Problems of Forensic Sciences 2009, vol. LXXVIII, 141–159

ISSN 1230-7483

APPLICATION OF BAYESIAN NETWORKS IN FORENSIC GENETICS AND CRIMINALISTICS

Grzegorz ZADORA, Paulina WOLAŃSKA-NOWAK

Institute of Forensic Research, Krakow, Poland

Education and Culture DG
Lifelong Learning Programme
Leonardo da Vinci

Abstract

The main aim of this work is to illustrate some applications of Bayesian networks built for evaluating the value of DNA and criminalistics evidences. Some issues are presented concerning relatedness testing (paternity and identity investigation), coherent evaluation of different items of genetic evidence (obtained using additional markers – ChrY and mitochondrial polymorphism), phenotype predicting on the basis of knowledge of SNP polymorphism in some chosen genes and analysis of glass evidence. The developed networks may be easily extended when new information in a particular case is available and/or it is characterised by serious uncertainty.

Key words

Bayesian networks; Evidence evaluation; DNA profiles; Glass evaluation.

Received 16 December 2008; accepted 16 February 2009

1. Introduction

The aim of the forensic expert is to evaluate evidence, e.g. information about determined physicochemical features of analysed glass samples or DNA profiles, in the context of two hypotheses. One is related to the opinion of the prosecutor (H_1) and the other one is related to the defence opinion (H_2). The most suitable form of various evidences evaluation for forensic purposes is the likelihood ratio:

$$LR \quad \frac{\Pr(E|H_1)}{\Pr(E|H_2)},$$

which is a well-documented measure of evidence value in the forensic field [2]. The decision rules are that values of *LR* above 1 support H_1, and values of *LR* below 1 support H_2. A value of *LR* close to 1 provides little support for either proposition. Also, the larger (lower) the value of the *LR*, the stronger the support of *E* for H_1 (H_2).

There are plenty of *LR* models which could be used for evaluation of the evidence value of samples which are the subject of forensic expert analysis [2, 8]. The disadvantage of application of these models is that they require a relatively large database in order to evaluate all parameters presented in the model. This is an especially crucial point when multivariate continuous data are evaluated [1, 3]. Moreover, there is a lack of commercial software which could allow calculation of *LR* in a relatively easy way.

The Bayesian networks approach (BN) is a type of graphical model whose elements are [17]: a) nodes, which represent an uncertain state of factors (e.g. variables) and hypotheses; b) arrows between nodes, which represent links among different factors (variables). Fundamental to the idea of a graphical model is the combination of simpler parts and basic connections used in BN (Figure 1). If there is an arrow pointing

from node A to node B it is said that A is a parent of B and B is a child of A. A joint probability distribution can be ascribed to each BN. The key feature of a BN is the fact that it provides a method for breaking down a joint probability distribution of many factors (variables) into a set of local distributions of a few factors (variables) within each set. The multiplication law allows us to break down a joint probability distribution with k factors (variables) asinto a product of k-1 conditional and a marginal distribution:

$$Pr(X_1,\ldots,X_k) \quad \prod_{i\ 2}^{k} Pr(X_i|X_1,..,X_{i\ 1})\ Pr(X_1).$$

The so-called chain rule for the BN follows directly from the Markov property, i.e. a BN can be factorized as the product, for all factors (variables) in the network, of their probabilities conditional on their parents (**PA**) only, i.e.

$$Pr(X_1,\ldots,X_p) \quad \prod_{i\ 2}^{k} Pr\ X_i|\boldsymbol{PA}(X_k)\ ,$$

Therefore, the follwing probabilities could be assigned to connections presented in Figure 1:

a) serial connection (Figure 1a):
   Pr(A,B,C) = Pr(A)Pr(B|A)Pr(C|B);
b) converging connection (Figure 1b):
   Pr(C,A,B) = Pr(A)Pr(B|A)Pr(C|A);
c) diverging connection (Figure 1c):
   Pr(A,B,C) = Pr(A)Pr(B)Pr(C|A,B).

The probability of any child-node in the BN being in one state or another without current evidence is described using a conditional probability table (CPT). Conditional probabilities represent likelihoods based on prior information or past experience, *e.g.* historical information obtained from a suitable database. A node with no parents also has a probability table, but it consists only of prior probabilities. Information put into CPT and tables with prior probabilities is called *a priori* beliefs. If the true state of a particular node is known then this information could be entered into the paticular node. It is also called hard evidence, i.e. evidence that a node is 100% in one state, and 0% in all other states. A flow of information *via* BN and changes of probability values of states of appropriate nodes is a result of entering hard evidence. The results, e.g. values of posterior probabilities of each state of node H, could be used to solve an analysed problem. Sometimes soft evidence, i.e. any evidence that is not hard evidence, could be used. In other words, evidence that a node is less than 100% in one state, and/or greater than 0% in other states. Soft evidence is often used for information about which there is some uncertainty,

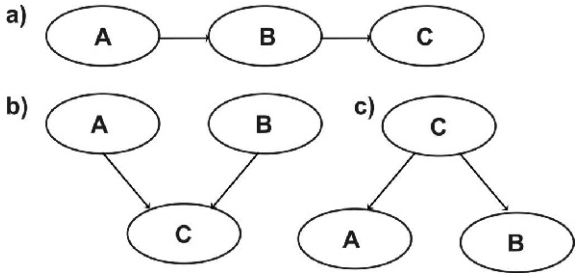

Fig. 1. Basic connections used in Bayesian networks: a) serial connection, b) converging connection, c) diverging connection.

such as from conflicting reports or an unreliable source.

There are various programmes (HUGIN [14], GeNIe [12]) which allow analysis of BN models and they do not require any special knowledge of programming. One of the disadvantages of application of the BN approach in the forensic sphere is that information obtained in the discrete type node responsible for evaluation of a hypothesis, after entering hard evidence provides probability in the form of $Pr(H_1|E)$ and $Pr(H_2|E)$. It was mentioned that information which a forensic expert could provide is $Pr(E|H_1)$ and $Pr(E|H_2)$, *e.g.* in the form of *LR*. A likelihood ratio could be also obtained by BN, when prior probabilities inserted into node H are taken into account, because it could be shown according to Bayes theory that:

$$\frac{Pr\ H_1|E}{Pr\ H_2|E} \quad \frac{Pr\ E|H_1}{Pr\ E|H_2} \quad \frac{Pr\ H_1}{Pr\ H_2} \quad LR \quad \frac{Pr\ H_1}{Pr\ H_2}.$$

For example, when $Pr(H_1) = Pr(H_2) = 0.5$ then *LR* value is equal to the posterior odds $\Pr\ H_1|E\ \big/ \Pr\ H_2|E$. Of course, this kind of calculation makes sense when two hypotheses are considered.

1.1. Physicochemical data evaluation

The application of the BN approach in criminalistics will be presented in the case of glass traces analysis. Glass is a material commonly used in many areas of human activity, therefore glass fragments are frequently present at the scene of events such as car accidents, burglaries, fights. This is why the importance of glass as evidence was recognised many years ago [9, 22]. There are various aims of analysis of small glass fragments for forensic purposes. One such aim is to answer the following question – could two glass samples have originated from the same object? – when comparing two glass samples. This is so-called "source level" analysis [1, 3, 22]. Another aim of glass analysis

of evidences for forensic purposes is related to the so-called "activity level", e.g. to answer the following question – could the suspect have broken this glass object? A solution to the second problem requires both objective data (e.g. the number of glass fragments which were recovered from the suspect's clothes and which revealed $LR > 1$ or $LR < 1$ – after comparative analysis) and the expert's subjective knowledge, which allows presentation of relations between objective data. The issue of the evaluation of the evidence value of glass fragments on the so-called "activity level" could also be solved by application of the *LR* approach, then H_1 – the suspect was the person who broke the window; H_2 – another person was the person who broke the window. A well-known model based on univariate data is described in [9]. It required knowledge about the number of glass fragments which were recovered from the suspect's clothes and how many of them revealed $LR > 1$ or $LR < 1$ when they were compared to the control sample. The process of grouping of glass fragments is an integral part of this approach. The general scheme of the grouping process is: 1) rank glass samples in ascending order according to their mean *RI* values; 2) calculate distances between them and find distances which are greater than the critical value of distance, 3) calculate a likelihood ratio on the "source level" (appropriate equations are in [9]). The grouping procedure described above can only be applied to cases where glass fragments are described by one variable. It is not suitable for the evaluation of glass fragments described by several features (e.g. elemental composition) as it is not possible to rank samples described by a vector of mean values. Therefore, a model based on Bayesian networks has been proposed.

1.2. DNA profiles evaluation

Identification of human remains, analysing criminal samples and paternity testing are currently conducted using DNA profiles comprising several genetic markers, such as short tandem repeats (also called STRs) or mitochondrial polymorphic regions. Sometimes there is a need to involve polymorphic STRs dispersed on the Y or X chromosome to improve the results of forensic genetic analysis [11]. Considering the importance of DNA evaluation to the legal system, further research into using Bayesian networks seems to be essential. They reduce the amount of confusion that can occur, by presenting important relationships between evidences in a logical way. Once a network has been created, it can be used again and again in similar cases [13].

The computational technology of Bayesian network may be implemented in the evaluation of a likelihood ratio, which can pass from an initial pedigree representation of a forensic identification problem to an appropriate graphical representation. Complex genetic pedigree fits particularly well into the Bayesian network model. Nuclear family relationships constitute natural modular building blocks of the representation: parents and children have become part of the general terminology of the Bayesian network – for non-genetic applications as well. A graph associated with a genetic pedigree is a set of nodes representing the variables (genotypes and alleles) and a set of directed arrows representing the links between these variables. The simple relationship graph is a standard representation of the transmission of genes from parents to offspring. The overall probability structure is completely determined by specifying the conditional probability tables for each variable given its "parents". Mendelian laws of inheritance and logical relationships between genes and genotypes give the conditional probability tables (CPT). After entering the case evidence at the relevant genotype nodes, one can "propagate" it throughout the network. The required likelihood ratio, based on the data for one marker, is obtained from the posterior distribution at node "query". Dawid et al. proposed basic network fragments focusing on individual genes and genotypes [10]. Here the node *gt,* representing a genotype, is modelled as a logical combination of the alleles inherited from the mother and

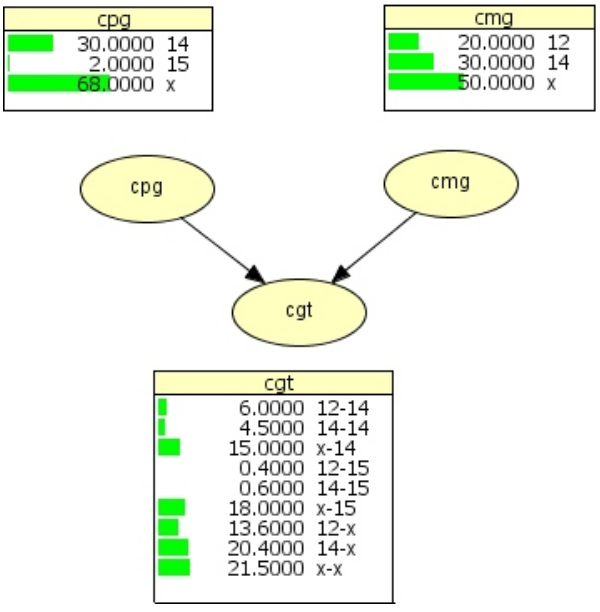

Fig. 2. Basic model useable for inferring a child's paternal and maternal genes (*cpg* and *cmg*), based on the child's genotype (*cgt*).

