RAIRO-Theor. Inf. Appl. **42** (2008) 323–333
DOI: 10.1051/ita:2007034

Available online at:
www.rairo-ita.org

PHENOTYPE SPACE AND KINSHIP ASSIGNMENT FOR THE SIMPSON INDEX

Bruce Litow[1] and Dmitry Konovalov[1]

Abstract. We investigate the computational structure of the biological kinship assignment problem by abstracting away all biological details that are irrelevant to computation. The computational structure depends on phenotype space, which we formally define. We illustrate this approach by exhibiting an approximation algorithm for kinship assignment in the case of the Simpson index with *a priori* error bound and running time that is polynomial in the bit size of the population, but exponential in phenotype space size. This algorithm is based on a relaxed version of the assignment problem, where fractional assignments (over the reals) are permitted.

Mathematics Subject Classification. 68X30, 68W25, 92D25.

1. Introduction

Kinship assignment

Kinship assignment is a fundamental problem in population biology. This is reflected in active research into this problem. Our work has been motivated by several studies, *e.g.*, [1,2,4,8,11,17,20]. Some of our research has been reported in [12–15]. Over a population sample of n individuals drawn from a single species one wants to assign them to kinship groups. A kinship group (hereafter: group) is admissible if all of its members could possibly be related. The specification of relatedness will depend on the particular application. For example, relatedness might mean that individuals are offspring of the same pair of parents, or a more complicated meaning might be involved, *e.g.*, cousins, etc. The crucial point is that an individual x is excluded from a group if there exists a member y such that

Keywords and phrases. Population biology, kinship assignment complexity, Tarski algebra, phenotype space.

[1] Bioinformatics Applications Research Centre, School of MP & IT, James Cook University, Townsville, Qld. 4811, Australia; bruce.litow@jcu.edu.au; dmitry.konovalov@jcu.edu.au

Article published by EDP Sciences

it is impossible for x and y to be related. The rules for determining admissibility are typically derived from heredity models. However, in most cases there will be very many possible ways to achieve admissible groups. Thus, a measure of goodness for any given assignment is to some extent arbitrary when we do not know true parentage. We show that the computational structure of kinship assignment depends critically on the rules for forming admissible groups, and that these rules can be presented abstractly, independent of specific biological content. That is, we present the general form of the constraints on groups without recourse to explicit genetic assumptions. As an example of this general view, we present a kinship assignment algorithm $\mathcal{A}$ for a goodness measure known as the Simpson index that generates an admissible set of groups (called an admissible partition of the population) whose Simpson index is within a small distance of the optimal Simpson index. Our main result concerning algorithm $\mathcal{A}$ is

Theorem 1. *Algorithm $\mathcal{A}$ has running time dominated by $(\log(n) + M)^{M^{O(1)}}$, where $M = H \cdot m$, H is the number of phenotypes, and m is the size of phenotype space. M is independent of n. The Simpson index of the admissible partition returned by algorithm $\mathcal{A}$ is within $3M/n^2$ of the optimal Simpson index.*

In Section 4 we give an improved version of algorithm $\mathcal{A}$.
Ignoring the factor 3 in the error bound of the theorem, if we work with a population size $n > 2^{k/2} \cdot \sqrt{M}$, algorithm $\mathcal{A}$ would deliver an assignment with a Simpson index within $1/2^k$ of optimal. For example, for a population of 10 000, and a Simpson index within 1% of optimal, so $k = 7$, M must be bounded above by 2^{20}. Assuming Mendelian heredity and independence of loci, kinship assignment within the 1% tolerance can be achieved for 5 loci, with 4 alleles per locus. Of course, directly running algorithm $\mathcal{A}$ for such M is impossible.

