Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

PedExpert: a computer program for the application of Bayesian networks to human paternity testing

R.R. Gomes[1,2], S.V.A. Campos[2] and S.D.J. Pena[2,3]

[1]Departamento de Bioquímica e Imunologia,
[2]Departamento de Ciência da Computação,
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[3]GENE - Núcleo de Genética Médica, Belo Horizonte, MG, Brasil

Corresponding author: S.D.J. Pena
E-mail: spena@dcc.ufmg.br

Genet. Mol. Res. 8 (1): 273-283 (2009)
Received December 15, 2008
Accepted January 5, 2009
Published March 10, 2009

ABSTRACT. PedExpert is a Windows-based Bayesian network software, especially constructed to solve problems in parentage testing that are complex because of missing genetic information on the alleged father and/or because they involve genetic mutations. PedExpert automates the creation and manipulation of Bayesian networks, implementing algorithms that convert pedigrees and sets of indispensable information (genotypes, allele frequencies, mutation rates) into Bayesian networks. This program has a novel feature that can incorporate information about gene mutations into tables of conditional probabilities of transmission of alleles from the alleged father to the child, without adding new nodes to the network. This permits using the same Bayesian network in different modes, for analysis of cases that include mutations or not. PedExpert is user-friendly and greatly reduces the time of analysis for complex cases of paternity testing, eliminating most sources of logical and operational error.

Key words: Bayesian networks; Probabilistic expert systems; Algorithm; Software; Paternity testing; DNA

INTRODUCTION

The need to establish paternity relationships often arises legally, in paternity courts, socially, in private paternity disputes, or medically, in, for example, prenatal diagnosis, genetic counseling and transplantation. In the United States alone, 420,740 “legal” paternity tests were carried out in 1996 (AABB, 2006).

Paternity testing is predicated on a Popperian logical asymmetry between verification and falsification (Pena and Chakraborty, 1994). Exclusions of paternity are logically irrefutable and were found in 89,890 (25.85%) of the cases examined in the United States in 2006. Proof of paternity, on the other hand, depends on statistical inference. This can be achieved by calculating a ratio of the likelihood of obtaining the observed set of findings given that the alleged father is the true one (X) over the same likelihood in the hypothesis of a random man being the father (Y). Such likelihood ratio (X/Y), called the paternity index (PI), provides the odds for paternity of the alleged father. Data from the American Association of Blood Banks (AABB, 2006) indicate that for all the non-excluded cases, high levels of combined paternity index could be obtained, providing strong evidence for paternity.

Data from the American Association of Blood Banks (AABB, 2006) also indicate the occurrence of a remarkable technological convergence in DNA paternity testing - 98.53% of the American cases were performed using multiplex microsatellite (short tandem repeat) genetic markers, tested using the polymerase chain reaction. This technology has been shown to be perfectly capable of resolving all simple cases in which the three parties (father, mother and child) are available for testing.

A different and much more complex situation arises when the father cannot be studied directly, generally because he is deceased. The basic strategy then is to try to reconstitute his genetic profile from living relatives, which can be children, sibs or parents. Although such reconstitution can occasionally be achieved using algebraic calculations, this approach can be rather complex from a logical stand point, time-consuming and error-prone. Dawid et al. (2002) have led the way in demonstrating how such cases can be solved using expert probabilistic systems, more commonly called Bayesian networks.

In their article, Dawid et al. (2002) indicate solutions for diverse complex problems in paternity testing without, however, formalizing algorithms for the construction of Bayesian networks. Their flexible approach makes possible the use of general Bayesian network software packages such as GeNIe (http://genie.sis.pitt.edu/) and Hugin (http://www.hugin.com/) to solve the problems. However, the use of these generic programs for the analysis of complex cases of paternity testing demands the construction of multiple tables for each DNA marker used and is still predisposed to errors and very labor-intensive (Cowell, 2003).

Thus, we have written, PedExpert, a Windows-based Bayesian network software especially constructed for solving problems in parentage testing that are complex because of missing genetic information on the alleged father and/or because they involve genetic mutations.

PedExpert has the advantage of creating the structure of Bayesian networks directly from family pedigrees, which are easily built and understood and constitute part of the day-to-day routine of geneticists. Internally, Bayesian networks will contain genetic data obtained from the available individuals in the pedigree structure and will have embedded tables of conditional probabilities constructed according to Mendelian principles making use of allele frequencies and mutation rates of the DNA polymorphic loci typed.