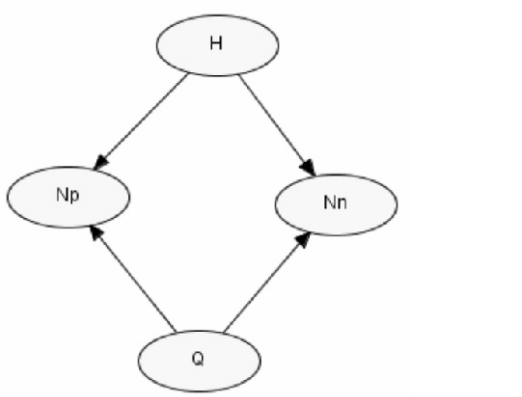

Fig. 3. The proposed Bayesian networks model with examples of conditional probability tables. Node *H*: H_p – the suspect was the person who broke the window; H_d – another person was the person who broke the window; node N_p: the number of glass fragments which revealed $LR > 1$; node N_n: the number of glass fragments which revealed $LR < 1$; node *Q*: the applied *LR* model for comparison problem gives correct or incorrect answers.

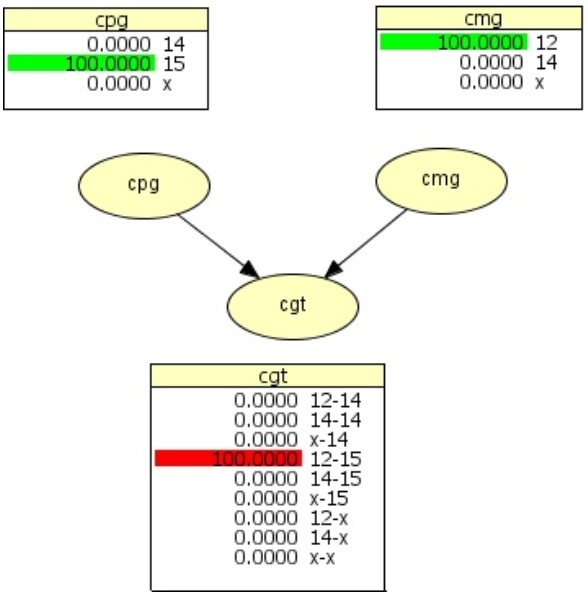

Fig. 4. Basic model after entering information about child genotype (12–15). It is certain that the child inherited allele 12 from the mother and allele 15 from his/her biological father.

father, respectively. These parentally inherited genes are represented by the nodes *mg* and *pg*, which means: "maternal gene" and "paternal gene" (Figure 2). For simplicity, the definition of evidence is now restricted to the typing results of a single locus or marker. A particular application is relatedness testing as discussed by Dawid et al. [10]. These authors have shown how appropriate graphical structures for Bayesian networks can be derived from initial pedigree representations of forensic identification problems (Figure 3 and 4).

Evaluating the probability that the alleged father is the true father requires, for every possible father, the ratio of the likelihood of the observed DNA profiles (if he was the true father), to the likelihood of the above mentioned profiles (if the alleged father was not the true father). These likelihood ratios (paternity indexes), depend on the amount of shared ancestry among the mother, the alleged father and the alternative possible father [18] (Figure 5).

The advantage of the logical approach is that the likelihood ratio can be put in the context of an entire case. However, in real casework it is difficult to simplify the problem to two hypotheses aligned with the prosecution and defence. In identification cases or incestuous paternity investigations the judge should consider more than two hypotheses as to the real perpetrator. Supposing that the suspect has some relatives who are related to a lesser or greater degree, then the number of defence hypotheses increases. This is provided by the general form of Bayes's theorem. Hence, the defence hypothesis can be divided into sub propositions, which are mutually exclusive and exhaustive.

In paternity cases, there could be, for example:

- H_1: the alleged man is the father of the child;
- H_2: the brother of the alleged man is the biological father of the child;
- H_3: the cousin of the alleged man is the biological father of the child;
- H_4: the biological father of the child is unknown and unrelated to the alleged man.

Following Champod, they are called sub propositions [17]. Such an approach requires some assumptions as to the prior knowledge about possible paternity or identity of chosen person. In a routine criminal paternity investigation (for example incestuous paternity) or in cases of human remains identification, it is common practice to use a panel of 16 autosomal markers. However, in some identification cases, when there are no or few reference materials from potentially close relatives of unknown body remains, the obtained likelihood ratios are too low to assess the posterior probability of identity as proved. So there is a need to broaden the range of routinely used markers. Mitochondrial, Y-chromosome or X-chromosome markers can be useful, dependent on the case. As a result, several different pieces of evidence could be put together, providing a single combined likelihood ratio. An autosomal DNA estimate is formed by the product rule, but mtDNA or ChrY-STR and ChrX-STR estimates are the frequencies in the appropriate databases, so multi-

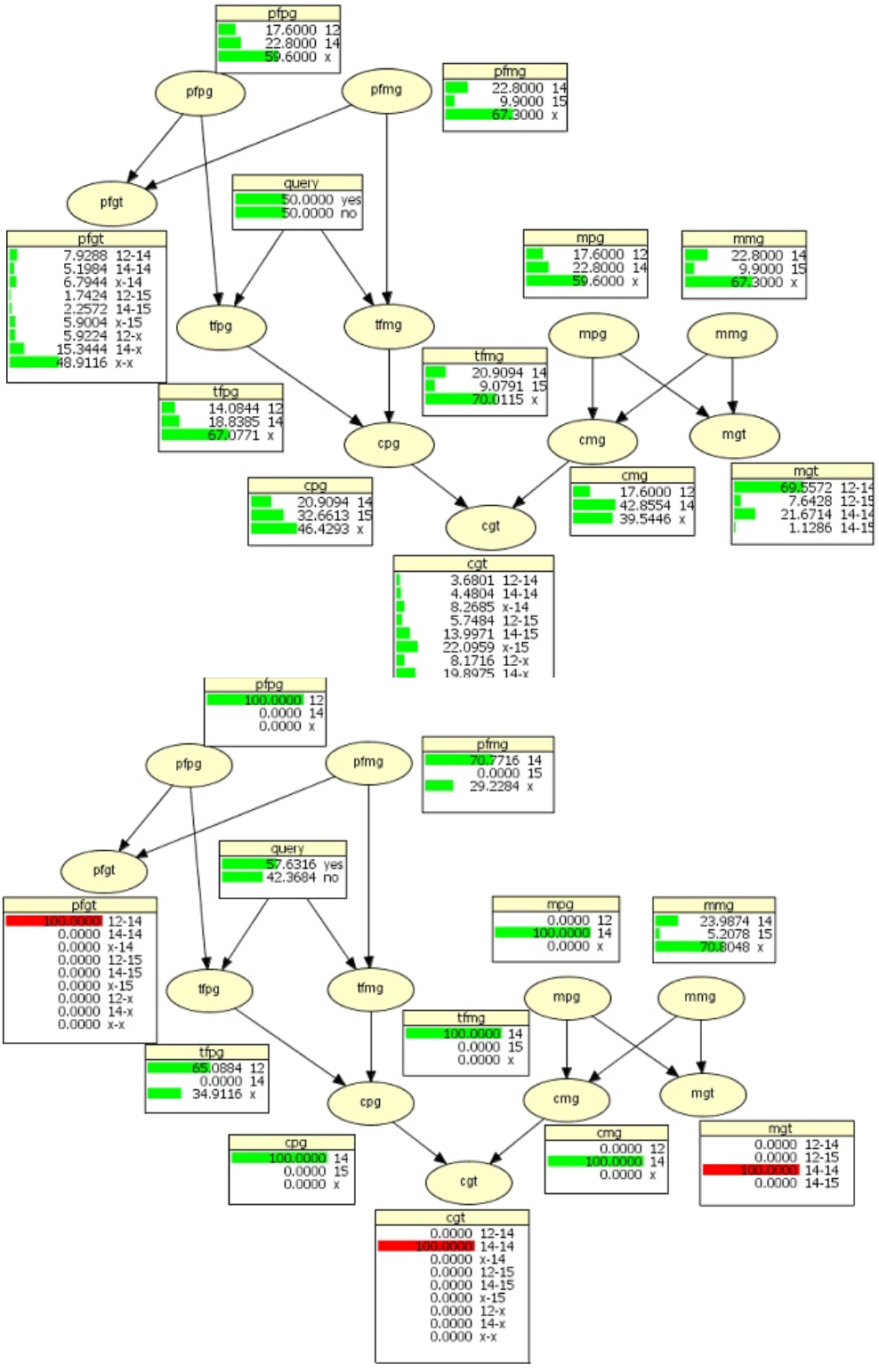

Fig. 5. Bayesian network for evaluating paternity index on the basis of knowledge of mother, child and putative father genotypes: A – before entering any information, B – after entering information about genotypes of involved individuals in a single marker. *f* denotes father, *p* – paternal, *m* – mother (if in the first place) or maternal (if in second place), *c* – child, *gt* – genotype. For example *fmg* means father maternal gene, *cgt* – child's genotype.

plying them may provide the first estimate of the joint probability. One of the problems with this approach is that the information inherent in the Y chromosome haplotype and mitotype contains strong information as to the population origin of the involved persons. So an approach was proposed that proceeds by adjusting the coancestry coefficients value upwards to account for the matches already observed [8].

Another application of Bayesian networks in forensic genetics is the possibility of prediction of human pigmentation based on genetic data. Detailed knowledge concerning gene variants and their associations with physical traits is necessary to allow phenotype prediction that could be applied to forensic genetics. Pigmentation is a complex physical trait with multiple genes involved. The newest data indicate that the highest association for blue – non blue eye colour appeared for a single SNP in intron 86 of *HERC2* – rs12913832. The other examined variant of *OCA2* rs1800407 is involved in green iris colour determination [4].

Forensic biological traces collected at crime scenes are normally subjected to human identification, which is based on analysis of highly variable STR systems. Obtained crime scene DNA profiles are compared with reference profiles and an identified match is understood as strong evidence supporting the hypothesis of common origin of the analysed samples. If no suspect is involved in a case, obtained crime scene profiles are usually loaded into a national DNA database. However, if no match is found, the investigation may significantly slow down. The crime scene samples may, however, also be used for intelligence purposes. Genetic prediction of physical traits is another emerging area for forensic science, and doubtlessly this may become very useful at the investigation stage.