Markers and phenotypes

A genetic marker can be any number of things, from individual bases to alleles corresponding usually to genes. Although our approach does not depend on specification of what constitutes a marker, it may make things more concrete if we give an instance. Consider a diploid species, and some number r of independent (unlinked) loci on its genome. An individual will have the string of codominant markers $(u_1, v_1), \ldots, (u_r, v_r)$, where at locus i, (u_i, v_i) is a pair (one from each parent) of alleles (gene variants). If there are h_i alleles at locus i, then there will be $h_i \cdot (h_i + 1)/2 = h_i \cdot (h_i - 1)/2 + h_i$ unordered pairs of alleles (note: $u_i = v_i$ is possible). Thus, there are $H = \prod_{i=1}^{r} h_i \cdot (h_i + 1)/2$ strings of markers. For application to kinship assignment, an individual is completely characterized by its string of markers, which we call its phenotype. We do not enter into the complex issue of how to define the concept phenotype from the standpoint of evolutionary biology.

Since an individual reduces to its phenotype, a group can be viewed as a set of phenotypes. If Mendelian heredity and independence of loci are assumed, then within an admissible group at each locus there are at most 4 possible markers over

all members. Thus, there are at most H^4 admissible groups (an overestimate for biologically significant cases). Notice that this number is independent of n. A club is defined to be a set of phenotypes corresponding to an admissible group. The phenotype space is defined to be the set of clubs. We see that kinship assignment amounts to partitioning the population into clubs. It is only in applying a goodness measure to partitions that n enters into consideration. In the literature there is some confusion about which space is being partitioned in connection with kinship assignment. The space of partitions of individuals is irrelevant to the problem. It is the partition of phenotype space into clubs that matters. The number of indviduals in each club is determined by the goodness measure, not by the genetics and heredity data. This will be made quite clear in our example using the Simpson index.

We give a brute force assignment algorithm, using phenotype space which already shows that the set of partitions of the population does not enter into the computation. List all partitions of phenotype space, throwing away all those containing inadmissible groups. Now, for each surviving partition, list all possible apportionments of individuals to each partition, based on their phenotypes. This method has a running time dominated by $P \cdot n_1^{b_1} \cdots n_H^{b_H}$, where P is the number of partitions of phenotype space, and b_i is the number of clubs containing phenotype i. Notice that $n = n_1 + \cdots + n_H$, and we assume $n_i > 0$ for all phenotypes i. For a simple comparison, assume that $n_i = n/H$ and $b_i = m/H$. It is easy to check that if $n > H^{2H}$, then the brute force running time dominates that of $\mathcal{A}$. More important, the running times of both the brute force method and algorithm $\mathcal{A}$ do not have n in the exponent.

Simpson index

The Simpson index is well known in population biology, but we review some of its basic features. The Simpson index of a partition $S_1, \ldots, S_r$ of a set S of size n is defined to be

$$\frac{1}{n(n-1)} \sum_{i=1}^{r} |S_i| \cdot (|S_i| - 1).$$

Note that the Simpson index assumes values in the closed interval $[0, 1]$, and indeed it is usually regarded as a probability of relatedness. Each group S_i is regarded as consisting of related individuals. The probability that two individuals drawn independently and uniformly from S are related, *i.e.*, in the same group S_i is

$$\frac{|S_i|}{n} \cdot \frac{|S_i| - 1}{n - 1}.$$

Thus, the probability that two individuals are in the same group is the sum over the above probability for each group (disjoint events), but this is the Simpson index. Assuming that only admissible partitions are considered, one partition is a better relatedness hypothesis than another if its index is larger. In effect, one seeks to maximize legitimate probability of relatedness. Of course there may be

many admissible partitions with maximum index. We use a variant of the Simpson index, denoted by SI. SI is defined to be $\sum_{i=1}^{r} |S_i|^2$. It is easy to check that SI is maximized iff the corresponding Simpson index is maximized. It is also true that SI for two partitions are equal iff their Simspon indices are equal.

We regard two partitions of S, $S_1, \ldots, S_r$ and $T_1, \ldots, T_s$ as different if $r \neq s$, or for $r = s$, $|S_1|, \ldots, |S_r|$ is not a permutation of $|T_1|, \ldots, |T_r|$. It is clear that different partitions in this sense can have the same Simpson index. The next result shows that the Simpson index falls far short of uniquely identifying partitions.

