observed values.

If there is a marked discrepancy (as determined by a χ-square test with a cutoff of $p=0{\cdot}10$), the hypothesis is rejected. This genotype forecasting and phenotype comparison is done for all the possible hypotheses. Those remaining after this are then evaluated by the decision maker. If all the hypotheses have been eliminated, the program iterates the hypothesis testing using the less stringent rules. If no more tenable hypotheses are found, the program will give up.

3. Decision making

Most often, at this point, only one plausible hypothesis remains and no decision is necessary. If there is more than one possible mode of transmission, a decision as to the most plausible and an ordering of the remaining ones are done. Several heuristics are used to select the most plausible. For example, if possible recessive transmission is compared with another mode, and some of the affected individuals are products of consanguinous marriages, the recessive mode is considered most likely. If none of the heuristics are applicable, or they conflict, the ordering of the probability of the hypothesis explaining the data (as determined by the χ-square test of the preceding section) is used to rank the hypotheses. If these probabilities are not too different, then the respective modes of transmission are considered equally likely.

THE PROGRAM'S GENETICS COUNSELING

The program gives the possible modes of inheritance, and the probability of a particular set of parents having affected offspring, based upon a given mode of transmission. It also returns a list of carriers of the abnormal gene. The modes of transmission and the estimate of the likelihood of having affected offspring is of foremost importance in counseling the prospective parents. The list of carriers of the gene may be used for further study of the family. A closer investigation of these people may be indicated in order to determine if some subclinical form of the trait is present. From further investigation a better understanding of the trait and its genetic mechanisms may be acquired.

The program cannot take into account the subjective aspects of counseling. The information related by the program must still be interpreted and used by the human counselor. The manner in which the information is related to the parents is of utmost importance. Perhaps the program can be used to help parents better understand the probabilistic aspects of the counseling. Monte Carlo simulations could generate chances in a game to learn what $p=0{\cdot}x$ means as a subjective probability. However, because of the complexities involved, the subjective interactions are best left to the human counselor.

DISCUSSION AND SUMMARIES

The Genetics Counselor is a program specifically designed to perform in a highly-specialized area. It makes no pretence to be a general problem-solver. Certainly, its capabilities can be extended to solve more complex genetics problems, e.g., multiple alleles, or linkage problems.

The task which the program accomplishes is based upon a relatively small set of possibilities. However, it is operating in an area where human intellect and insight are generally required. It would take a human 15–30 minutes to do the same job as the program now accomplishes in 1·5 to 3·0 minutes. (This time will decrease by a factor of ten when the program is compiled.)

The operation and strategies of the program can be summarized in the following LISP function. (This function is not used in the program but represents its conceptual organization. This program provides for the input of data, the organization of the family tree, a search for consanguinity and selection of plausible hypotheses. The portion between L1 and L2 deletes those hypotheses which do not fit the data well and makes a decision as to the ordering of the hypotheses. It then returns the pertinent information.)

```
DEFINE (((GENETICSCOUNSELOR (LAMBDA (FAMILY DATA
        RULES PARENTS) (PROG (X L)
      (MAPCAR FAMILY (FUNCTION (LAMBDA (J) (PUTPROP J
        DATA))))
      (FINDRELATIVES FAMILY)
      (SETQ FAMILYTREE NIL)
      (FINDCONSANGUINITY)
      (CSETQ POSSIBLEHYPOTHESES (TESTHYPOTHESES
        RULES FAMILY))
      (COND ((NULL POSSIBLEHYPOTHESES) (GO L1)))
L2          (FINDGENOTYPE POSSIBLEHYPOTHESES FAMILY)
            (SETQ L POSSIBLEHYPOTHESES) L3 (SETQ X (CAR L))
            (COND ((LESSP (CHISQUARE (EXPECTEDPHENOTYPE X)
              DATA) 0.1) (DELETE X POSSIBLEHYPOTHESES)))
            (SETQ L (CDR L)) (COND ((NOT (NULL L)) (GO L3)))
            (COND ((NULL POSSIBLEHYPOTHESES) (GO L1)))
            (CSETQ POSSIBLEHYPOTHESES (DECISION
              POSSIBLEHYPOTHESES))
            (RETURN (LIST POSSIBLEHYPOTHESES
              (PROBAFFECTEDCHILD PARENTS)))
L1          (CSETQ POSSIBLEHYPOTHESES (TESTHYPOTHESES
              (MODIFY RULES FAMILY)))
            (COND ((NULL POSSIBLEHYPOTHESES) (RETURN
              (LIST (QUOTE $$$EITHER THE DATA DOES NOT FIT
              THE HYPOTHESES$) (QUOTE $$$ OR I DO NOT KNOW
              ENOUGH $) )) ))
            (GO L2) ))) ))
```

The program has not as yet been tried in a clinical environment. However, it has been applied to sample cases selected in a somewhat random fashion. Its performance in selecting the correct mode of inheritance has been good. It usually eliminates all but one or two possibilities and selects the correct one. In cases where ambiguities are present it will reduce the selection to these ambiguous ones.

The portion of the program concerned with organizing the data into family tree structures, and finding consanguinity, is being applied to population studies. The population of small Indian villages in Guatemala is being organized into family structures. This is a fairly large undertaking because the size of the populations range from 100–1600 individuals. Ultimately, we will apply the entire genetics program to these populations to investigate the transmission of several characteristics, particularly blood groups.

Computer-based genetics counseling will certainly allow for broader dissemination of the service. This is important in an era where diagnosis, treatment, and prevention of genetic disease are feasible. In a large university medical center, there are usually well-trained geneticists who engage in counseling. However, in the general medical community there is an inadequate supply. Computer-based genetic counseling can be used in conjunction with trained paramedical personnel to deliver better this service to the community at large.

BIBLIOGRAPHY

Ledley, R.S. (1969) Practical problems in the use of computers in medical diagnoses. *Proc. I.E.E.E.*, **57**, 1900–18.

Stern, Curt (1960) *Principles of human genetics*. San Francisco: W.H. Freeman.

APPENDIX: GLOSSARY

Alleles, alternative forms of genes which may be present at the same locus.

Autosomes, chromosomes not concerned with primary sex determination.

Chromosomes, long strands of DNA which contain the genetic information.

Coefficient of inbreeding, probability that the genes present at the same locus on the paired chromosomes are descendent from a common ancestor.

Consanguinity, marriage between persons with a common ancestor.

Dominance, expression of phenotype when gene is present either on one or both pairs of chromosomes.

Genotype, the genetic composition of an individual.

Heterozygote, person who possesses different alleles at a given locus.

Homozygote, person who possesses identical alleles at a particular locus.

Linkage, genes localizable to the same chromosome.

Locus, location of a gene on the chromosome.

Penetrance, proportion of individuals of a given genotype who manifest the corresponding trait.

Phenotype, observable characteristic, mapping of the genotype into the environment.

Recessiveness, expression of a phenotype only when allele is present in the homozygous state.

Sex-limited, autosomal trait, which for some reason is expressed differently in the two sexes.

Sex-linked, trait produced by gene that is located in the X chromosome.