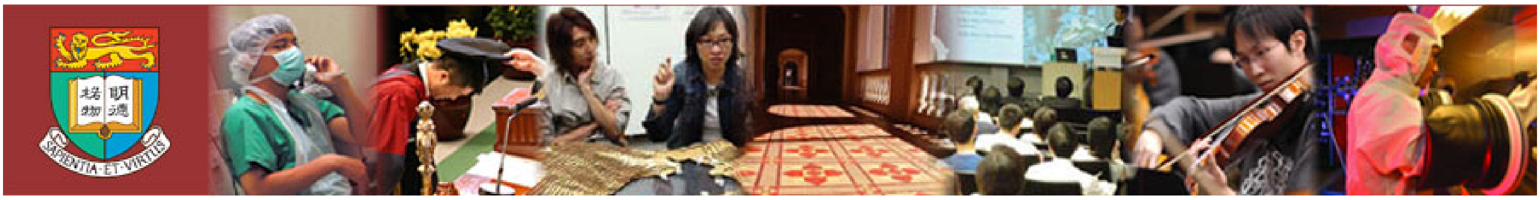

Title	Interpreting DNA mixtures with relatives of a missing suspect
Author(s)	Hu, YQ; Fung, WK; Choy, YT
Citation	The 1st International Conference on Remote Sensing, Environment and Transportation Engineering (RSETE 2011), Nanjing, China, 24-26 June 2011. In Proceedings of the 1st RSETE, 2011, p. 7649-7652
Issued Date	2011
URL	http://hdl.handle.net/10722/160497
Rights	Creative Commons: Attribution 3.0 Hong Kong License

Interpreting DNA Mixtures with Relatives of a Missing Suspect

Yue-Qing Hu
Institute of Biostatistics, School of Life Sciences
Fudan University, Email:yuehu@fudan.edu.cn

Wing K. Fung, Yan Tsun Choy
Department of Statistics and Actuarial Science
The University of Hong Kong

***Abstract*—Recent advances in DNA profiling have been proven extremely useful for forensic human identification. DNA mixtures are commonly found in serious crimes such as rape as well as voluminous crimes like theft. In this paper, one general formula is obtained for the evaluation of DNA mixtures when the suspect is unavailable for typing, but one maternal and one paternal relatives of the suspect are typed instead. In principle, closer relatives of the suspect will provide more genetic information on the genotype of the unavailable suspect. The effect of the relatives' DNA profiles on the interpretation of DNA mixtures is illustrated with case example.**

I. Introduction

DNA profiling or DNA fingerprinting has become a very popular and powerful method for human identification in forensic medicine since its inception two decades ago. It is found useful in a large variety of criminal offences including serious crimes such as murder and rape, as well as voluminous crimes like theft etc. Statistical treatment of DNA evidence has been commonly considered [1,2].

In forensic DNA analysis, the assessment of DNA profiles from biological samples containing a mixture of DNA from more than one person is often considered. For example, in a rape case the sample may contain materials from the victim, her consensual sexual parter(s) and/or the perpetrator(s); in a multiple murder case the knife may have blood specimen from the victims and/or the murder; in a theft case the DNA mixture is resulted when the thieves and the owners touch the door handle and leave their DNA there. The statistical assessment of such forensic DNA mixture problems has been regarded as a complex problem by the U.S. Second National Research Council Report on the evaluation of DNA evidence [3] (hereafter called NRC II).

In forensic science, it is common to have two propositions (hypotheses), the prosecution and defence propositions, giving two alternative explanations to the evidence. In that regard, the likelihood ratio (of two probabilities under the propositions) is often used to quantify the weight of DNA evidence. Weir et al. [4] and Fukshansky and Bär [5] obtained one general formula for the evaluation of the likelihood ratio when all persons involved came from the same population which was in Hardy-Weinberg equilibrium. The evaluations of the likelihood ratio of the DNA evidence under various situations such as different ethnic groups and subdivided populations have also been considered [6,7].