To graphically represent the relationships between the observed data we developed a Bayesian network [15] enabling inferring as to the probability of individual hair, skin and iris colour on the basis of a known person's genotype in the range of *MC1R*, *HERC*2 rs12913832 and *OCA2* rs1800407. Such a network can be used to follow how a change in the level of certainty in one variable (e.g. the kind of *MC1R* or *HERC2* genotype) may change the level of certainty for other variables (e.g. probability of a given hair colour of an unknown donor of a crime sample).

The main aim of this paper concerning DNA evidence interpretation was to illustrate some applications of Bayesian networks built for evaluating the value of DNA evidence. Some issues are presented concerning relatedness testing (paternity and identity investigation), coherent evaluation of different items of genetic evidence and even phenotype predicting on the basis of genetic data.

2. Methods

2.1. Bayesian Network – a model for glass analysis on the activity level

The model presented in Figure 3 was used in the evaluation of glass evidence on the activity level, i.e. could the suspect have broken this glass object? Analysis of the proposed model was performed with Hugin Researcher v. 6.9 software [14]. In the prepared model each glass sample is separately compared to the control sample by a suitable *LR* model, i.e. analysis on the source-level. The number of glass fragments which reveal $LR > 1$ (N_p) and $LR < 1$ (N_n), when they are compared to the control sample, are obtained as a final result of the source-level evaluation [1, 2]. Node N_p had four states: 0 – non glass fragments recovered from suspect's clothes revealed $LR > 1$ on the source level, 1, 2, 3 – one, two, three and more than three glass fragments recovered from suspect's clothes revealed $LR > 1$ on the source level. Node N_n had three states: 0 – non glass fragments recovered from suspect's clothes revealed $LR < 1$ on the source level, 1, 2 – one, two or more than two glass fragments recovered from suspect's clothes revealed $LR < 1$ on the source level. The appropriate conditional probabilities are presented in Tables I–II. Node *H* has two states: H_1 – a suspect broke the glass object, H_2 – another person broke the glass object. It was assumed $Pr(H_1) = Pr(H_2) = 0.5$. Node *Q* also has two states: *Q* – the applied *LR* model for evaluation of physicochemical data determined for glass samples gave correct answers, $\overline{Q}$ – the applied *LR* model for evaluation of physicochemical data determined for glass samples did not give correct answers, i.e. provided misleading evidence. Taking into account the authors knowledge, $Pr(Q)$ and $Pr(\overline{Q})$ could be assumed as equal to 0.9 and 0.1.

TABLE I. CONDITIONAL PROBABILITIES FOR NODE N_p

States of N_p node	$Pr(Q)$		$Pr(\overline{Q})$	
	$Pr(H_1)$	$Pr(H_2)$	$Pr(H_1)$	$Pr(H_2)$
3	0.2	0.001	0.001	0.0001
2	0.5	0.004	0.004	0.0004
1	0.2	0.045	0.045	0.0055
0	0.1	0.95	0.95	0.99

TABLE II. CONDITIONAL PROBABILITIES FOR NODE N_n

States of N_n node	$Pr(Q)$		$Pr(\overline{Q})$	
	$Pr(H_1)$	$Pr(H_2)$	$Pr(H_1)$	$Pr(H_2)$
2	0.01	0.05	0.001	0.0001
1	0.04	0.20	0.004	0.0004
0	0.95	0.75	0.995	0.9995

For example, if we assume that a suspect broke the glass object, H_1 is true, then it is very probable to find on the suspect's clothes 2 glass fragments $Pr(N_p \quad 2|H_1,Q)$, which revealed $LR > 1$ when analysis on the source level was carried out and when the applied LR model provided correct answers. It was assumed, taking into account the authors' knowledge, that all the aforementioned could happen in 50% of analysed cases, i.e. $Pr(N_p \quad 2|H_1,Q)$ 0,5. In the same situation, $LR < 1$ for two analysed samples seems to be a very rare event. Thus, it was assumed that $Pr(N_p \quad 2|H_1,Q)$ 0.005.

When H_2 is correct then $Pr(N_p \quad 2|H_1,Q)$ is a conditional probability that N_p glass fragments revealed similarity to the control sample by chance, i.e. they originate from glass object(s) other than the glass object broken during the crime. Moreover, $Pr(N_p \quad 2|H_1,Q)$ is also related to the situation where recovered glass fragments were present on the suspect's clothes by chance. The authors' knowledge suggests that these probabilities are higher when Q is true than when Q is false.

2.2. Bayesian network – models for DNA analysis

Bayesian networks for evaluation of combined likelihood ratio based on different kinds of genetic evidences were performed with Hugin Researcher v. 6.9 software [14]. Conditional probability tables assigned to the particular nodes required some probabilistic assumptions. According to knowledge from our practice, the mean values were set to certain parameters. In our routine casework involving identity and paternity investigation we use AmpF*l* Identifiler, AmpF*l* Yfiler, AmpF*l* MiniFiler STR kits and Big Dye Terminator Cycle Sequencing Ready Reaction Kit (Applera). Calculation of paternity index was performed according to Dawid, Mortera, Pascali, and Van Boxel [10]. Each individual's genotype (genotype nodes, such as putative father – *pfgt*) was desegregated into its constituent, unobserved, paternally and maternally inherited genes (allele nodes – such as *pfpg* – putative father paternal gene). The hypothesis node "query" embodies H_1 when it takes the value "true" and H_2 when "false". We assumed both the hypotheses as equally probable, so that after propagation of evidence, the ratio of their posterior probability yields the paternity ratio based on a given marker. Prior information about allele frequencies in the South Poland population was taken from a previous study [19].

The network for predicting physical traits consisted of three gene nodes: "*MC1R*", "*OCA2*", "*HERC2*" and three phenotype ones: "hair_colour", "skin_colour" and "eye_colour". For example, it was assumed for the sake of simplicity that hair colour is

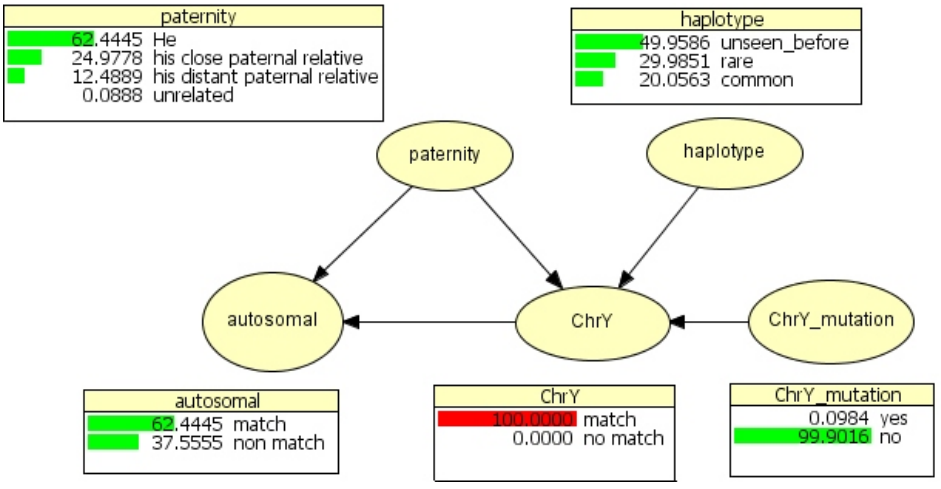

Fig. 6. The proposed Bayesian network for evaluating hypotheses as to alleged father paternity using two kinds of DNA markers: autosomal STR and ChrY ones. On the left: the results of entering "hard evidence" (that is the observed no exclusion in autosomal markers) and propagating it throughout the network. It should be remarked that the probability of the match in Chr Y markers simultaneously grew to nearly 100%.

TABLE III. RESULTS OF SENSITIVITY ANALYSIS – INFLUENCE OF CHANGES MADE IN THE NODE N_p ON VALUES OF POSTERIOR ODDS $\frac{Pr(H_1|N_p,N_n)}{Pr(H_2|N_p,N_n)}$. IT WAS ASSUMED THAT: $N_n = 0$, $Pr(Q) = 0.9$, $Pr(Q) = 0.1$, $Pr(H_1) = Pr(H_2) = 0.5$

States of N_p node	Time			
	2 h		8 h	
	$Pr(N_p\|H_1,Q)$	$\frac{Pr(H_1\|N_p,N_n)}{Pr(H_2\|N_p,N_n)}$ *	$Pr(N_p\|H_1,Q)$	$\frac{Pr(H_1\|N_p,N_n)}{Pr(H_2\|N_p,N_n)}$
3	0.2	12.6	0.01	249.8
2	0.5	12.6	0.04	156.2
1	0.2	5.7	0.15	5.7
0	0.1	0.2	0.80	0.2

TABLE IV. RESULTS OF SENSITIVITY ANALYSIS – INFLUENCE OF CHANGES MADE IN THE NODE N_p ON VALUES OF POSTERIOR ODDS $\frac{Pr(H_1|N_p,N_n)}{Pr(H_2|N_p,N_n)}$. IT WAS ASSUMED THAT: $N_p = 0$, $N_n = 0$, $Pr(H_1) = Pr(H_2) = 0.5$

Q	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
$\overline{Q}$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\frac{Pr(H_1\|N_p,N_n)}{Pr(H_2\|N_p,N_n)}$ *	158	156	155	150	146	140	133	123	107	77	10

TABLE V. CONDITIONAL PROBABILITY TABLE FOR THE NODE "AUTOSOMAL" OF THE NETWORK PRESENTED IN FIG. 7

Chr Y	Match				No match			
mt DNA	Match		No match		Match		No match	
identity	He	His maternal relative	His paternal relative	Unrelated	He	His maternal relative	His paternal relative	Unrelated
Match	1	1	1	$1.0\ 10^{-9}$	0	0	$1.0\ 10^{-6}$	0
No match	0	0	0	1	1	1	0.99999	1

determined by two factors: the *MC1R* genotype and the *HERC2* rs12913832. The overall probability structure was completely determined by specifying the conditional probability tables for each node, based on the obtained results of genetic data. Population stratification (proportions of individuals with different hair and different kinds of iris and skin in the analysed sample of 388 individuals from South Poland) was used as a priori information. Conditional probability tables were constructed according to Branicki et al. [5, 6]. For example, the *HERC2* node represents all probabilities for *Pr*(*HERC2* | hair, skin, eye colour). After setting hard evidence (known SNP individual variants) on any of the gene node states, e.g. the crime sample has *MC1R* genotype – 160/160 and propagating it through the network, one can ascertain the probabilities of particular hair colours of the sample donor. The names of states and prior probabilities of the particular states for each variable are shown on monitor windows accompanying each node.