GENERAL SOFTWARE DESCRIPTION

PedExpert operates in the GeNIe (GraphicalNetwork Interface)/SMILE (Structural Modeling, Inference, and Learning Engine) environment developed by the Decision Systems Laboratory of the University of Pittsburgh (http://genie.sis.pitt.edu/). It was implemented on the Microsoft.NET platform using the development environment Borland Delphi 2006.NET. More specifically, PedExpert makes use of SMILE.NET, a version of SMILE embedded in a DLL (Dynamic link library) and containing the main classes and methods of SMILE API. SMILE.NET was created for the Microsoft.NET platform utilizing Microsoft Visual C++, but it can be used with any programming language supported by Microsoft.NET.

PedExpert automatically constructs and manipulates Bayesian networks, implementing algorithms that have the function of converting pedigrees and the sets of indispensable information (genotypes, allele frequencies, mutation rates) into Bayesian networks that can be run on the SMILE environment. The database of PedExpert uses the Paradox software, currently commercialized by Corel Corporation.

A technical article providing a detailed description of the algorithms used in PedExpert is currently being finalized (Gomes RR, Campos SVA and Pena SDJ, unpublished results).

USING THE SOFTWARE

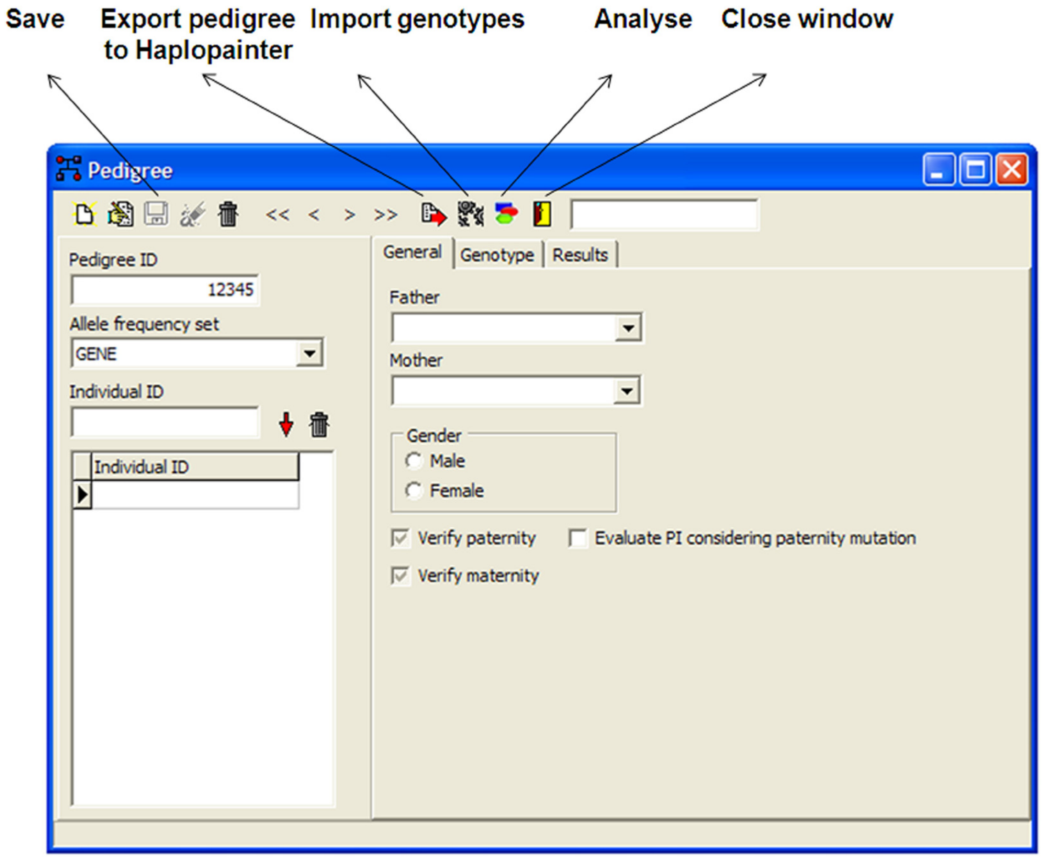
Inputting locus information

After loading PedExpert the user is shown a window that has only three menu options: File, Data and Close. When the Data menu is activated, a drop-down menu appears with five choices: New allele frequency set, Loci/Alleles, Import allele frequency, Import mutation rates, and Parameters. The user can input as many sets of marker loci as desired and these are kept stored in the program's database. The sets can include microsatellites, indels, single nucleotide polymorphisms, or combinations of these. The allele frequencies at each locus are entered into the database, as are the paternal mutation rates by using the respective tags. In Parameters, the user can input a default allele frequency and a default mutation rate. These are useful when a given locus has not had its mutation rate established or when alleles that are not in the database appear. These basic genetic data will be utilized by the program to build the conditional probability tables.