Sometimes, relatives of the victims, perpetrators and/or suspects are involved. For example, relatives of the victims are commonly involved in the case of rape, theft and murder. Brookfield [8] discussed the effect of relatives being the possible source of crime scene on the likelihood ratio. Evett [9] obtained a formula for the likelihood ratio which was used to assess the weight of DNA evidence in a case where the defence was "It was my brother." Belin et al. [10] described a method that summarized DNA evidence by addressing the possibility that a relative of the accused individual is a source of a crime sample. In the case of DNA mixed stains, Fukshansky and Bär [11] and Hu and Fung [12] investigated the evaluation of DNA mixture when a relative of a typed suspect was a potential perpetrator or when the suspect was unavailable for typing and his/her relative was typed instead. Statistical evaluations of DNA mixture when there were two groups of relatives had been considered [13].

This paper considers the evaluation of DNA mixtures with missing suspects; the suspects' relatives are, however, available for typing. To see the possible effects of the DNA profiles of the relatives, we motivate with a simple example having mixture $\{A_1, A_2, A_3, A_4\}$ contributed by one perpetrator and the victim with genotype A_3A_4. We let p_i be the allele frequency of type A_i. One man is suspected of having contributed to the mixed stain by the police but he is missing. Suppose that his father F or his paternal grandfather GF is available for typing. With regard to two competing propositions about whether or not the suspect is the perpetrator, i.e.

$$H_p: \text{ The suspect is the perpetrator;}$$
$$H_d: \text{ One unknown man is the perpetrator,}$$

the likelihood ratio (LR) is evaluated as follows:

(a) $LR = P(S = A_1A_2|F = A_1A_1)/(2p_1p_2) = 1/(2p_1)$ if the father F of the suspect S is typed as A_1A_1;

(b) $LR = P(S = A_1A_2|GF = A_1A_1)/(2p_1p_2) = 1/2 + 1/(4p_1)$ if the grandfather GF is typed instead and his genotype is A_1A_1.

It is interesting to note that the first LR may be bigger or smaller than the second one, depending on the value of p_1, the allele frequency of type A_1. Furthermore, it is not certain in this situation if the father/grandfather provides DNA evidence in favor of the suspect (i.e. $LR < 1$).

Suppose we have the situation that one more relative, say

978-1-4244-9171-1/11/$26.00 ©2011 IEEE

the mother M of the suspect, is typed and her genotype is A_2A_3, then the LR becomes

(a) $P(S = A_1A_2|M = A_2A_3, F = A_1A_1)/(2p_1p_2) = 1/(4p_1p_2)$ if M and F are typed;

(b) $P(S = A_1A_2|M = A_2A_3, GF = A_1A_1)/(2p_1p_2) = 1/(8p_2) + 1/(8p_1p_2)$ if M and GF are typed.

The first LR is always larger than the second one in this situation.

In this paper, we obtain a general formula for evaluating the likelihood ratio in the interpretation of DNA mixtures when the suspect is unavailable but his/her two relatives are typed instead. This general formula is also extended to deal with two sets of relatives where two unknown contributors are related and one unknown contributor is related to a typed person [12,13]. These formulae can be implemented easily by computer. One example is shown to illustrate the usefulness of the derived formulae. A few concluding remarks are finally given.

II. Evaluation of DNA mixtures

In some practical cases, it may be encountered that a suspect (denoted by S) is identified but his genotype cannot be typed for some reasons, e.g. escaped away. Instead, a relative of S is available for typing and Hu and Fung [12] have derived a general formula for the evaluation of LR in such a situation. In principle, there will be more information on the genotype of the unavailable suspect if closer relatives of S are typed. Suppose that two relatives of S, one maternal (denoted by R_1) and one paternal (denoted by R_2), are typed for reference. The maternal relative R_1 of S means that R_1 is a relative of the mother of S and is unrelated to the father of S. The paternal relative R_2 of S is defined similarly. In this section we consider the way to incorporate the genetic information of two relatives in the evaluation of DNA mixtures.

As we did in [12], let M be the set of distinct alleles found in the mixed stain, and K be the collection of genotypes of the typed people including R_1 and R_2. The proposition about whether the suspect was a contributor to the mixed stain is expressed as follows:

$$H : \text{The suspect and } x-1 \text{ unknowns were contributors.} \quad (1)$$

The known contributor(s) to the mixed stain must also be declared in the proposition H although we omit the details here for brevity. Notice that a typed person may be a contributor in one hypothesis and not be so in another. So a known contributor must be a typed person, but not vice versa. Of course, all unknown contributors are untyped. In evaluating the likelihood ratio about two competing propositions, it is necessary to calculate the probability

$$P(\text{Evidence}|H) = P(M, K|H) = P(K|H)P(M|K, H).$$

Since the probability $P(K|H)$ does not depend on the hypothesis H, it appears simultaneously in both the numerator and denominator of the likelihood ratio expression and so will be canceled out. Thus we only need to focus on the calculation of the conditional probability $P(M|K, H)$. We assume Hardy-Weinberg equilibrium and all people involved except S, R_1, and R_2 are biologically unrelated.