Theorem 2. *Let $L = a_1, \ldots, a_r$ be a list of complex numbers. The list $L' = b_1, \ldots, b_r$, where $b_k = \sum_{i=1}^{r} a_i^k$ determines L up to permutation, and no proper sublist of L' can do this.*

We omit a proof as this is a well known result that can be obtained from the classical theory of linear equations, based on explicit formulae for the coefficients of the characteristic polynomial of a matrix in terms of the traces of its powers. See [6].

2. Kinship assignment in terms of phenotype space

Recall that a group is admissible if its members could possibly be related according to heredity rules, given their phenotypes. A club is an admissible group seen only in terms of the phenotypes of its members. A kinship assignment for a population of n individuals is defined to be a mapping of individuals to clubs under the requirement that the phenotype of an individual is an element of the club to which it is mapped. For a given goodness measure μ over partitions P of the set of individuals, an optimal kinship assignment is a partition P that corresponds to a partition of phenotype space, *i.e.*, its groups correspond to clubs, and such that $\mu(P)$ is optimal. If μ is SI, then $\mu(P)$ is maximal. We proceed to formalize this for our algorithm $\mathcal{A}$.

An instance of kinship assignment $S, \Lambda, C, \mathcal{P}(S), \iota, \mu$ consists of:

- S is a set of n individuals (population sample).
- $\Lambda = \{\lambda_1, \ldots, \lambda_H\}$ is a set of labels (phenotypes).
- $C = \{C_1, \ldots, C_m\}$, where $C_i \subseteq \Lambda$ is a set of clubs. C is the phenotype space. We point out that typically one imposes the constraint that C should be closed in the sense that if $X \subseteq C_i$, $|X| > 2$, then $X = C_j$ for some j. That is, a subset of at least size 3 of a club is a club.
- ι: $S \to \Lambda$ is the mapping that assigns an individual to its phenotype. We extend ι to subsets S' of S via $\iota(S') = \bigcup_{a \in S'} \iota(a)$.
- $\mathcal{P}(S)$ is the set of partitions over S. A partition of S is said to be admissible if its groups correspond to clubs.
- μ: $\mathcal{P}(S) \to \mathbb{R}$ subject to $\mu(A) \geq 0$ and if $\mu(A) > 0$, then $\iota(S_i)$ is a club for $i = 1, \ldots, r$, where $A = \{S_1, \ldots, S_r\}$. Note that μ is the 'goodness' measure on partitions and only admissible partitions can receive positive measure.

- A kinship assignment solution is an admissible partition A such that $\mu(A) \geq \mu(B)$ for any other partition B.

The details of how phenotype space, and phenotypes are constructed, which is a problem in biology, has been abstracted away in this scheme. Of course, these details will determine the number of phenotypes and the clubs. Once these have been determined, the biology is irrelevant to the computation.

3. An algorithm for kinship assignment

This section is devoted to proving Theorem 1. Algorithm $\mathcal{A}$ is framed in Tarski algebra, which makes possible streamlined descriptions of otherwise quite complicated statements in real algebraic geometry. See [3] for a current and comprehensive discussion of this research field.