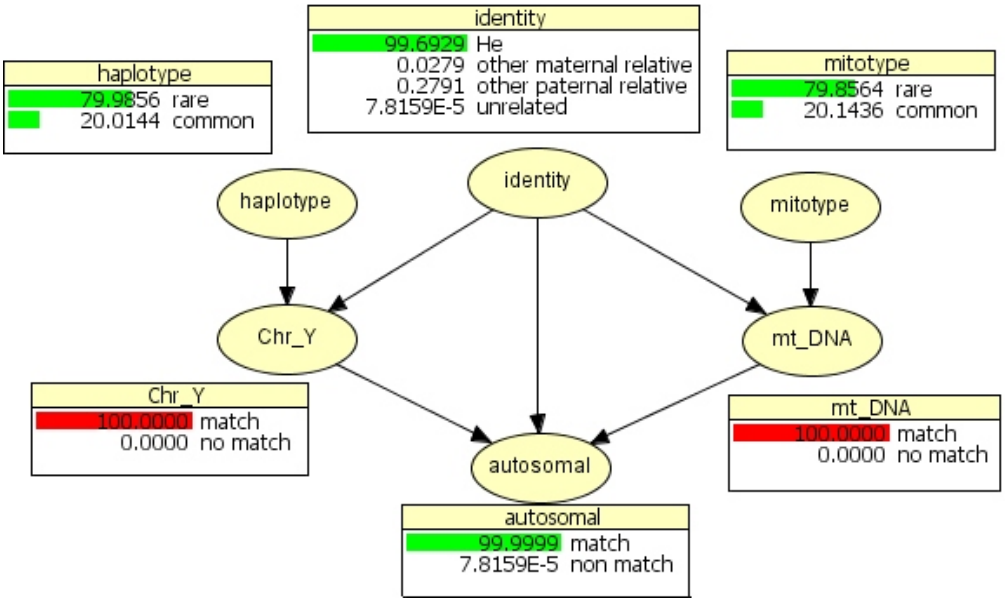

Fig. 7. Bayesian network for evaluating a hypothesis about identity of human remains after propagation of the "hard evidences" on the "mt DNA" and "ChrY" nodes, which means that after examination of some relatives using only those DNA markers there is a high probability of a match. One can see that simultaneously the probability of a match in autosomal markers is increased [21].

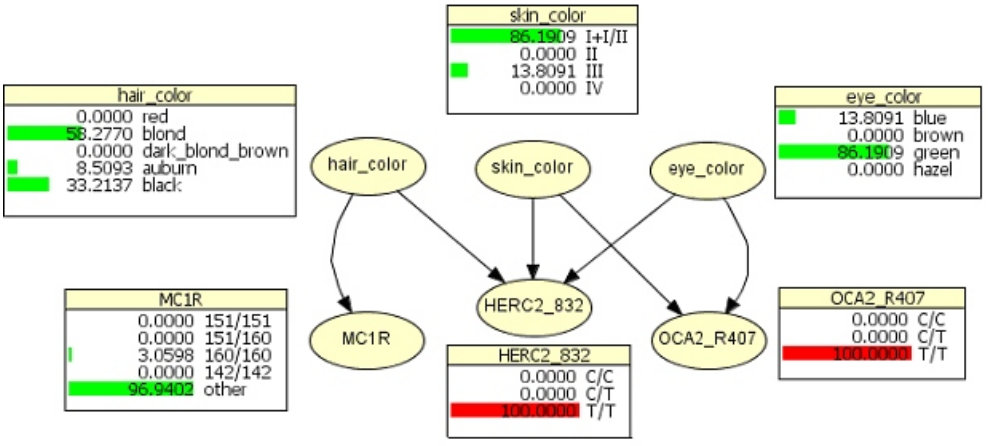

Fig. 8. Bayesian network after incorporating information: DNA sample is T/T in *HERC2* gene and T/T in *OCA2*.

3. Results and discussion

3.1. Bayesian network – a model for glass analysis on the activity level

Sensitivity analysis of the proposed model was performed. Two factors were the subject of analysis. The first one was related to the influence of various opinions about the quality of the applied *LR* model for analysis on the source level, i.e., The results of such analysis are presented in Table III. It could easily be observed that an increase in strength of opinion about the correctness of the applied *LR* model is supported by increasing of:

$$\frac{Pr(H_1 \mid N_p, N_n)}{Pr(H_2 \mid N_p, N_n)},$$

which is equal to *LR* as it was assumed that $Pr(H_1) = Pr(H_2) = 0.5$.

The second analysed problem was related to changes in $Pr(H_1|N_p, N_n)$ and $Pr(H_2|N_p, N_n)$ which could be observed when various sets of $Pr(N_p|H_1, Q)$ were used. The various proposed sets of $Pr(N_p|H_1, Q)$ could be related to changes in other factors that are important for glass fragments evaluation, e.g. passage of time between the incident and when the suspect was arrested. In general, the number of glass fragments recovered from suspect clothes decreases with passage of time. The obtained results of:

$$\frac{Pr(H_1|N_p, N_n)}{Pr(H_2|N_p, N_n)} \quad LR$$

(Table IV) confirmed that the proposed BN model worked correctly. When the number of recovered glass samples which revealed $LR > 1$ relative to the control sample increases, then *LR* also increases. The influence of time is also correctly modelled by the BN model, as *LR* values were higher when calculated for "8 h" than for "2 h". This makes sense as recovery of three or more glass fragments after eight hours is a rarer event than after two hours. It is a well-known rule that the rarer the situation, the higher the evidence value, e.g. expressed by *LR* value.

3.2. Bayesian network – models for DNA analysis

3.2.1. Paternity investigation model

Hypothesis node "paternity" has four states: "he", "his close paternal relative", "his distant relative" and "unrelated" in relation to the alleged man that could be the biological father. The Y chromosomal match node is influenced by the "haplotype" node representing the kind of haplotype ("unseen before", "rare", "common" in the population). It is also influenced by information about the possible mutation process during spermatozoa production. Probability of autosomal evidence (no exclusion of paternity in 16 STR markers) given the chosen hypotheses was assumed to be non-independent of the knowledge of the results of Chr Y typing, hence there is an arrow (influence) between the "Chr Y" node and the "autosomal" node. The overall network structure is illustrated by Figure 6.

3.2.2. Identity investigation model

Hypothesis node "identity" has four uniformly probable states: "he", "his maternal relative", "his paternal relative" and "unrelated". Autosomal match probabilities (Table V) depend on the kind of accessible reference material; the closer the possible relatives, the stronger the impact on the truth of the hypothesis that the remains originate from "he". The proposed structure of the network is shown in Figure 7.

Interest in the probabilistic evaluation of DNA evidence has grown considerably during the last decade, particularly in respect of relatedness testing, including complicated incest paternity disputes and identity cases [17, 20]. Bayesian networks can be built to represent the above-mentioned problems. Such networks allow one to obtain the correct likelihood ratio for the chosen hypotheses based on all available evidences. The described Bayesian networks seem to be usable in addressing a wide range of different scenarios simply by changing the assumed conditional probabilities assigned to the particular nodes to what is appropriate to the specified case. They account for different prosecution and partitioned defence hypotheses, including maternal and paternal relatives as a possible source of DNA material and occurrence of a mutation event. The overall probability structure is completely determined by specifying the conditional probability tables for each variable given its "parents". The proposed Bayesian networks are rather simple and can be extended in various ways depending on the information as to the circumstances of the real casework. Our preliminary experiences with implementation of Bayesian networks in evaluation of the strength of DNA evidence resulting from the analysed different DNA markers seem to provide reliable combined likelihood ratios given the chosen hypotheses and could be an additional and efficient tool in forensic practice.

Correct evaluation of genotyping results is a fundamental issue in forensic examinations. Homozygous *MC1R* genotypes for so-called R polymorphisms significantly support the hypothesis that the sample donor has red hair colour and fair skin. Thus, an important aim of the present study was to propose an inferring model based on the results of the performed genetic analysis in the Bayesian network (Figure 8).

The proposed model can provide scientific justification for the results or conclusions of *MC1R*, *HERC2* (rs12913832) and *OCA2* (rs1800407) genotypes examinations. In addition, this model can also be used by other forensic practitioners, who can add information on the influence of other genes on visible traits.

Evidence in the case of a variable is a statement of the certainties of its states. If the variable is instantiated, it is called "hard" evidence; otherwise it is called "soft" evidence. A detailed study performed recently by Branicki et al [4] on *OCA2* rs1800407 revealed that the variant T/T of this SNP has significant correlation with green eye colour. Hence, if one obtains a DNA sample with this gene variant (Figure 8)

and e.g. T/T variant of *HERC2*, the operational advice to police may be: "the perpetrator should have blond hair colour with about 58%, fair skin with 86% and green eye colour with about 86% certainty". When, for example, we obtain a sample with *MC1R* genotype 151/160, then the operational advice to police is fairly certain: "the perpetrator should have dark red or red hair".

4. Conclusion

The developed networks may be easily extended when new information in a particular case is available and/or it is characterised by great uncertainty [7]. The last factor is especially important when data inserted in a BN were estimated only on the basis of expert knowledge, which is characterised by relatively large subjectivity. If we have additional knowledge about, for example, other genes influencing human phenotype features, it will be utilised for further updating our present knowledge on relevant factors and can provide scientific justification for the conclusions or findings of genotypes examinations.

However, there are some limitations to applying the Bayesian network model despite its advantages. Such a model involves an individual's subjective judgement in determining causal relationships and preliminary assumptions. For example, the prior distribution of the *MC1R* genotype and different hair colour were assessed on the basis of examining 388 individuals; on the other hand, we are aware that this sample may not be representative of the whole population. Other serious problems concerning combining different pieces of genetic evidence are emerging from possible dependences between distribution of Y-chromosome haplotypes, mitochondrial mitotypes and autosomal markers in relevant populations. Hence, further research and analyses are needed to enhance the inference capability and accuracy of the Bayesian network model before it can be applied in routine forensic casework.

Acknowledgments
The research was financially supported by Leonardo da Vinci – "Uncertainty in forensic evidence evaluation".