The mixture M is always taken to be composed of the alleles of the known and unknown contributors declared in H. Let U be the alleles set M with the removal of the alleles of the known contributors declared in H, then the alleles set U must be explained by the x unknown contributors declared in H. We assign capital letters with subscript g to represent the genetic profiles (distinct alleles) hereafter. For example, S_g and X_{g0} are the genetic profiles of the suspect S and $x-1$ unknown contributors in the hypothesis H and so $S_g \cup X_{g0}$ is the genetic profile of the x unknowns. Using Eq. (1) in [12] and considering the relationship among the typed persons and unknown contributors, we have

$$\begin{aligned} & P(M|K, H) \\ = \; & P(U \subset S_g \cup X_{g0} \subset M|K) \\ = \; & W(M) - \sum_{i \in U} W(M \setminus \{i\}) + \sum_{i,j \in U} W(M \setminus \{i, j\}) \\ & + \cdots + (-1)^{|U|} W(M \setminus U), \end{aligned} \quad (2)$$

where $|U|$ is the cardinality of set U and

$$W(D) = P(S_g \cup X_{g0} \subset D|K) \quad (3)$$

is defined for any subset D of M.

Since the independence among multiple loci is taken and the overall likelihood ratio can be achieved by multiplication across loci, we focus on the calculation of $P(M|K, H)$ at one locus and notice that the relevant M, K, U, and the allele frequencies will remain the same within this locus. It is convenient to write the mixture as $M = \{1, 2, \ldots, m\}$ with allele frequencies p_1, p_2, ..., p_m, respectively. For any given set $D \subset M$, and non-negative integer n, define

$$\begin{aligned} Q(n, D) = & \; s^n - \sum_{i \in D} (s - p_i)^n + \sum_{i,j \in D} (s - p_i - p_j)^n \\ & + \cdots + (-1)^{|D|} (s - \sum_{i \in D} p_i)^n, \end{aligned} \quad (4)$$

where $s = \sum_{i \in M} p_i$. It is noted that the quantity $Q(n, D)$ is determined by not only n and D, but also the frequencies of alleles in set M, which remain the same within that locus and hence we can view $Q(n, D)$ as a function of n and D only. Note $Q(0, \phi) = 1$ and $Q(n, D) = 0$ if $n < |D|$. The implementation of $Q(n, D)$ by computer is easy. As a special case, the quantity $Q(2x, U)$ is just the probability that the x unrelated unknown contributors have each allele present in set U and each allele of these contributors must be present in set M [4,5].

Notice that the kinship coefficients $(k_0, 2k_1, k_2)$ between two individuals are the probabilities that these two persons share 0, 1, and 2 ibd (identical by descent) alleles, respectively [14]. It is concluded that the suspect S and the maternal relative R_1 cannot share two ibd alleles, and so do S and R_2. Thus let $(k_0^{SR_1}, 2k_1^{SR_1}, 0)$ denote the kinship coefficients between individuals S and R_1, and $(k_0^{SR_2}, 2k_1^{SR_2}, 0)$ the kinship coefficients between S and R_2.

Theorem 1 Let $r_{11}r_{12}$ and $r_{21}r_{22}$ be the genotypes of the maternal relative R_1 and paternal relative R_2 of the suspect S, and R_1 and R_2 be biologically unrelated, then the conditional probability associated with the hypothesis H in Eq. (1) is