Tarski algebra

The first order theory of the field of reals is often called Tarski algebra (TA) in honor of Alfred Tarski who gave a quantifier elimination decision method for it in 1931. See [18,19]. Details about Tarski algebra can be found in [3,10,16]. Background notions from logic can be found in [7]. We require just a few summary facts about TA. A TA prenex formula $B(y_1, \ldots, y_j)$ in the free variables $y_1, \ldots, y_j$ has the form $Q_1 x_1 \cdots Q_k x_k A(x_1, \ldots, x_k, y_1, \ldots, y_j)$, where $Q_1, \ldots, Q_k$ are either $\forall$ or $\exists$ and $A(x_1, \ldots, x_k, y_1, \ldots, y_j)$ is a quantifier free Boolean in just $\vee, \wedge$ over atomic formulae. An atomic formula has the form $P = 0, P > 0, P < 0$, where P is an integer coefficient polynomial in $x_1, \ldots, y_j$. A term is an arithmetic expression built up from $0, 1, +, \times$. Using conventional binary notation an integer m can be represented by a term of size $O(\log^2 m)$. However, admitting brackets into the language, by Hörner's method, if $m = a_0 + 2 \cdot a_1 + \cdots + a_k \cdot 2^k$, then one can represent M as $a_0 + 2 \cdot (a_1 + 2 \cdot (\ldots 2 \cdot (a_{k-1} + 2 \cdot a_k) \ldots)$, which has $O(\log m)$ size. A TS formula B without free variables is called a sentence. TA formulae $B(y_1, \ldots, y_k), C(y_1, \ldots, y_k)$ are said to be equivalent iff

$$\forall y_1, \ldots, y_k B(y_1, \ldots, y_j) \Leftrightarrow C(y_1, \ldots, y_k)$$

is a true sentence interpreted over the real numbers. The main result that we need concerning TA is

Theorem 3. *A prenex TA formula $B(y_1, \ldots, y_j)$ can be converted to an equivalent quantifier free formula $C(y_1, \ldots, y_k)$ in time $b^{j^{a+O(1)}}$, where b is the size of $B(y_1, \ldots, y_j)$, and a is the number of quantifier alternations in the formula.*

Theorem 3 is a refinement of Tarski's result that develops out of work by Collins and Grigoriev. For some historical perspective, see [5,9]. See [3] for details and sharper results.

If we have a family of formulae for which a is fixed, then each formula can be converted to an equivalent quantifier free formula in EXP-POLY time.

The next lemma will be used in the development of algorithm $\mathcal{A}$. An interval can be empty, a point, closed, open, or half open.

Lemma 1. *Let $F(z)$ be a TA formula in the single free variable z. The set of z such that $F(z)$ is a finite union of intervals.*

Proof. By Theorem 3, $F(z)$ is equivalent to a Boolean $\bigvee_p \bigwedge_q F_{p,q}(z)$ where $F_{p,q}(z)$ is an atomic formula. It is clear that the set of z such that $F_{p,q}(z)$ is an interval, and $F_p = \bigwedge_q F_{p,q}$ is a finite intersection of intervals, which is again an interval. □

The relaxed algorithm

We first look at a relaxed version of kinship assignment. Relaxed means that an individual can be fractionally assigned to a club. If an individual's phenotype is a member of clubs $C_{i_1}, \ldots, C_{i_k}$, and $a_1, \ldots, a_k$ are nonnegative reals such that $a_1 + \cdots + a_k = 1$, then we can assign a_i of the individual to C_i. We formalize this.

Let n_i be the number of elements of S mapped by ι to λ_i, for $i = 1, \ldots, H$. Note that $\sum_{i=1}^{H} n_i = n$. Define $A_{i,j}$ by

$$A_{i,j} = \begin{cases} 0 \text{ if } \lambda_i \notin C_j \\ 1 \text{ if } \lambda_i \in C_j. \end{cases}$$

The array $A_{i,j}$ keeps track of phenotype membership in clubs. For $i = 1, \ldots, H$, $j = 1, \ldots, m$ introduce the real variables $x_{i,j}$. We interpret $A_{i,j} \cdot x_{i,j}$ as the fraction of the population having phenotype λ_i that is assigned to club C_j. We have relaxed kinship assignment by allowing for fractions that do not correspond to a whole number of individuals. We do require that $x_{i,j} \geq 0$. The fact that all individuals having phenotype λ_i are divided among all clubs is expressed by $U_i = 0$, where