Entering data

When the File menu is activated, a drop-down menu appears with two choices: New pedigree and Open pedigree. PedExpert stores all previous runs in the Paradox database management system under a numeric code and all previously analyzed cases can be recalled by Open pedigree.

If New pedigree is chosen, a new window opens (Figure 1) and two fields have to be immediately completed in the front-end. A Pedigree ID field has to be typed in and an Allele frequency set choice box allows the user to choose from whichever set of loci/alleles/mutation rates are available in the database. With that finished, the case has to be saved immediately, which makes available all the other keys in the front-end.



Pedigree ID: 12345

Allele frequency set: GENE

Individual ID

Father

Mother

Gender
○ Male
○ Female

☑ Verify paternity
□ Evaluate PI considering paternity mutation
☑ Verify maternity

Figure 1. Screen shot of PedExpert's front end. The function of some of the main icons is shown. PI = paternity index.

Running the program

In order to better explain how the program performs the paternity analysis, we use as example a hypothetical case whose pedigree is shown in Figure 2. This is a complex paternity case involving a child (possible son = PS), indicated by an arrow in the pedigree, who wants to know if he is the son of a deceased man (possible father = PF). The mother of the child (mother of possible son = MPS) is alive and available for testing, as are one daughter of the alleged father (D), two brothers (B1 and B2) and the widow (W). The parents of the deceased possible father (GF and GM) are both also deceased.

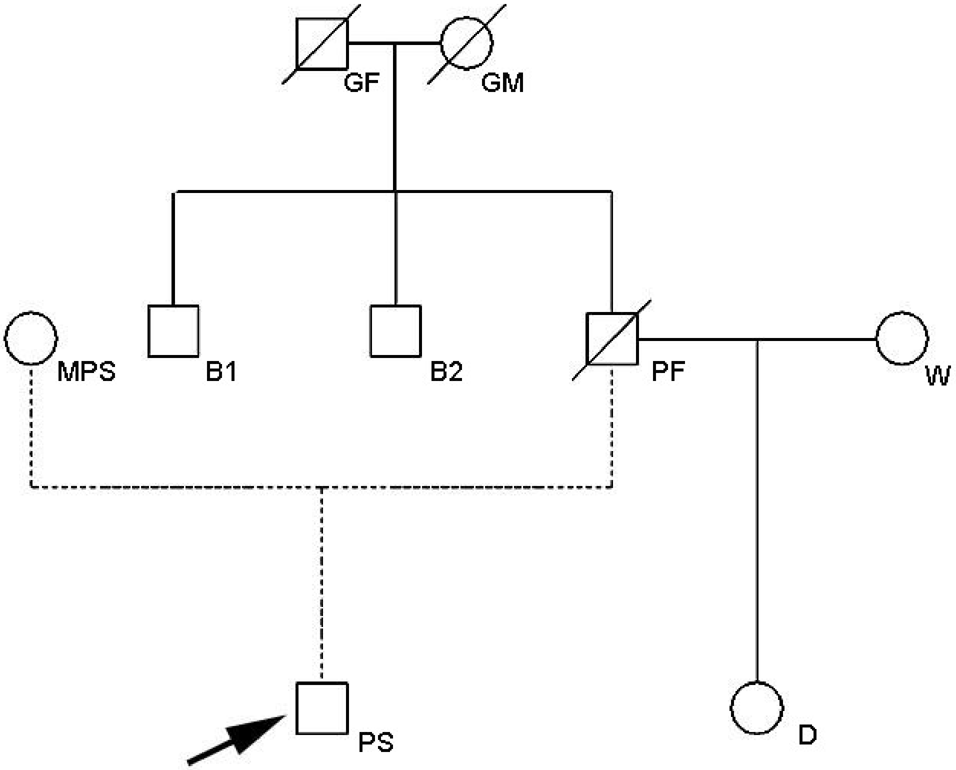


Figure 2. Pedigree of a hypothetical case involving paternity testing after the death of the alleged father. The possible son, product of an illegitimate relationship (dotted lines) is indicated with an arrow. To solve this paternity test one should infer the possible genotypes of the deceased possible father from genotypic information from his matrimonial daughter, the widow, and two brothers.