$$\begin{aligned}
&P(M|K,H) = k_0^{SR_1} k_0^{SR_2} Q(2x, U) \\
&+k_0^{SR_1} k_1^{SR_2} \{I_M(r_{21}) Q(2x-1, U \setminus \{r_{21}\}) \\
&\ +I_M(r_{22}) Q(2x-1, U \setminus \{r_{22}\})\} \\
&+k_1^{SR_1} k_0^{SR_2} \{I_M(r_{11}) Q(2x-1, U \setminus \{r_{11}\}) \\
&\ +I_M(r_{12}) Q(2x-1, U \setminus \{r_{12}\})\} + k_1^{SR_1} k_1^{SR_2} \\
&\{I_M(r_{11}) I_M(r_{21}) Q(2x-2, U \setminus \{r_{11}\} \cup \{r_{21}\}) \\
&\ +I_M(r_{11}) I_M(r_{22}) Q(2x-2, U \setminus \{r_{11}\} \cup \{r_{22}\}) \\
&\ +I_M(r_{12}) I_M(r_{21}) Q(2x-2, U \setminus \{r_{12}\} \cup \{r_{21}\}) \\
&\ +I_M(r_{12}) I_M(r_{22}) Q(2x-2, U \setminus \{r_{12}\} \cup \{r_{22}\})\},
\end{aligned} \tag{5}$$

where I is the indicator function, e.g. $I_M(r_{11}) = 1$ if $r_{11} \in M$ and 0 otherwise.

For the sake of clarity, the proof of Theorem 1 is omitted here. Detailed expressions of $P(M|K,H)$ for all possible combinations of the genotypes of R_1 and R_2 can be obtained by hand. Since the indicator function takes the value of 0 or 1, it can be seen from Eq. (5) that the conditional probability $P(M|K,H)$ is just a linear combination of quantities $Q(n,D)$ for various n and D, which facilitates the computation by a computer program.

Remark 1. If each allele of R_1 and R_2 is not present in the mixture M, we have $P(M|K,H) = k_0^{SR_1} k_0^{SR_2} Q(2x, U)$. So the smaller values of $k_0^{SR_1}$ and $k_0^{SR_2}$ imply the weaker evidential strength of the DNA mixture against the suspect S.

Remark 2. If S and R_2 are unrelated, i.e. the individual R_2 provides no genetic information about whether the suspect S is the contributor of the mixed stain, we have $P(M|K,H) = k_0^{SR_1} Q(2x, U) + k_1^{SR_1}\{I_M(r_{11}) Q(2x-1, U \setminus \{r_{11}\}) + I_M(r_{12}) Q(2x-1, U \setminus \{r_{12}\})\}$, which corresponds to the case that the suspect is unavailable and one maternal or paternal relative of the suspect is typed instead. Particularly, if only the mother of the suspect S, R_1, is typed for reference, then $P(M|K,H) = \{I_M(r_{11}) Q(2x-1, U \setminus \{r_{11}\}) + I_M(r_{12}) Q(2x-1, U \setminus \{r_{12}\})\}/2$. Symmetrically, if only the father of the suspect S, R_2, is typed for reference, then $P(M|K,H) = \{I_M(r_{21}) Q(2x-1, U \setminus \{r_{21}\}) + I_M(r_{22}) Q(2x-1, U \setminus \{r_{22}\})\}/2$. For more general results, see [12].

Remark 3. If S and R_1, S and R_2 are both pairwise unrelated, i.e. individuals R_1 and R_2 provide no biological information about whether the suspect S is the contributor of the mixed stain, then $P(M|K,H) = Q(2x, U)$, which is just the result reported in [4,5]. In other words, the conditional probability $P(M|K,H)$ for the hypothesis "H: There are x unknown contributors" is just $Q(2x, U)$.

Remark 4. If R_1 and R_2 are parents of the suspect S, i.e. $k_1^{SR_1} = k_1^{SR_2} = 1/2$, $k_0^{SR_1} = k_0^{SR_2} = 0$, then $P(M|K,H) = [I_M(r_{11}) I_M(r_{21}) Q(2x-2, U \setminus \{r_{11}\} \cup \{r_{21}\}) + I_M(r_{11}) I_M(r_{22}) Q(2x-2, U \setminus \{r_{11}\} \cup \{r_{22}\}) + I_M(r_{12}) I_M(r_{21}) Q(2x-2, U \setminus \{r_{12}\} \cup \{r_{21}\}) + I_M(r_{12}) I_M(r_{22}) Q(2x-2, U \setminus \{r_{12}\} \cup \{r_{22}\})]/4$. It could happen that $P(M|K,H) = 0$ and so the suspect is excluded as being a contributor of the mixed stain when the suspect's parents are typed.