$$U_i = \sum_{j=1}^{m} x_{i,j} - 1.$$

The overall partition constraint is then $U = 0$, where

$$U = \sum_{i=1}^{H} U_i^2.$$

Notice that this forces $U_i = 0$ and depends on a basic property of $\mathbb{R}$, namely that for $x \in \mathbb{R}$, $x^2 \geq 0$. If $U = 0$ we will say that the $x_{i,j}$ form a relaxed admissible partition. For $j = 1, \ldots, m$ define

$$V_j = \sum_{i=1}^{H} A_{i,j} \cdot n_i \cdot x_{i,j} \ .$$

Each summand $A_{i,j} \cdot n_i \cdot x_{i,j}$ is the possibly fractional number of individuals having phenotype λ_i that are assigned to club C_j, so V_j is the fractional number of individuals assigned to C_j, *i.e.*, the size of a group in a relaxed partition. The relaxed SI is then

$$V = \sum_{j=1}^{m} V_j^2.$$

Since the number of partitions is finite for kinship assignment it is clear that there is a maximum value of the Simpson index over all admissible partitions. We now show that the maximum SI exists over all relaxed admissible partitions. We adopt the convention that $\mathbf{x}$ denotes the list of all $x_{i,j}$ variables. The TA formula $F_1(y)$ in the single free variable y given by

$$\exists\, \mathbf{x} \bigwedge_{i,j} x_{i,j} \geq 0 \wedge U = 0 \wedge y = V,$$

expresses that fact that there exists a relaxed admissible partition whose SI is y. The domain $\mathcal{F}$ of relaxed admissible partitions is obviously compact, since it is determined by $\bigwedge_{i,j} x_{i,j} \geq 0 \wedge U = 0$, and V is continuous over $\mathcal{F}$, so by standard analysis, V assumes its maximum over $\mathcal{F}$. From this observation the TA formula $F_2(y)$ given by $\forall\, z\ F_1(y) \wedge F_1(z) \Rightarrow z \leq y$ expresses the fact that y is the maximum SI for any relaxed admissible partition. It is interesting to note that Theorem 3 immediately implies that y is a real algebraic number. Also note that y is an upper bound on the optimal SI for the kinship assignment problem.

Let y_+ and y_- be the maximum and minimum SI, respectively over all relaxed admissible partitions. Notice that the admissible partition in which each individual is assigned to a different group has Simpson index of 0. The corresponding SI is n. Because we are mapping to clubs it is likely that $y_- > n$. Given positive integer k, we show how to compute the integer D such that $|y_+ - D/2^k| < 1/2^k$. The same method can be applied to y_-. Let $G_{<,u}$ and $G_{=,u}$ be the sentences $\forall\ y\ F_2(y) \Rightarrow u < y$ and $\forall\ y\ F_2(y) \Rightarrow u = y$, respectively, where u is an integer term. Proceeding by binary search, starting from the term u for 2^{k-1}, we can in $O(k)$ tests with $O(k^2)$ size terms find D. Each of the $O(k)$ sentences has size $O(M + k^2 + \log n)$ and has $O(1)$ quantifier alternations and $O(H)$ variables. Observe that the occurrence of $\log n$ in this bound is due solely to encoding the n_i as terms. By Theorem 3, we can find D in time $(\log n + k + M)^{M^{O(1)}}$.

We conduct a simple sensitivity analysis for V. It will be convenient to linearly reorder the variables $x_{i,j}$ as $z_1, \ldots, z_M$, where $M = H \cdot m$. We re-index the array $A_{i,j}$ as just A_i, $i = 1, \ldots, M$. Also, redefine n_i for $i = 1, \ldots, M$ to be the number of individuals whose phenotype corresponds to z_i. Assume $0 \leq \delta < 1$, and for all $i = 2, \ldots, M$, $z_i = z_i'$ and $|z_1 - z_1'| < \delta$. We have

$$|V(\mathbf{z}) - V(\mathbf{z}')| \leq A_1 \cdot n_1 \cdot (2 \cdot z_1 \cdot \delta + \delta^2).$$

Since $A_1 \cdot n_1 \leq n$, and $z_1 \leq 1$, this yields

$$|V(\mathbf{z}) - V(\mathbf{z}')| < 2n \cdot (\delta + \delta^2).$$

If we require that $2n \cdot (\delta + \delta^2) < 1/2^k$, then choosing $\delta < 1/(2n \cdot 2^{k+1})$ certainly satisfies the required inequality. Thus, δ need only be an $O(k + \log n)$ bit rational.