From this pedigree, PedExpert constructs a Bayesian network using SMILE.NET provided commands. Such network can be opened on GeNIe and is schematically shown in Figure 3. In reality, it is necessary to set up a Bayesian network for each locus that is typed, since the genetic parameters (allele frequencies and mutation rate) are locus-specific. However, the network structure remains the same, since it depends only on the pedigree structure. This need of having a different network for each locus is the main reason why using generic software is slow and labor-intensive.

The Bayesian network is composed of nodes, each one representing an allele belonging to a person. Thus, the Bayesian network has a structure analogous to a pedigree, which is evident from comparison of Figure 2 and Figure 3 (we have put dotted squares indicating the different individuals in the Bayesian network). The people in whom genetic tests were actually performed and whose results will be entered into the network can be identified by having associated input genotype nodes, shown in yellow.

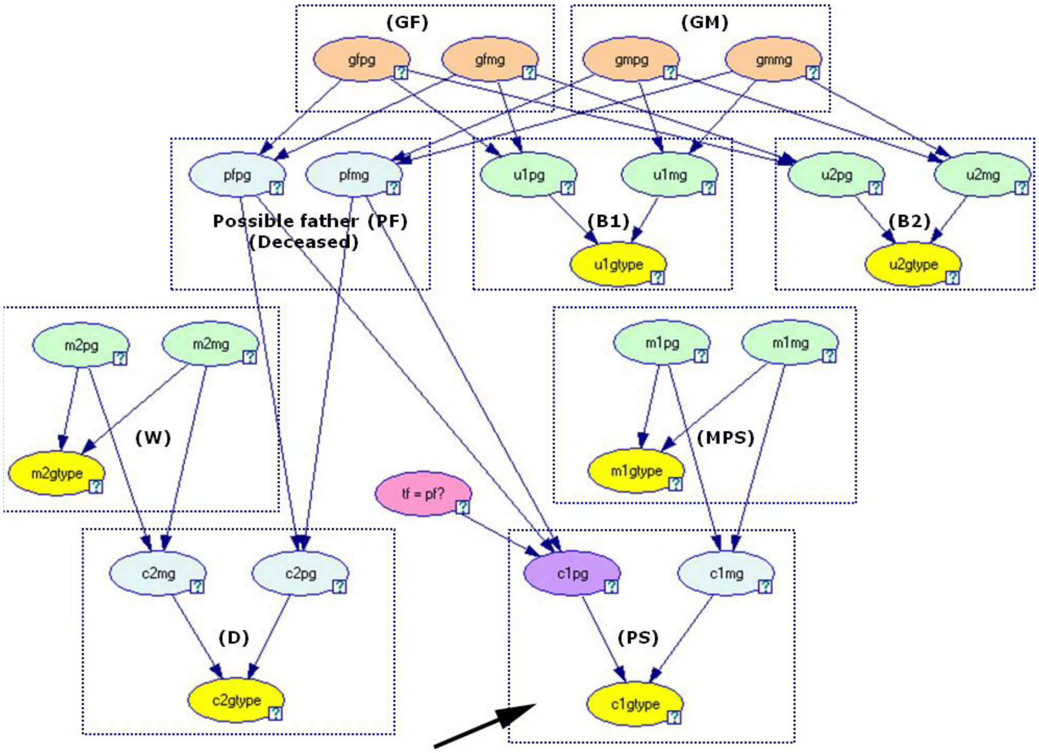
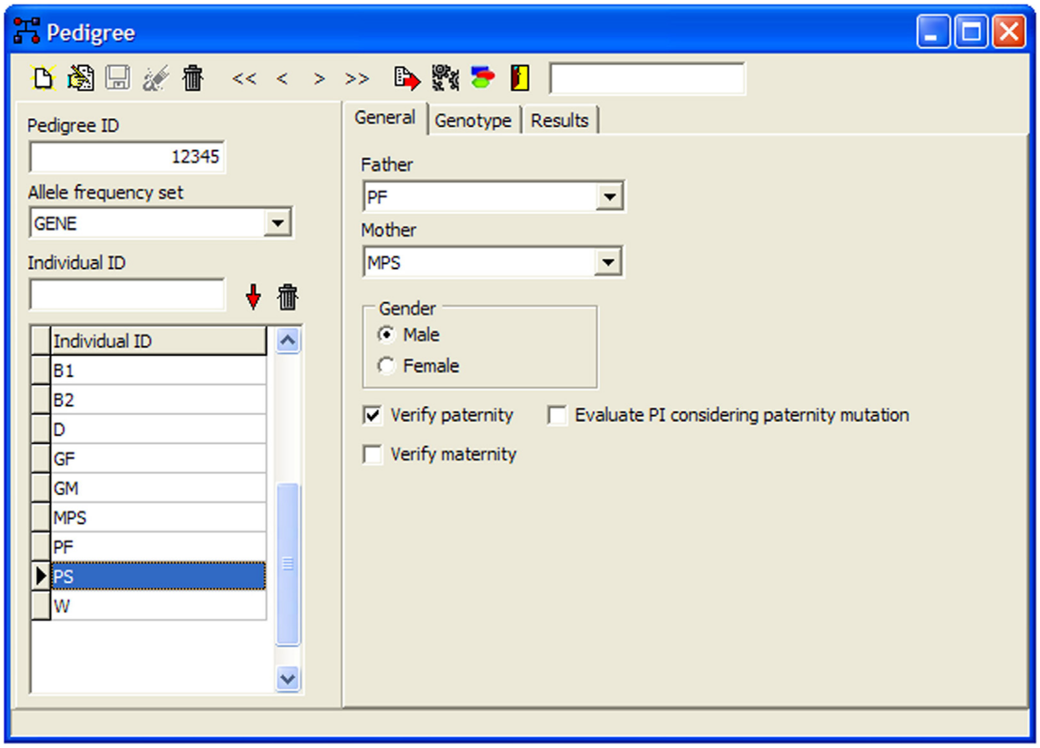