Example. In order to investigate whether the genotypes of the relatives of the suspect provide useful information in interpreting DNA mixtures when the suspect's genotype is unavailable, we reanalyze the rape case reported in [12], in which the suspect was typed. For illustration, we assume that the suspect was not typed and instead, some close relatives of the suspect, such as the mother, father, and/or paternal grandfather were typed. Their genotypes as well as the mixture, the victim's genotype (V) at three loci D3S1358, vWA, and FGA are listed in Table 1. Regarding to this case, the prosecution and defense propositions are proposed as follows: H_p: The victim and the suspect were contributors of the mixed stain; H_d: The victim and one unknown were contributors.

Table 1. Alleles and genotypes for the mixture M, victim V, maternal relative R_1 (mother), and paternal relative R_2 (father or grandfather) of the suspect for a rape case in Hong Kong. The suspect S was later typed for comparison purpose.

Locus	M	V	R_1	R_2	S
D3S1358	14		14		14
	15	15			
	17			17	17
	18	18			
				19	
vWA	16		16	16	16
	18	18		18	
FGA	20	20	20		
	24	24			
	25		25	25	25
				26	

For simplicity, we consider the following scenarios about which relative(s) of the suspect is(are) typed.

Scenario 1: suspect untyped but his mother typed;
Scenario 2: suspect untyped but his mother and paternal grandfather typed;
Scenario 3: suspect untyped but his mother and father typed;
Scenario 4: suspect typed.

Scenario 4 is introduced only for the comparison of the effect of the biological information of the suspect's relatives on interpreting DNA mixtures. As we can see, scenario 2 provides more biological information about the suspect than scenario 1, and so about scenarios 3 and 2, and scenarios 4 and 3. To illustrate the application of our developed formulae in the calculation of $LR = P(M|K,H_p)/P(M|K,H_d)$, we take locus FGA as an example. At locus FGA, we see from Table 1 that $M = \{20, 24, 25\}$ and $V = 20/24$. Under hypothesis H_d, the only known contributor to the mixed stain is the victim, so the allele set $U = \{25\}$ and $P(M|K, H_d) = Q(2, \{25\})$ by Remark 3. In scenario 1 in which the suspect is not typed, a maternal relative R_1 (mother) is typed instead and is given as 20/25, and so using results in Remark 2 with $U = \{25\}$, $x = 1$, $r_{11} = 20$ and $r_{12} = 25$, we obtain $P(M|K, H_p) = [Q(1, \{25\}) + Q(1, \phi)]/2$. In scenario 2 having genotypes of the mother and paternal grandfather typed (Table 1), using Eq. (5) in Theorem 1 with $U = \{25\}$, $x = 1$, $r_{11} = 20, r_{12} = 25$, $r_{21} = 25$, $r_{22} = 26$, $k_0^{SR_1} = 0$, $k_1^{SR_1} = k_0^{SR_2} = 1/2$, and $k_1^{SR_2} = 1/4$, we obtain $P(M|K, H_p) = [Q(1, \{25\}) + Q(1, \phi) + 1]/4$. In scenario 3 with genotypes of the mother and father, using results in Remark 4 with $U = \{25\}$, $x = 1$, $r_{11} = 20, r_{12} = 25$, $r_{21} = 25$ and $r_{22} = 26$, we obtain $P(M|K, H_p) = Q(0, \phi)/2 = 1/2$. In scenario 4 where the suspect is later available for typing, using result in Remark 3 with $U = \phi$ and $x = 0$, we obtain $P(M|K, H_p) = Q(0, \phi) = 1$. So the likelihood ratios for these four scenarios are obtained accordingly. Table 2 shows the likelihood ratios under these four scenarios for the three loci as well as for the overall one. The overall LRs corresponding to scenarios 1, 2, 3 and 4 are in the ratios of 1:5.5:15.9:63.5.

Table 2. Likelihood ratios of "H_p : the victim and the suspect were contributors of the mixed stain" versus "H_d : the victim and one unknown were contributors" in example about the rape case in Hong Kong.