References

1. Aitken C. G. G., Lucy D., Zadora G. [et al.], Evaluation of transfer evidence for three level multi-variate data with the use of graphical models, *Computational Statistics & Data Analysis* 2006, 50, 2571–2788.
2. Aitken C. G. G., Taroni F., Statistics and the evaluation of evidence for forensic scientists. Statistics in Practice, John Wiley & Sons, Chichester 2004.
3. Aitken C. G. G., Zadora G., Lucy D. A., Two level model for evidence evaluation, *Journal of Forensic Science* 2007, 52, 412–419.
4. Branicki W., Interaction between *HERC2*, *OCA2* and *MC1R* may influence human pigmentation phenotype [manuscript in preparation].
5. Branicki W., Brudnik U., Kupiec T. [et al.], Association of polymorphic sites in the *OCA2* gene with eye colour using the tree scanning method, *Annals of Human Genetics* 2008, 72, 184–192.
6. Branicki W., Brudnik U., Kupiec T., [et al.], Determination of phenotype associated SNPs in the *MC1R* gene, *Journal of Forensic Science* 2007, 52, 349–354.
7. Branicki W., Wolańska-Nowak P., Brudnik U. [et al.], Forensic application of a rapid test for red hair colour prediction and sex determination, *Problems of Forensic Sciences* 2007, 69, 37–51.
8. Buckletown J., Triggs C. M., Walsh S. J., Forensic DNA evidence interpretation, CRC Press, Boca Ratton 2005.
9. Curran J. M, Hicks T. N., Buckleton J. S., Forensic interpretation of glass evidence, CRC Press, Boca Raton 2000.
10. Dawid A. P., Mortera J., Pascali V. L. [et al.], Probabilistic expert systems for forensic inference from genetic markers, *Scandinavian Journal of Statistics* 2002, 29, 577–595.
11. Evett W., Weir B. S., Interpreting DNA evidence, Statistical genetics for forensic scientists, Sinauer Associates Inc., Sunderland 1998.
12. GeNIe http://genie.sis.pitt.edu.
13. Hepler A.B., Dawid P., Leucari V., Object-oriented graphical representations of complex patterns of evidence, Law, *Probability and Risk* 2007, 6, 275–293.
14. Hugin, www.hugin.com.
15. Jensen F. V., Bayesian networks and decision graphs, Springer-Verlag New York, 2001.
16. Steffen L., Lauritzen S. L. Sheehan N. A., Graphical models for genetic analysis, *Statistical Science* 2003, 18, 489–514.
17. Taroni F., Aitken C. G. G., Garbolino P. [et al.], Bayesian Networks and probabilistic inference in forensic science, John Wiley & Sons, Chichester 2006.
18. Wolańska-Nowak P., Application of subpopulation theory to evaluation of DNA evidence, *Forensic Science International* 2000, 113, 63–69.
19. Wolańska-Nowak P., Branicki W., Kupiec T., STR data for SGM Plus and penta E and penta D loci in a population sample from south Poland, *Forensic Science International* 2002, 127, 237–239.
20. Wolańska-Nowak P., Branicki W., Parys-Proszek A., Usefulness of bayesian networks – a forensic genetics per-

spective [in:] Chemometrics – methods and aplications, Institute of Forensic Research Publishers, Kraków 2006.
21. Wolańska-Nowak P., Branicki W., Parys-Proszek A., Examples of combining genetic evidence – Bayesian network approach, *Forensic Science International: Genetics Supplement Series 1* 2008, 669–670.
22. Zadora G., The role of statistical methods in assessing the evidential value of physico-chemical data, *Problems of Forensic Sciences* 2006, 55, 91–103.

Corresponding author
Grzegorz Zadora
Instytut Ekspertyz Sądowych
ul. Westerplatte 9
PL 31-033 Kraków
e-mail: gzadora@ies.krakow.pl

ZASTOSOWANIE MODELI SIECI BAYESOWSKICH W GENETYCE SĄDOWEJ I KRYMINALISTYCE

1. Wstęp

Rolą biegłego sądowego jest ocena, czy dostarczony do analizy dowód (E), np. dane fizykochemiczne opisujące okruchy szkła lub oznaczony profil DNA, bardziej wspiera hipotezę prokuratora (H_1) czy też hipotezę obrony (H_2). Z punktu widzenia wymiaru sprawiedliwości najlepszą metodą oceny wartości dowodu jest zastosowanie ilorazu wiarygodności (*LR*) [2], tj.:

$$LR \quad \frac{Pr(E|H_1)}{Pr(E|H_2)}.$$

Iloraz wiarygodności interpretuje się następująco: gdy jego wartość jest większa od 1, to dowód E wspiera hipotezę H_1, a gdy jego wartość jest mniejsza od 1, to wspiera hipotezę H_2. W interpretacji stosuje się też prostą zależność, że im większa jest od 1 (mniejsza od 1) wartość *LR*, tym mocniejsze jest wsparcie dla hipotezy H_1 (H_2).

Istnieją różne modele *LR*, które mogą być użyte do oceny wartości dowodowej wyników badań różnego rodzaju materiału dowodowego [2, 8]. Jedną z niedogodności zastosowania tych modeli jest to, że wymagają one, aby dostępna była stosunkowo duża liczba danych o analizowanych obiektach, szczególnie w przypadku analizy danych wielowymiarowych, w których występują korelacje pomiędzy zmiennymi. Odpowiednia liczba danych wymagana jest w celu wykonania rzetelnej estymacji parametrów (średnie, wariancje, kowariancje) występujących w stosowanych modelach *LR* [1, 3]. Ponadto brak jest komercyjnych programów komputerowych umożliwiających obliczanie *LR*, szczególnie przy analizie danych fizykochemicznych.

Sieć bayesowska (BN) jest zbiorem węzłów reprezentujących zmienne, które są połączone za pomocą strzałek [17]. Strzałki te ilustrują zależności pomiędzy zmiennymi. Węzły mogą być reprezentowane przez dane dyskretne lub ciągłe. Fundamentalną ideą metod graficznych jest ich konstruowanie z prostszych części. Złożone sieci baysowskie składają się z podstawowych struktur zaprezentowanych na rycinie 1. Jeżeli w węzeł B wchodzi strzałka od węzła A, to wówczas węzeł B zwany jest węzłem-dzieckiem, a węzeł A zwany jest węzłem-rodzicem. Użyteczną cechą modeli graficznych jest to, że umożliwiają one rozłożenie rozkładu łącznego prawdopodobieństwa *k* zmiennych na zbiór lokalnych rozkładów o mniejszej wymiarowości, a konkretnie podział łącznego rozkładu prawdopodobieństwa dla *k* zmiennych na iloczyn *k*-1 rozkładów warunkowych i brzegowych

$$Pr(X_1,\ldots,X_k) \quad \prod_{i\ 2}^{k} Pr(X_i|X_1,..,X_{i\ 1})\ Pr(X_1).$$

Z własności Markowa można wyprowadzić, że w przypadku BN opartych na $A_1,.., A_p$ węzłach, rozkład łączny prawdopodobieństwa $Pr(A_1,.., A_p)$ może być wyrażony jako iloczyn

$$Pr(X_1,\ldots,X_p) \quad \prod_{i\ 2}^{k} Pr\ X_i|\mathbf{PA}(X_k)\ ,$$

gdzie (**PA**) oznacza węzły źródłowe węzła X_i. Tym samym, można zapisać, że dla:

a) połączenia szeregowego (rycina 1a):
   Pr(A,B,C) = Pr(A)Pr(B|A)Pr(C|B);
b) połączenia rozproszonego (rycina 1b):
   Pr(C,A,B) = Pr(A)Pr(B|A)Pr(C|A);
c) połączenia skupiającego (rycina 1c):
   Pr(A,B,C) = Pr(A)Pr(B)Pr(C|A,B).

Prawdopodobieństwo, że dany węzeł-dziecko w analizowanej BN znajduje się w danym stanie, jest opisany prawdopodobieństwem warunkowym umieszczonym w tzw. tablicy prawdopodobieństw warunkowych (z ang. conditional probabilisty tables; *CPT*). Ocena wartości tego prawdopodobieństwa warunkowego oparta jest na informacji pochodzącej z doświadczenia (tzw. wiedza historyczna) osoby tworzącej taką sieć i która często zebrana jest w różnego rodzaju bazach danych. Węzeł źródłowy również posiada *CPT*, ale do opisu stanów takiego węzła potrzebna jest informacja o prawdopodobieństwie bezwarunkowym.

Jeżeli znany jest aktualny stan, w jakim występuje konkretny węzeł, to taka informacja może być wprowadzona do tego węzła. Jest to tzw. twardy dowód, tj. dowód, że sytuacja, że węzeł występuje w danym stanie, jest w 100% pewna. Po prowadzeniu takiego dowodu następuje przepływ informacji przez poszczególne węzły sieci i jako rezultat uzyskujemy wartości prawdopodobieństw *a posteriori* w węźle *H*. W niektórych sytuacjach do sieci wprowadzony może być tzw. miękki dowód, tj. dowód, gdy ocenia się, że nie można być pewnym, że węzeł występuje w danym stanie ze 100% pewnością. W takim przypadku wprowadzona wartość prawdopodobieństwa do węzła jest mniejsza niż 1 (100%).

Istnieją liczne programy (HUGIN [14], GeNIe [12]), które mogą być zastosowane w celu analizy modeli BN i one nie wymagają żadnej wiedzy z zakresu programowania, w przeciwieństwie do prób zastosowania modeli *LR*. Jedną z niedogodności stosowania BN w naukach sądowych jest to, że informacja uzyskiwana w węźle dla da-

nych dyskretnych, odpowiedzialnym za analizę rozważanych hipotez, po wprowadzeniu dowodu podawana jest w postaci następujących prawdopodobieństw warunkowych, tj. $Pr(H_1|E)$ i $Pr(H_2|E)$.

Jak już wspomniano, zadaniem biegłego sądowego jest ocena wartości dowodowej dostarczonych danych (E) w kontekście rozważanych hipotez, tj. prawdopodobieństw warunkowych $Pr(H_1|E)$ i $Pr(H_2|E)$ w formie LR. Iloraz wiarygodności można również wyznaczyć, stosując informację uzyskaną po wykonaniu analizy opartej na modelach BN, tj. gdy uwzględnimy informację o prawdopodobieństwach *a priori* wprowadzonych do węzła H, ponieważ możemy wówczas zastosować poniższą zależność wynikającą z teorematu Bayesa:

$$\frac{Pr\ H_1|E}{Pr\ H_2|E} \quad \frac{Pr\ E|H_1}{Pr\ E|H_2} \quad \frac{Pr\ H_1}{Pr\ H_2} \quad LR \quad \frac{Pr\ H_1}{Pr\ H_2}. \quad \{1\}$$

Dla przykładu, gdy $Pr(H_1) = Pr(H_2) = 0{,}5$, to wówczas wartość LR jest równa stosunkowi prawdopodobieństw *a posteriori* $Pr\ H_1|E \ / Pr\ H_2|E$. Oczywiście taki sposób wyznaczania LR może być wykorzystany, gdy rozważane są tylko dwie hipotezy w węźle H.