The basis of algorithm $\mathcal{A}$ is a simple deflation technique based on Theorem 3. Each of the M variables is replaced by a suitable rational, and successively all quantifiers are eliminated. The only care required is in keeping track of cumulative errors introduced by replacing variables with rationals. We describe this technique for the relaxed problem.

Let k be a positive integer and choose integer $g = \lceil \log 2n \rceil + \lceil \log M \rceil + k$. Let D be the integer such that $0 \leq y_+ - D/2^k < 1/2^k$ and $D/2^k \geq y_-$. We show how to compute integers $d_1, \ldots, d_M$ such that $\mathbf{z}$ given by $z_i = d_i/2^k$ is a relaxed admissible partition, and its SI is within $1/2^k$ of y_+. Let $F_{3,1}(z_1)$ be the TA formula (compare with $F_1(z_1)$)

$$\exists\, z_2, \ldots, z_M \bigwedge_i z_i \geq 0 \wedge U = 0 \wedge y_+ = V.$$

By the restrictions on D and continuity of V over the domain $\mathcal{F}$, $F_{3,1}(z_1)$ for at least one real z_1. By Theorem 3, $F_{3,1}(z_1)$ can be converted in time $(k + \log(n) + M)^{M^{O(1)}}$ into a Boolean as in Lemma 1, and from this Boolean, d_1 can be computed in time polynomial in the time bound (all atomic formulae have sizes bounded above by the time bound). To see this, by Lemma 1, the solution set of this Boolean is a finite union of intervals. This means that we can take the fraction, $d_1/2^g$, to be an approximation of either an endpoint or midpoint of an interval which can be derived from polynomial time methods for approximating zeros of polynomials.

The process is iterated. $F_{3,2}(z_2)$ is the TA formula

$$\exists\, z_3, \ldots, z_M \bigwedge_{i>1} z_i \geq 0 \wedge U = 0 \wedge D/2^k = V,$$

where $d_1/2^k$ substitutes for z_1. From the sensitivity analysis with $\delta = 1/2^g$, and choice of g, the distance between y_+ and the SI associated to the partition using $d_1/2^g$ rather than the z_1 being approximated is less than $1/(M \cdot 2^k)$. Now $d_2/2^g$ is determined, and $F_{3,3}(z_3)$ is use, with z_1 and z_2 being replaced by $d_1/2^g$ and $d_2/2^g$, respectively. The distance between y_+ and the new SI is less than $2/(M \cdot 2^k)$. After M stages we have a distance of less than $1/2^k$, as required. The time bound for this process is straightforward by Theorem 3: $M \cdot k \cdot (\log n + k + M)^{M^{O(1)}} = (\log n + k + M)^{M^{O(1)}}$.

ALGORITHM $\mathcal{A}$

It is not difficult to convert a relaxed admissible partition that achieves SI close to y_+ into an admissible partition with a bounded distance from the maximal SI. The first step of $\mathcal{A}$ has already been described, so we have a relaxed admissible partition $\mathbf{x}$ such that its SI is within $1/2^k$ of y_+. We describe the second, conversion step.