Figure 3. A Bayesian network corresponding to pedigree of Figure 2. The network is composed of nodes, each one representing an allele belonging to a person. Thus, the Bayesian network has a structure analogous to a pedigree, as seen from comparison with Figure 1 (we have put dotted squares indicating the different individuals in the network). The people in whom genetic tests were actually performed and whose results will be entered into the network can be identified by having associated input genotype nodes, shown in yellow. For abbreviations, see legend to Figure 4.

Thus, the next step in running PedExpert is to input the DNA typing data, which can be easily done by using the Import genotypes icon and pasting the genetic data from an Excel sheet. The program infers the sex of the tested individuals from the results of the *Amelogenin* locus. With that completed, the tested individuals will appear in the PedExpert window (Figure 4). The persons that were not tested, i.e., the possible father (PF) and his parents (GF and GM) have to be entered manually, with their respective sexes.

Next, we have to assign a father and a mother to the family members, thus establishing the network structure. For instance, in Figure 4 it is indicated that the child (PS) has MPS as mother and PF as father (this relationship is the one that will be in fact tested, as indicated by clicking the box Verify paternity). The choice of mother and father for all individuals is made using drop-down menus in the buttons Father and Mother (Figure 4). After we assign the mother and father to the individuals, we can use PedExpert to automatically export the data to the public domain pedigree-drawing software Haplopainter



Pedigree ID: 12345

Allele frequency set: GENE

Individual ID

Individual ID
B1
B2
D
GF
GM
MPS
PF
PS
W

Father: PF

Mother: MPS

Gender
● Male
○ Female

☑ Verify paternity ☐ Evaluate PI considering paternity mutation

☐ Verify maternity

Figure 4. Screen shot of PedExpert's front end after entry of the family information, i.e., the genotypes of tested individuals and the identity of untested family members. The identities of all individuals are the following: PF = possible father; PS = possible son; MPS = mother of possible son; D = daughter of the alleged father; W = widow; B1 and B2 = brother 1 and brother 2; GF and GM = parents of the deceased possible father. The individual whose paternity is being tested (PS) is identified by checking the box "Verify Paternity."

(Thiele and Nürnberg, 2005), using the icon Export pedigree to Haplopainter. The resulting pedigree, which is shown in Figure 5, contains the same information as the one in Figure 2, and is very useful for making sure that all family relationships have been correctly specified. Now, we can calculate likelihood ratios propagating information at internally created networks by clicking on the Analyse icon.

The results of the network run will appear on the tab Results (Figure 6). The likelihood ratio (paternity index) for each locus is then displayed on a table, at the bottom of which are shown the Combined Paternity Index (PI) and the probability of paternity, calculated from the PI using Bayes theorem with an *a priori* probability of 0.5. The complete contents can be exported to the Windows transfer buffer using the icon on the left top corner of the table, and from there pasted onto an Excel spreadsheet for further calculations, if necessary, or to a report form.

The analysis of a complex case such as this one using 30 microsatellite loci, applying the methodology proposed by Dawid et al. (2002) and utilizing the GeNIe software, would take ca. 2.5 h even when managed by an experienced user. The analysis of the same case using PedExpert takes less than 5 min!