1.1. Ocena wartości dowodowej okruchów szkła

Zastosowanie BN w celu rozwiązania problemów z zakresu interpretacji danych analizowanych w kryminalistyce będzie omówione na przykładzie interpretacji okruchów szkła. Szkło jest materiałem powszechnie stosowanym w wielu dziedzinach życia człowieka. Dlatego też jest często spotykane na miejscu takich zdarzeń, jak wypadek samochodowy, włamania czy bójki. Fakt ten powoduje również, że szkło jest jednym z często analizowanych materiałów dowodowych [9, 22]. Różne są cele analizy okruchów szkła dla potrzeb wymiaru sprawiedliwości. Problem porównywania okruchów szkła jest związany z następującym pytaniem zadawanym przez przedstawicieli wymiaru sprawiedliwości – czy dwie porównywanie próbki szkła mogą pochodzić z tego samego obiektu? Jest to tzw. analiza na „poziomie źródła" [1, 3, 22]. Innym zadaniem związanym z oceną fragmentów szkła dla potrzeb wymiaru sprawiedliwości jest problem analizy na tzw. „poziomie aktywności", tj. udzielenie odpowiedzi na pytanie, czy podejrzany jest osobą, która rozbiła obiekt szklany na miejscu zdarzenia? Rozwiązanie tego drugiego problemu wymaga połączenia informacji o danych uzyskanych w sposób obiektywny (np. informacje o liczbie okruchów szkła ujawnionych na odzieży podejrzanego, dla których uzyskano $LR > 1$ lub $LR < 1$, gdy porównywano je do próbki porównawczej) z wiedzą biegłego dotyczącą zależności występujących pomiędzy obiektywnymi danymi, która to wiedza ze swej natury jest wiedzą subiektywną. Problem analizy na poziomie aktywności może być rozwiązany poprzez zastosowanie modelu LR, w którym H_1 – to podejrzany jest osobą, która rozbiła okno na miejscu zdarzenia; H_2 – inna osoba rozbiła okno na miejscu zdarzenia. Model, który pozwala na taką ocenę, jest opisany w literaturze przedmiotu [9]. Jego zastosowanie wymaga wiedzy o liczbie fragmentów szkła odzyskanych z odzieży podejrzanego, które wykazują wartość $LR > 1$ lub $LR < 1$, gdy zestawiono ich właściwości fizykochemiczne z próbką porównawczą, tj. na poziomie „źródła". Tym samym proces grupowania jest integralną częścią tego modelu i składa się on z następujących etapów: 1) pogrupowania próbek szkła według wzrastającej średniej wartości współczynnika załamania światła wyznaczonej dla nich (RI); 2) obliczenia odległości (miara podobieństwa) pomiędzy nimi, w a następnie wyszukanie odległości, które są większe niż wartość krytyczna tej odległości uzyskana z odpowiednich tablic; 3) obliczenie wartość ilorazu wiarygodności na „poziomie źródła", np. według modelu opisanego w [9]. Jednak ta procedura grupowania może być zastosowana wyłącznie w przypadku, gdy fragmenty szkła są opisane przez jedną zmienną. Tym samym nie może być zastosowana w przypadku analizy próbek opisanych przez większą liczbę zmiennych (np. skład pierwiastkowy), ponieważ nie można we wspomniany wyżej sposób dokonać grupowania wektorów zmiennych. Dlatego też zaproponowano model oparty na sieciach bayesowskich (BN).

1.2. Wartość dowodu z profilowania DNA

Identyfikacja ludzkich szczątków, analiza śladów z miejsca zdarzenia oraz dochodzenie spornego ojcostwa obecnie są przeprowadzane za pomocą testowania DNA przy użyciu wielu markerów genetycznych, takich jak krótkie tandemowo powtarzające się sekwencje (STRs – ang. short tandem repeats) lub polimorficzne regiony mitochondrialnego DNA. W pewnych przypadkach powstaje konieczność wprowadzenia polimorficznych STR-ów rozlokowanych na chromosomach X lub Y celem wzmocnienia wartości wyników analizy genetycznej dla celów sądowych [11]. Rozważając znaczenie analizy DNA dla systemu prawnego, dalsze badania z zastosowaniem sieci bayesowskich wydają się bardzo obiecujące. Pozwalają one na zredukowanie niepewności, która może powstawać przy okazji prezentowania istotnych zależności pomiędzy wartością różnego rodzaju dowodów w sposób logiczny. Jeżeli utworzy się jedną sieć, to wówczas może być ona wykorzystana wielokrotnie dla podobnych przypadków [13].

Technologia obliczeniowa sieci bayesowskiej może być zastosowana do obliczania ilorazu wiarygodności (ang. likelihood ratio), co przeprowadza się począwszy od wstępnej reprezentacji problemu identyfikacji dla celów sądowych aż do utworzenia właściwej sieci. Skomplikowane drzewa genetyczne znakomicie „wpasowują"

się w model sieci bayesowskiej. Zależności rodzinne badane przy pomocy markerów jądrowych stanowią naturalne modularne bloki tej reprezentacji: rodzice i dzieci stały się częścią ogólnej terminologii sieci bayesowskiej, także dla niegenetycznych zastosowań. Graf związany z drzewem genetycznym jest zbiorem węzłów reprezentujących zmienne (genotypy lub allele), a także zbiorem skierowanych krawędzi ilustrujących związki pomiędzy tymi węzłami. Najprostszy graf przedstawia standardową transmisję genów od rodziców do ich dzieci. Całkowite prawdopodobieństwo takiej struktury determinowane jest przez zdefiniowanie tablic prawdopodobieństw warunkowych (CPT, ang. conditional probability tables) dla każdej zmiennej przy założeniu prawdopodobieństw ich zmiennych „rodzicielskich". Mendlowskie reguły dziedziczenia oraz logiczne zależności pomiędzy genami a genotypami stanowią podstawę wyznaczania warunkowych prawdopodobieństw. Po wprowadzeniu dowodu wynikającego z badanej sprawy do odpowiedniego węzła genotypowego, można ten „dowód" propagować przez sieć. Oczekiwany iloraz wiarygodności wynikający z analizy jednego markera można otrzymać z rozkładu *a posteriori* w węźle *querry*. Dawid i in. zaproponowali podstawowe fragmenty sieci skupiające się na indywidualnych genach i genotypach [10]. Zgodnie z powyższym, węzeł *gt*, oznaczający genotyp, jest modelowany jako logiczna kombinacja alleli możliwych do odziedziczenia po matce i ojcu. Te dziedziczone po rodzicach allele są reprezentowane przez węzły *mg* (ang. maternal gene) i *pg* (ang. paternal gene). Wstępne informacje na temat frekwencji występowania alleli matczynych i ojcowskich są zebrane w tabelach towarzyszących wymienionym wyżej węzłom (tzw. parentalnych). Początkowa informacja na temat częstości występowania wszystkich możliwych genotypów dziecka jest zawarta w tablicy prawdopodobieństw warunkowych poniżej węzła *cgt* (ang. child's genotype) (rycina 2). Częstości poszczególnych alleli zostały otrzymane na podstawie badań populacji Polski południowej [19]. Dla uproszczenia, definicja dowodu jest tu ograniczona do wyników analizy genetycznej jednego lokus, czy markera. Przykład zastosowania tej metody do badania pochodzenia został przedyskutowany przez Dawida i in. [10]. Autorzy pokazali, jak odpowiednie graficzne struktury sieci bayesowskiej można otrzymać z początkowej reprezentacji identyfikacji człowieka dla celów sądowych (rycina 3).

Oszacowanie prawdopodobieństwa, że pozwany jest biologicznym ojcem dziecka wymaga:

1. wyznaczenia wartości ilorazu dwóch prawdopodobieństw warunkowych: uzyskania dowodu (brak wykluczenia pozwanego jako ojca), jeżeli prawdą jest, że pozwany jest biologicznym ojcem dziecka do prawdopodobieństwa uzyskania tegoż dowodu, jeżeli brak wykluczenia pozwanego jest dziełem przypadku. Jest to definicja tzw. współczynnika ojcostwa (ang. paternity index);
2. oszacowania prawdopodobieństwa *a priori* ojcostwa pozwanego wynikającego z innych niż dowód z badania DNA przesłanek w sprawie.

Dowodem jest tu analiza profili genetycznych matki, dziecka i pozwanego, która wskazuje, iż pozwanego nie można wykluczyć jako biologicznego ojca spornego dziecka.

Takie ilorazy wiarygodności (zwane inaczej współczynnikami ojcostwa) zależą od możliwego pokrewieństwa pomiędzy matką dziecka a pozwanym czy też alternatywnym ojcem dziecka [18] (rycina 4). Przewagą logicznego podejścia do wartościowania dowodu jest to, że iloraz wiarygodności można wyznaczać w kontekście okoliczności całej sprawy. Jednakże w rzeczywistych sprawach czasem trudno jest sprowadzić problem do stawiania tylko do dwóch hipotez, czyli hipotezy oskarżenia (podejrzany jest winny) i obrony (podejrzany jest niewinny). W sprawach dotyczących identyfikacji nieznanych szczątków albo w sprawach o prawdopodobne kazirodztwo sąd może zażyczyć sobie rozważenie więcej niż dwóch możliwych wytłumaczeń uzyskanych wyników analiz dotyczących rzeczywistego sprawcy. Zakładając, że podejrzany może mieć więcej lub mniej bliskich krewnych, liczba możliwych hipotez obrony wzrasta. Zapewnia to ogólny wzór teorematu Bayesa. Stąd hipoteza obrony może zostać podzielona na kilka „podhipotez" (subpropozycji), które są wzajemnie wykluczające się oraz wyczerpujące.

W przypadkach dochodzenia spornego ojcostwa mogą takie hipotezy wyglądać, jak następuje:

- H_1: pozwany jest ojcem dziecka;
- H_2: brat pozwanego jest biologicznym ojcem dziecka;
- H_3: kuzyn pozwanego jest biologicznym ojcem dziecka;
- H_4: biologicznym ojcem dziecka jest nieznany i niespokrewniony z pozwanym mężczyzna.

Zgodnie z Champod [17], wymienione hipotezy mogą zostać nazwane subpropozycjami. Takie podejście wymaga przyjęcia pewnych założeń co do wiedzy *a priori* o możliwym ojcostwie czy też o pochodzeniu nieznanych szczątków ludzkich.

W rutynowych kryminalnych sprawach dochodzenia spornego ojcostwa (jak w przypadku podejrzenia o kazirodztwo) lub w razie identyfikacji nieznanych szczątków ludzkich, zwykle stosuje się panel 16 autosomalnych markerów do wyznaczenia profilu DNA człowieka. Jednakże w pewnych przypadkach, kiedy brak lub do dyspozycji jest niewiele materiału porównawczego od potencjalnych bliskich krewnych, otrzymane ilorazy wiarygodności mogą być zbyt niskie, ażeby uznać prawdopodobieństwo identyfikacji *a posteriori* jako wystarczający dowód identyfikacji.

Powstaje zatem konieczność rozszerzenia zakresu rutynowo stosowanych markerów. W zależności od rodza-

ju sprawy, użyteczne mogą być markery mitochondrialnego DNA oraz charakterystyczne markery chromosomów Y lub X. W wyniku przeprowadzonych badań kilka różnych rodzajów dowodów można zebrać razem, co może zapewnić pojedynczy iloraz wiarygodności. Oszacowanie wyniku badania autosomalnych markerów DNA przeprowadza się, stosując regułę mnożenia, jednakże oszacowania częstości występowania mitochondrialnch czy związanych z chromosomem Y haplotypów oszacowuje się na podstawie odpowiedniej bazy danych. Dlatego przemnożenie tych wyników stanowi wstępne przybliżenie łącznego ilorazu wiarygodności. Jednym z problemów dotyczących tego podejścia jest to, że informacja zawarta w rodzaju haplotypu chromosomu Y czy mitotypu może być silnie zależna od pochodzenia etnicznego danej osoby. Dlatego zaproponowano podejście, które prowadzi przez dostosowanie otrzymywanych prawdopodobieństw zgodności przy zastosowaniu współczynnika odziedziczalności (ang. coancestry coefficient) [8].