Locus	Scenario			
	1	2	3	4
D3S1358	15.15	23.42	31.70	63.40
vWA	4.29	8.99	13.70	13.70
FGA	3.80	6.42	9.04	18.07
Overall	247	1,351	3,923	5,694

We expect a decrease in the likelihood ratio for the case of missing suspect compared to that for the case of available suspect. From the LRs listed in Table 2, we understand that two relatives provide more genetic information than one relative, and the closer the relative, the more information he/she provides for forensic DNA analysis. In the case having no genotype from the suspect, it is worthy to type one or two close relatives of the suspect for getting relevant genetic information in the evaluation of DNA mixtures.

Lastly, suppose we consider the alternative proposition as follows: H_d': two unknowns were contributors of the mixed stain. Table 3 lists the corresponding likelihood ratios which also display the same pattern as we described above.

III. Conclusion

A general formula (Eq. (5), Theorem 1) is obtained for the evaluation of DNA mixtures when the suspect is not available and two of his/her relatives are typed instead. The two relatives are biologically unrelated, and so one of them can be a maternal relative and the other a paternal relative. Although in principle our theorem can be applied to any two unrelated relatives, we recommend typing close relatives since it provides more accurate genetic information about the suspect and thus can raise the power of the DNA test. It is noted that the relatives can provide genetic information favorable or unfavorable to the suspect. Currently, we are working towards relaxing the restriction that the two relatives need to be biologically unrelated and we hope to report the findings in the future. Moreover, we investigate the evaluation of DNA mixtures when the suspect is missing and his/her two relatives are typed, and one unknown contributor is related to a typed person or two unknowns are biologically related. The corresponding formulae for calculating the likelihood ratio are reported.

Table 3. Likelihood ratios of "H_p : the victim and the suspect were contributors of the mixed stain" versus "H_d' : two unknowns were contributors" in example.

Locus	Scenario			
	1	2	3	4
D3S1358	68	105	143	285
vWA	37	78	119	119
FGA	75	127	180	359
Overall	191,468	1,049,574	3,047,021	12,188,087

Acknowledgment

This work was partially supported by the Scientific Foundation of Fudan University.

References

[1] D.J. Balding, Estimating products in forensic identification using DNA profiles, J. Am. Stat. Assoc. 90 (1995) 839–844.

[2] L.A. Foreman, A.F.M. Smith, I.W. Evett, Baysian analysis of DNA profiling data in forensic identification applications (with discussion), J. R. Statist. Soc. A 160 (1997) 429–469.

[3] National Research Council Commitee on DNA Forensic Science, The Evaluation of Forensic DNA Evidence, National Academy Press, Washington, DC, 1996.

[4] B.S. Weir, C.M. Starling, L.I. Stowell, K.A.J. Walsh, J. Buckleton, Interpreting DNA mixtures, J. Forensic Sci. 42 (1997) 213–222.

[5] N. Fukshansky, W. Bär, Interpreting forensic DNA evidence on the basis of hypotheses testing, Int. J. Legal Med. 111 (1998) 62–66.

[6] N. Fukshansky, W. Bär, Biostatistical evaluation of mixed stains with contributors of different ethnic origin, Int. J. Legal Med. 112 (1999) 383–387.

[7] J.M. Curran, C.M. Triggs, J. Buckleton, B.S. Weir, Interpreting DNA mixtures in structured populations, J. Forensic Sci. 44 (1999) 987–995.

[8] J.F.Y. Brookfield, The effect of relatives on the likelihood ratio associated with DNA profile evidence in criminal cases, J. Forensic Sci. Soc. 34 (1994) 193–197.

[9] I.W. Evett, Evaluating DNA profiles in case where the defense is "It is my brother", J. Forensic Sci. Soc. 32 (1992) 5–14.

[10] T.R. Belin, D.W. Gjertson, M.Y. Hu, Summarizing DNA evidence when relatives are possible suspects, J. Am. Stat. Assoc. 92 (1997) 706–716.

[11] N. Fukshansky, W. Bär, Biostatistics for mixed stains: the case of tested relatives of a non-tested suspect, Int. J. Legal Med. 114 (2000) 78–82.

[12] Y.Q. Hu, W.K. Fung, Interpreting DNA mixtures with the presence of relatives, Int. J. Legal Med. 117 (2003) 39–45.

[13] Y.Q. Hu, W.K. Fung, Evaluation of DNA mixtures involving two pairs of relatives, Int. J. Legal Med. 119 (2005) 251–259.

[14] C.W. Cotterman, Relatives and human genetic analysis, Sci. Month. 53 (1941) 227–234.