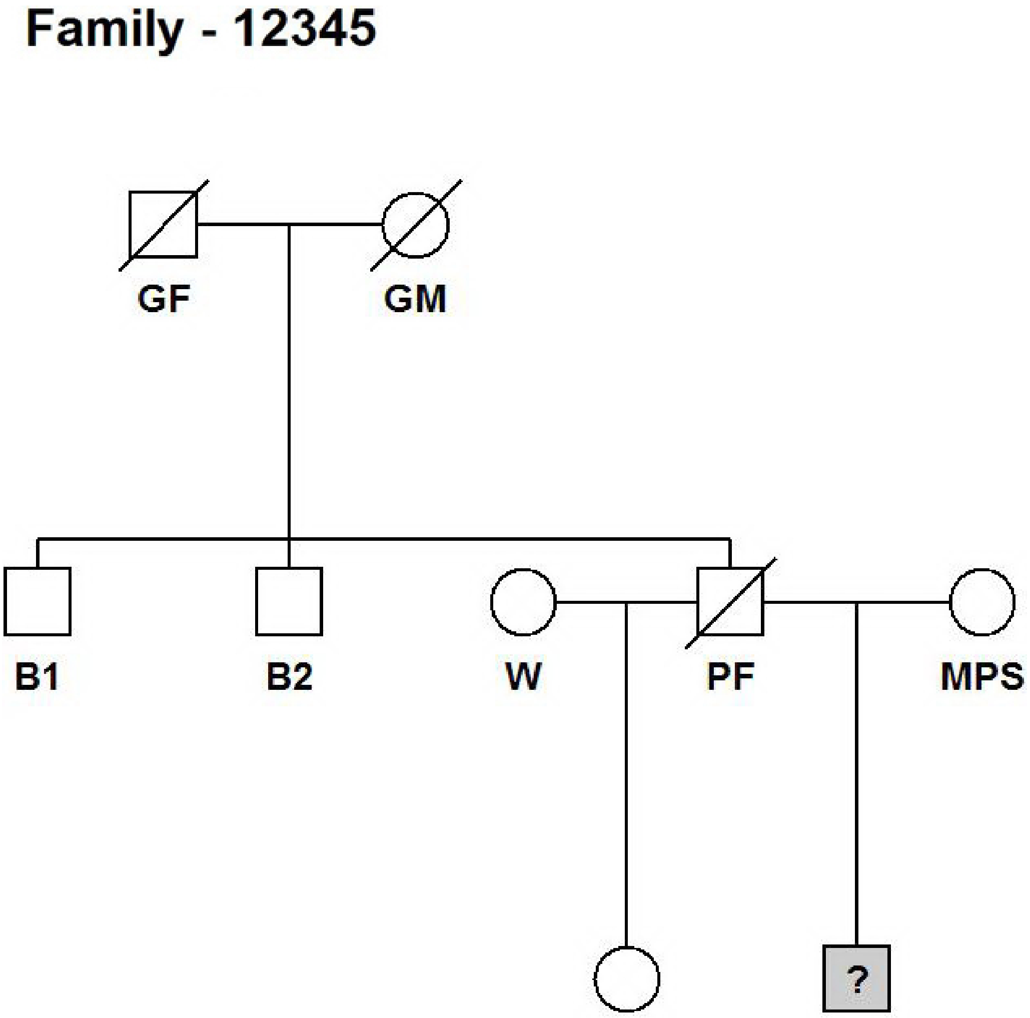
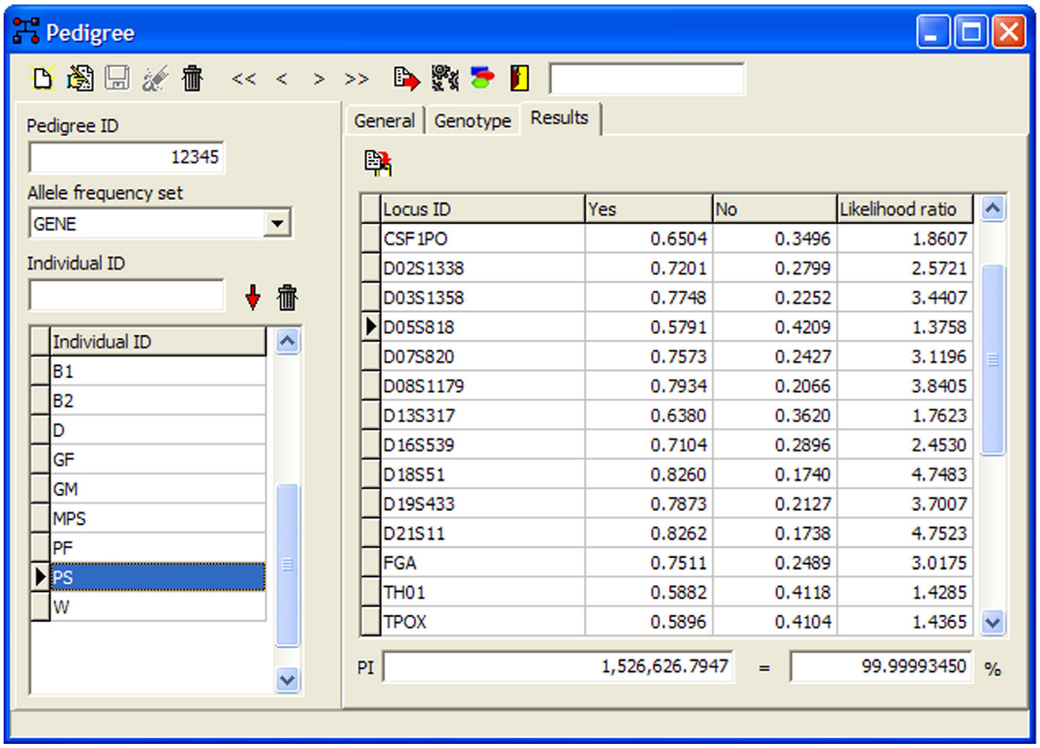