Innym zastosowaniem sieci bayesowskich w genetyce sądowej jest możliwość przewidywania cech pigmentacyjnych człowieka na podstawie wyników badań genetycznych. Skrupulatna wiedza dotycząca zmienności genów oraz jej związku z cechami fizycznymi jest niezbędna do przewidywania fenotypu człowieka, co mogłoby być przydatne w genetyce sądowej. Pigmentacja jest złożoną fizyczną cechą, w wynik której zaangażowane jest wiele genów. Najnowsze dane wskazują, że najwyższa asocjacja dla niebieskich – nieniebieskich kolorów oczu występuje dla pojedynczego SNP (ang. single nucleotide polymorphism) w intronie 86 genu *HERC2* rs12913832. Inny badany wariant genu *OCA2* rs1800407 jest związany z determinacją zielonego koloru tęczówki oka [4]. Ślady biologiczne dla celów sądowych zgromadzone na miejscu zdarzenia są zwykle poddawane identyfikacji przy użyciu wysoce zmiennych markerów typu STR. Otrzymane profile DNA są porównywane z profilami DNA próbek referencyjnych, a uzyskana zgodność jest silnym dowodem podtrzymującym hipotezę o wspólnym pochodzeniu analizowanych próbek. Jeżeli w danej sprawie nie wytypowano podejrzanego, otrzymane profile DNA są przesyłane do państwowej bazy danych profili DNA. Niestety jeżeli nie uzyska się ich zgodności, to śledztwo ulega znacznemu spowolnieniu. Jednak uzyskane próbki mogą być dalej użyteczne dla celów operacyjnych. Genetyczne przewidywanie fizycznych cech człowieka staje się nową dziedziną nauk sądowych i niewątpliwie może się ona okazać niezwykle użyteczna przy prowadzeniu dochodzenia.

Ażeby graficznie przedstawić zależności pomiędzy obserwowanymi danymi, zaprojektowano sieć bayesowską umożliwiającą wnioskowanie co do prawdopodobieństwa koloru włosów, skóry i tęczówki oka danej osoby na podstawie jej znanego genotypu w zakresie *MC1R*, *HERC2* rs1291332 i *OCA2* rs1800407. Taka sieć umożliwia śledzenie, jak zmiana wiedzy o prawdopodobieństwie stanu jednej zmiennej (np. rodzaj genotypu *MC1R* czy też *HERC2*) może zmieniać poziomy pewności innych zmiennych (np. prawdopodobieństwo danego koloru włosów nieznanego dawcy próbki DNA z miejsca zdarzenia).

Głównym celem tej pracy dotyczącej interpretacji dowodu z badania DNA było zilustrowanie pewnych aplikacji sieci baysowskich zaprojektowanych dla oszacowania wartości dowodu badania DNA. Zaprezentowano pewne problemy odnoszące się do testowania pokrewieństwa (dochodzenie spornego ojcostwa oraz identyfikacja osób), koherentne oszacowanie różnych rodzajów wyników profilowania DNA, a także możliwość przewidywania fenotypu człowieka na podstawie wyników badań jego genomu.

2. Metody

2.1. Sieć bayesowska – ocena wartości dowodowej okruchów szkła na poziomie aktywności

Model zaprezentowany na rycinie 3 zastosowano w celu analizy problemu oceny dowodu w postaci okruchów szkła na poziomie aktywności, tj. czy podejrzany mógł rozbić konkretny obiekt szkła? Analiza zaproponowanego modelu została przeprowadzona z wykorzystaniem programu Hugin Researcher v. 6.9 [14]. W proponowanym modelu każdy okruch szkła jest najpierw pojedynczo porównywany do próbki kontrolnej za pomocą odpowiedniego modelu *LR*, tj. wykonywana jest analiza na poziomie „źródła". Liczbę próbek, dla których stwierdzono $LR > 1$ i $LR < 1$, gdy porównywano je do próbki porównawczej za pomocą modeli *LR* opublikowanych w [1, 2], oznaczono odpowiednio N_p i N_n. Węzeł N_p miał cztery stany: 0 – dla żadnego fragmentu szkła ujawnionego na odzieży podejrzanego uzyskano $LR > 1$ na poziomie źródła, 1, 2, 3 – jeden, dwa, trzy i więcej fragmentów szkła ujawnionych na odzieży podejrzanego, dla których uzyskano $LR > 1$ na poziomie „źródła". Węzeł N_n miał trzy stany: 0 – dla żadnego z fragmentów szkła ujawnionych na odzieży podejrzanego nie uzyskano $LR < 1$ na poziomie „źródła", 1, 2 – jeden, dwa i więcej fragmentów szkła ujawnionych na odzieży podejrzanego, dla których uzyskano $LR < 1$ na poziomie „źródła". Przypisane tym stanom prawdopodobieństwa warunkowe zaprezentowane są w tabelach I oraz II. Węzeł *H* ma dwa stany: H_1 – podejrzany rozbił obiekt szklany, H_2 – inna osoba rozbiła ten obiekt szklany. Założono, że $Pr(H_1) \quad Pr(H_2) \quad 0{,}5$. Węzeł *Q* również miał dwa stany: *Q* – „tak" (ang. „yes"), tj. stosowany model *LR* do oceny wartości fizykochemicznych działa poprawnie, $\overline{Q}$ – „nie" (ang. „no") – stosowany model *LR* do oceny wartości fizykochemicznych nie działa poprawnie. Zało-

żono na podstawie wiedzy biegłego, że $Pr(Q)$ i $Pr(\overline{Q})$ są równe odpowiednio 0,9 i 0,1.

Na przykład gdy założono, że podejrzany stłukł obiekt szklany, tj. H_1 jest prawdziwa, to wówczas jest bardzo prawdopodobne, że na jego odzieży można znaleźć 2 fragmenty szkła $Pr(N_p \quad 2|H_1,Q)$, które wykażą $LR > 1$, gdy analiza na poziomie „źródła" będzie wykonana i gdy zastosowany model *LR* dostarczył poprawnej odpowiedzi. Na podstawie wiedzy autorów założono, że taka sytuacja wystąpiła w 50% analizowanych przypadków, tj. $Pr(N_p \quad 2|H_1,Q) \quad 0{,}5$. W tej samej sytuacji uzyskanie $LR < 1$ dla obu analizowanych fragmentów wydaje się bardzo rzadkim zdarzeniem. Dlatego też założono, że $Pr(N_p \quad 2|H_1,Q) \quad 0{,}005$. Gdy H_2 jest poprawna, to wówczas wartość $Pr(N_p \quad 2|H_1,Q)$ zależy od liczby N_p fragmentów szkła, które wykazują podobieństwo do próbki porównawczej przez przypadek, tj. pochodzą one z innego obiektu szklanego niż obiekt szklany rozbity na miejscu zdarzenia. $Pr(N_p \quad 2|H_1,Q)$ jest również związane z sytuacją, że znalazły się one na odzieży podejrzanego przez przypadek. Ponadto na podstawie wiedzy autorów można zasugerować, że te prawdopodobieństwa są większe, gdy Q jest prawdzie.

2.2. Sieć bayesowska – modele dla wyników analizy DNA

Sieci bayesowskie do oszacowania łącznego ilorazu wiarygodności na podstawie wyników analizy genetycznej różnego rodzaju markerów oraz pozostałe sieci zaprojektowano przy użyciu programu komputerowego Huigin Researcher v. 6.9 [14]. Tablice prawdopodobieństw warunkowych związane z poszczególnymi węzłami sieci wymagają ustalenia pewnych probabilistycznych założeń. Zgodnie z wiedzą wynikającą z rutynowej praktyki, ustalono średnie wartości dla kilku parametrów. W rutynowej pracy ekspertowskiej dotyczącej dochodzenia spornego ojcostwa (w sprawach kryminalnych) lub identyfikacji osób stosuje się zestawy do analizy STR: Amp*Fl* Identifiler, Amp*Fl* Yfiler, Amp*Fl* Minifiler, a także Big Dye Terminator Cycle Sequencing Ready Reaction Kit do sekwencjonowania mitochondrialnego DNA (wszystkie firmy Applera). Wyznaczenie współczynnika ojcostwa przeprowadzono według Dawida i in. [10]. Indywidualny genotyp osoby (węzły genotypowe, jak na przykład prawdopodobny ojciec – ang. putative father – „pfgt") został rozdzielony na czynniki podstawowe, nieobserwowalne, ojcowsko lub w linii matczynej dziedziczone geny (węzły alleli takie, jak „pfpg"– ang. putative father paternal gene – ojcowskie geny prawdopodobnego ojca). Węzeł hipotez *query* zawiera dwie opcje: hipotezę H_1 zakładającą wartość: „prawda" i H_2 – zakładającą wartość „fałsz". Założono, że obydwie hipotezy są jednakowo prawdopodobne (*a priori*), a zatem po propagacji dowodu, iloraz ich prawdopodobieństw *a posteriori* stanowi współczynnik ojcostwa otrzymany dla pojedynczego markera genetycznego. Informację *a priori* o częstościach występowania w populacji poszczególnych alleli zaczerpnięto z poprzednich prac [19].

Sieć dla przewidywania cech fizycznych składa się z trzech węzłów genowych: „*MC1R*", „*OCA2*", „*HERC2*" oraz z trzech węzłów dotyczących rodzaju fenotypów: „kolor włosów" (ang. hair colour), kolor skóry" (ang. skin colour) i „kolor oczu" (ang. eye colour). Dla uproszczenia założono na przykład, że kolor włosów jest determinowany przez dwa czynniki: rodzaj genotypu w zakresie genu *MC1R* oraz genu *HERC2* rs12913832. Całkowite prawdopodobieństwo uzyskanej struktury jest determinowane przez specyfikację tablic prawdopodobieństw warunkowych dla każdego węzła na podstawie uzyskanych wyników badań genetycznych. Stratyfikacja populacji (proporcje występowania osób o różnych kolorach oczu, tęczówki oka i skóry w analizowanej próbie 388 osób z Polski południowej) została wykorzystana jako informacja *a priori*. Tablice prawdopodobieństw zostały ustalone zgodnie z pracami [5, 6]. Na przykład węzeł *HERC2* reprezentuje wszelkie prawdopodobieństwa dla $Pr(HERC2|\text{hair},\text{skin},\text{eye\_colour})$. Po ustaleniu tzw. „twardego dowodu", czyli stwierdzenia, że jeden ze stanów danej zmiennej jest pewny (na przykład: próbka z miejsca zdarzenia w genie *MC1R* ma genotyp rodzaju 160/160) można niejako „przesyłać" (propagować) tę wiedzę poprzez sieć. Co, z kolei pozwala ocenić prawdopodobieństwo posiadania przez dawcę próbki określonego koloru włosów.