Figure 5. PedExpert automatically exports the family data to the public domain pedigree-drawing software Haplopainter (Thiele and Nürnberg, 2005). The resulting pedigree contains the same information as the one in Figure 2, and is very useful for making sure that all family relationships have been correctly specified. For abbreviations, see legend to Figure 4.



Pedigree ID: 12345

Allele frequency set: GENE

Individual ID:

Individual ID
B1
B2
D
GF
GM
MPS
PF
PS
W

Locus ID	Yes	No	Likelihood ratio
CSF1PO	0.6504	0.3496	1.8607
D02S1338	0.7201	0.2799	2.5721
D03S1358	0.7748	0.2252	3.4407
D05S818	0.5791	0.4209	1.3758
D07S820	0.7573	0.2427	3.1196
D08S1179	0.7934	0.2066	3.8405
D13S317	0.6380	0.3620	1.7623
D16S539	0.7104	0.2896	2.4530
D18S51	0.8260	0.1740	4.7483
D19S433	0.7873	0.2127	3.7007
D21S11	0.8262	0.1738	4.7523
FGA	0.7511	0.2489	3.0175
TH01	0.5882	0.4118	1.4285
TPOX	0.5896	0.4104	1.4365

PI 1,526,626.7947 = 99.99993450 %

Figure 6. The results of the Bayesian network propagation will appear on the tab Results. The likelihood ratio (paternity index) for each locus is then displayed on a table, at the bottom of which are shown the Combined Paternity Index (PI) and the probability of paternity, calculated from the PI using Bayes' theorem with an *a priori* probability of 0.5. The complete contents can be exported to the Windows transfer buffer using the icon on the left top corner of the table, and from there pasted onto an Excel spreadsheet for further calculations, if necessary, or to a report form.

PEDEXPERT AND THE ANALYSIS OF PATERNITY CASES INVOLVING MUTATIONS

Microsatellite mutations are uncommon events, occurring on average once in every 670 paternal meioses (Leopoldino and Pena, 2003). However, since many loci are studied in each paternity case, single inconsistencies due to mutation are often seen. Laboratories must include the possibility of paternal mutations in the calculation of the paternity index in cases in which there are one or two inconsistencies (Brenner, 2004). On the other hand, single maternal inconsistencies, which are much rarer than the paternal ones, are not a problem since the examiner can simply eliminate from analysis the locus with the inconsistency.

Needless to say, inconsistencies due to mutations are also seen in complex paternity cases, and incorporating the possibility of their occurrence in the calculations of the paternity indices is exceedingly complicated. We have developed a novel and simple method to take into account the possibility of mutations in PedExpert by simply incorporating the mutation rates into the table that establishes the conditional probability of allele transmission from the PF to the PS.