Nazwy poszczególnych stanów i prawdopodobieństwa *a priori* poszczególnych stanów każdej zmiennej są zilustrowane na oknach towarzyszących każdemu węzłowi.

3. Rezultaty i dyskusja

3.1. Sieć bayesowska – ocena wartości dowodowej okruchów szkła na poziomie „źródła"

Sprawdzono poprawność działania zaproponowanej sieci, wykonując dwa eksperymenty. W pierwszym eksperymencie przeanalizowano wpływ zmian opinii odnoszącej się do poprawności działania zastosowanego modelu *LR* do interpretacji danych fizykochemicznych, tj. $Pr(Q)$. Rezultaty wykonanej analizy przedstawiono w tabeli III.

Na ich podstawie można stwierdzić, że model działa poprawnie, ponieważ im silniejsza opinia o poprawności działania modelu *LR*, tym większa wartość stosunku:

$$\frac{Pr(H_1|N_p,N_n)}{Pr(H_2|N_p,N_n)},$$

który jest równoważny wartości *LR* dla rozpatrywanych na poziomie „aktywności" hipotez, ponieważ założono, że $Pr(H_1) = Pr(H_2) = 0{,}5$.

Drugi eksperyment polegał na tym, że zmieniano wartości $Pr(H_1|N_p, N_n)$ i $Pr(H_2|N_p, N_n)$, które można było zaobserwować, gdy zastosowano różne zbiory $Pr(N_p|H_1, Q)$. Różnice w wartościach $Pr(N_p|H_1, Q)$ mogą być związane z wpływem różnych czynników istotnych do oceny wartości dowodowej okruchów szkła na poziomie aktywności, np. różnicy czasu pomiędzy zajściem zdarzenia, w czasie którego rozbito obiekt szklany, a czasem zabezpieczenia odzieży do badań. Liczba okruchów szkła odzyskiwanych z odzieży podejrzanego zmniejsza się wraz z upływem tego czasu. Uzyskane rezultaty:

$$\frac{Pr(H_1|N_p, N_n)}{Pr(H_2|N_p, N_n)} \quad LR$$

zaprezentowane w tabeli IV potwierdzają, że zaproponowany model BN działa poprawnie, ponieważ wzrost liczby okruchów wykazujących $LR > 1$ powoduje wzrost wartości *LR*. Ponadto wpływ czasu został poprawnie oceniony w proponowanym modelu BN w przypadku oceny wartości dowodowej (*LR*) trzech okruchów szkła w przypadku ich ujawnienia po upływie 8 h lub 2 h. Stwierdzono bowiem, że ta liczba okruchów szkła ujawniona po upływie 8 h ma większą wartość dowodową niż gdy ujawniona jest ona po 2 h. Jest to zgodne z logiką, ponieważ wraz z upływem czasu można ujawnić mniej okruchów. Tym samym rzadziej spotykana sytuacja daje większą wartość dowodową wyrażoną w postaci *LR* niż sytuacja spotkana częściej.

3.2. Sieć bayesowska – modele dla analizy DNA

3.2.1. Model dla spraw dochodzenia spornego ojcostwa

Węzeł hipotezy: *paternity* (ojcostwo) posiada cztery stany: *he* (on), *his close paternal relative* (jego bliski krewny w linii ojcowskiej), *his distant relative* (jego daleki krewny) oraz *unrelative* (niespokreniony) z pozwanym, który mógłby być biologicznym ojcem dziecka. Węzeł zgodności w zakresie chromosomu Y zależy od węzła *haplotype* (haplotyp), który reprezentuje rodzaj zgodnego haplotypu, czyli *unseen before* (niespotykany wcześniej w danej populacji), *rare* (rzadki), *common* (powszechny w populacji). Zależny jest on także od możliwej mutacji w procesie spermatogenezy. Założono, że prawdopodobieństwo wynikające z analizy autosomalnych markerów (brak wykluczenia pozwanego jako ojca w zakresie 16 markerów typu STR), pod warunkiem prawdziwości wybranych hipotez, nie jest niezależne od wiedzy wynikającej z analizy chromosomu Y, stąd skierowana krawędź (wpływ) pomiędzy węzłem „Chr Y" a węzłem „autosomal". Całkowitą strukturę sieci ilustruje rycina 6.

3.2.2. Model dla spraw dotyczących identyfikacji szczątków nieznanej osoby

Węzeł hipotezy identity posiada cztery jednakowo prawdopodobne stany: *he* (czyli osoba, która była wskazana w trakcie prowadzonego dochodzenia jako poszukiwana osoba), *his maternal relative* (jego krewny w linii matczynej), *his paternal relative* (jego krewny w linii ojcowskiej) i *unrelated* (nieskpokrewniony). Autosomalne prawdopodobieństwa zgodności (tabela V) zależą od rodzaju dostępnego materiału referencyjnego; im bliżsi są możliwi krewni, tym silniejszy jest wpływ na prawdziwość hipotezy, że szczątki pochodzą od wytypowanej w trakcie dochodzenia osoby *he*. Proponowana struktura sieci jest zilustrowana na rycinie 7.

Zainteresowanie probabilistycznym oszacowaniem dowodu z badania DNA zdecydowanie wzrosło w trakcie ostatniej dekady. Szczególnie łączy się to z zagadnieniami dotyczącymi testowaniem zależności rodzinnych, w tym skomplikowanych spraw związanych z dochodzeniem spornego ojcostwa czy identyfikacją szczątków nieznanej osoby [17. 20]. Można zbudować sieci bayesowskie, które będą reprezentować wyżej wymienione problemy. Tego rodzaju sieci pozwalają uzyskać właściwe ilorazy wiarygodności na bazie ustalonych hipotez z uwzględnieniem wszelkich dostępnych dowodów. Opisane sieci bayesowskie wydają się użyteczne do rozwiązywania wielu różnych scenariuszy poprzez zmianę poszczególnych prawdopodobieństw warunkowych przypisanych węzłom sieci adekwatnym dla danej sprawy. Odpowiadają one za różne hipotezy oskarżenia oraz podzielone hipotezy obrony, włączając w to różnych krewnych z linii matczynej lub ojcowskiej jako możliwe źródła materiału genetycznego, uwzględniając tym możliwość mutacji. Proponowane sieci bayesowkie są raczej proste i mogą być rozszerzone w różny sposób w zależności od informacji dotyczących okoliczności sprawy. Wstępne doświadczenia autorów niniejszej pracy z zastosowaniem sieci bayesowskich do oszacowania siły dowodu z badania DNA różnych markerów genetycznych wydają się zapewniać wiarygodne ilorazy wiarygodności na rzecz założonych hipotez oraz będą mogły stanowić dodatkowe i wydajne narzędzie w praktyce nauk sądowych.

Właściwe oszacowanie wyników genotypowania jest fundamentalnym problemem w badaniach sądowych. Homozygotyczne genotypy *MC1R*, tak zwane R-R, znacząco podnoszą prawdopodobieństwo, że dawca materiału genetycznego posiadał rude włosy oraz jasną skórę. Stąd jednym z celów niniejszej pracy było zaproponowanie modelu wnioskowania przy użyciu sieci bayesowskich na podstawie wyników z badań genetycznych (rycina 8).

Zaproponowany model zapewnia naukowe sprawdzenie wyników i wniosków otrzymanych z badania genotypów *MC1R*, *HERC2* (rs12913832) i *OCA2* (rs1800407). Co więcej, zaproponowany model może zostać wykorzystany przez innych praktyków i rozszerzony o dodatkową wiedzę dotyczącą wpływu innych genów na cechy fizyczne człowieka.

Dowód w odniesieniu do zmiennej stanowi stwierdzenie prawdziwości jednego z jej możliwych stanów. Jeżeli stan zmiennej zostaje ustalony, określa się to jako „silny" dowód, w przeciwnym wypadku traktuje się go jako „słaby". Szczegółowa praca przeprowadzona ostatnio przez Branickiego i współpracowników [4] w zakresie analizy genu *OCA2* rs1800407 wykazała, że wariant T/T tego zmiennego miejsca (SNP) ma znaczącą korelację z zielonym kolorem oczu. Stąd, jeżeli otrzyma się próbkę DNA posiadającą dany wariant genu, np. T/T w *HERC2*, to operacyjna rada dla policji może wyglądać następująco: „sprawca powinien posiadać jasne włosy z prawdopodobieństwem około 58%, jasną skórę z prawdopodobieństwem około 86% oraz zielone oczy z prawdopodobieństwem około 86%". Natomiast jeżeli otrzyma się próbkę DNA posiadającą w zakresie genu *MC1R* genotyp 151/160, to operacyjna hipoteza może brzmieć następująco i jest niemal pewna: „sprawca posiada ciemnorude lub rude włosy".

4. Wnioski

Zaproponowane sieci mogą być łatwo rozszerzone w przypadku ustalenia nowych informacji w danej sprawie i (lub) jeżeli charakteryzują się sporym zakresem niepewności [7]. Ten ostatni czynnik jest szczególnie istotny, gdy dane wprowadzone do sieci bayesowskiej są szacowane wyłącznie na podstawie doświadczenia eksperta. Mając do dyspozycji dodatkową wiedzę, na przykład dotyczącą zmienności nowych genów wpływających na ludzki fenotyp, można uzyskać większe możliwości predykcji poszczególnych cech.

Jednakże zastosowanie sieci bayesowskich, mimo wielu zalet, posiada także pewne ograniczenia. Przedstawiony model wprowadza indywidualne, subiektywne przekonania odnośnie do przyczynowych zależności a także założeń wstępnych. Na przykład rozkład *a priori* genotypów *MC1R* czy różnych kolorów włosów oszacowano na podstawie badania 388 osób; z drugiej strony, trzeb mieć świadomość, że ta próba może nie być reprezentatywna dla całej populacji. Innym poważnym problemem dotyczącym łączenia różnego rodzaju dowodów genetycznych jest powstająca możliwość zaistnienia zależności pomiędzy rozkładem haplotypów chromosomu Y, mitotypami oraz rozkładem markerów autosomalnych w danej populacji. Stąd istnieje konieczność przeprowadzenia dalszych badań, zarówno doświadczalnych, jak też teoretycznych, celem usprawnienia możliwości wnioskowania czy dokładności modelu bayesowskiego, zanim będzie on mógł być zastosowany w praktyce.

Podziękowania
Niniejsze badania zostały sfinansowane dzięki programowi Leonardo da Vinci „Źródła niepewności w oszacowaniu wartości dowodowej materiału dowodowego".