In this table, the probabilities are conditioned on two mutually excluding hypotheses: PF is the biological father and PF is not the biological father. As an example, let us assume that PF has genotype a1, a2 at a given locus, and that the obligatory paternal allele of PS is a3, i.e., there is an inconsistency at this locus. Under the hypothesis that PF is the biological father, the conditional probabilities would be 0.5 for transmission of a1 or a2 to the child and zero for transmission of a3. However, taking the probability of a mutation from a1 to a3 as u and of a mutation from a2 to a3 as v, we now see that the probability of transmission of allele a1 is (0.5-u), that the probability of transmission of allele a2 is (0.5-v), and that the probability of transmitting a3 now becomes (u + v). To implement this, we need a genetic rule to calculate the probabilities of conversion of a given allele into another by mutation.

As we have seen, the most commonly used genetic markers in paternity testing are microsatellites, which have multiple alleles characterized by the number of repetitions of a simple motif with 2-6 nucleotides (reviewed in Jeffreys and Pena, 1993; Valdes et al., 1993). It is a well-known fact that more than 90% of the mutations in microsatellites are single-step events (i.e., conversions of an allele with n repeats into alleles with n - 1 or n + 1 repeats) and that the remaining cases are two-repeat mutations (Leopoldino and Pena, 2003). Mutations involving a larger number of repeats are exceedingly rare and can be ignored for practical purposes. The model that best explains the interconversions of microsatellite alleles by mutations is the so-called stepwise mutation model, initially proposed by Kimura and Ohta (1978) and reviewed by Valdes et al. (1993).

In PedExpert, we assumed a symmetrical stepwise mutation model, which is supported by our experimental data (Leopoldino and Pena, 2003). Thus, if we call the mutation rate for the microsatellite μ, an allele with n repeats has a probability of 1 - μ of being transmitted intact, a probability of 0.45μ of being transmitted as an allele with n - 1 or n + 1 repeats, and a probability of 0.05μ of being transmitted as an allele with n - 2 or n + 2 repeats. This is the model that we incorporated in the mutation mode of PedExpert. For maximal flexibility, this mutation mode can be switched on or off by simply clicking an appropriate box in the program (see Figure 1).

DISCUSSION

PedExpert is useful, fast and user-friendly and presents a series of innovative features, among which we can mention:

1) Development of an algorithm that has the function of converting pedigrees into Bayesian networks. This algorithm uses information about family members, tested or not, and their genetic relationships with each other, to automatically create the structure of the Bayesian network, defining the number and types of nodes and their connections.

2) Development of algorithms necessary for the automatic creation of the tables of conditional probabilities associated with the nodes of the Bayesian network.

3) Development of algorithms capable of filling in the thousands of entries necessary in the several tables of conditional probabilities associated with the nodes of the Bayesian network.

4) Development of a new model that incorporates information about the possibility of

gene mutations in the tables of the conditional probabilities of transmission of alleles from the alleged father to the child, without necessity of adding new nodes to the network. This novel feature permits the choice of using the same Bayesian network in different modes, for analysis of cases taking the possibility of mutations into account or not, depending on the choice of the user.

Based on these characteristics, PedExpert greatly reduces the time of analysis of complex cases of paternity testing and eliminates most sources of logical and operational error. Thus, it emerges as a new tool of extraordinary usefulness in paternity testing.

ACKNOWLEDGMENTS

Research supported by FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

AABB (American Association of Blood Banks) (2006). Annual report summary for testing in 2006. Available at [http://www.aabb.org/Documents/Accreditation/Parentage_Testing_Accreditation_Program/rtannrpt06.pdf]. Accessed March 5, 2009.

Brenner CH (2004). Multiple mutations, covert mutations, and false exclusions in paternity casework. *Int. Congress Ser. (ICS - International Congress Series)* 1261: 112-114.

Cowell RG (2003). FINEX: a Probabilistic Expert System for forensic identification. *Forensic Sci Int.* 134: 196-206.

Dawid AP, Mortera J, Pascali VL and van Boxel D (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scand. J. Stat.* 29: 577-595.

Jeffreys AJ and Pena SDJ (1993). A Brief Introduction to Human DNA Fingerprinting. In: DNA Fingerprinting: State of the Science (Pena SDJ, Chakraborty R, Epplen JT and Jeffreys AJ, eds.). Birkhäuser Verlag, Basel, 1-20.

Kimura M and Ohta T (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc. Natl. Acad. Sci. U. S. A.* 75: 2868-2872.

Leopoldino AM and Pena SD (2003). The mutational spectrum of human autosomal tetranucleotide microsatellites. *Hum. Mutat.* 21: 71-79.

Pena SD and Chakraborty R (1994). Paternity testing in the DNA era. *Trends Genet.* 10: 204-209.

Thiele H and Nürnberg P (2005). HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics* 21: 1730-1732.

Valdes AM, Slatkin M and Freimer NB (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133: 737-749.