Recall that n_i is the number of individuals with phenotype i. We revert to regarding $\mathbf{x}$ as the list $x_{i,j}$. Let b_i be the number of clubs in which phenotype i is a member. We can write $x_{i,j} = a_{i,j}/n_i + \delta_{i,j}$, where $a_{i,j}$ is a nonnegative integer, and $0 \le \delta_{i,j} < 1/n_i$. Since $\mathbf{x}$ is admissible ($U_i = 1$ for each i),

$$\sum_{j=1}^{m} (A_{i,j} \cdot a_{i,j}/n_i + A_{i,j} \cdot \delta_{i,j}) = 1.$$

It follows from this that

$$\sum_{j=1}^{m} A_{i,j} \cdot \delta_{i,j} = f_i/n_i,$$

where $f_i < b_i$ is a nonnegative integer. Notice here that $\sum_{j=1}^{m} A_{i,j} = b_i$. Let $j_1, \ldots, j_{b_i}$ be the indices for which $A_{i,j} = 1$. By re-indexing, assume that for $j = j_1, \ldots, j_{b_i}$, the $a_{i,j}$ are in descending order by size. For $j = j_1, \ldots, j_{f_i}$, assign $a_{i,j} + 1$ individuals to club C_j, and for $j = j_{f_i+1}, \ldots, j_{b_i}$, assign $a_{i,j}$. Observe that

$$\sum_{j=j_1,\ldots,j_{b_i}} (a_{i,j} + 1) = n_i,$$

so that we obtain an admissible partition in this way. Call this partition $\mathbf{y}$, noting that each $y_{i,j}$ is a nonnegative integer.

We measure the distance between the SI for the relaxed admissible partition $\mathbf{x}$ and the admissible partition $\mathbf{y}$ we have just constructed. We now have that $|x_{i,j} - y_{i,j}| < 1/n_i$. From this we get, Recalling the sensitivity analysis,

$$|V(\mathbf{x}) - V(\mathbf{y})| \le 2 \sum_i n_i \cdot \sum_j A_{i,j} \cdot x_{i,j}/n_i + \sum_i n_i \cdot \sum_j A_{i,j}/n_i^2.$$

The RHS is bounded above by $2\sum_i b_i + \sum_i b_i/n_i$. It is clear that $3M$ is a crude upper bound for this sum because $\sum_i b_i \le H \cdot m = M$. Reverting to the Simpson index, the distance becomes $3M/(n(n-1)$. This completes the proof of Theorem 1.

4. AN IMPROVEMENT TO ALGORITHM $\mathcal{A}$

We make an improvement on algorithm $\mathcal{A}$ by reducing its running time to $(M + k + \log n)^{H^{O(1)}}$ with error $1/2^k$. In general we expect that $H = o(\sqrt{M})$. We begin by noting that the relaxed kinship assignment problem, ignoring the $x_{i,j} \ge 0$

constraints, can be treated by the Lagrange multiplier method, *i.e.*, we can inspect the system of M equations $\partial_{i,j}V - \lambda\partial_{i,j}U = 0$, where $\partial_{i,j}$ abbreviates $\frac{\partial}{\partial x_{i,j}}$. Since both V and U are quadratic forms, both sides of each equation are linear in all variables. We show that in H stages, we can eliminate all of the $x_{i,j}$ variables at the cost of introducing new variables $\lambda_1, \ldots, \lambda_H$. The $x_{i,j}$ will be expressed as rational functions in these λ variables.

Consider eliminating all variables belonging to phenotype 1, so $x_{1,1}, \ldots, x_{1,b}$, where double subscripting of the club index has been dropped to simplify notation. We have the H equations

$$(1 - a_i \cdot \lambda)x_{i,j} = \lambda[x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{i,p}] - [\ldots, x_{i',j}, \ldots].$$

Here $a_i \neq 0$, $[x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{i,p}]$ is a linear form in the indicated variables, coming from U_i, and $[\ldots, x_{i',j}, \ldots]$ is a linear form coming from V_j. Thus, $x_{i,1}$ is expressible as a linear form A in $x_{i,2}, \ldots, x_{i,p}$ and some variables belonging to phenotypes $i' \neq i$. Examining the equation for $x_{i,2}$, and substituting A in the expression for $x_{i,2}$ on the LHS, we get a nontrivial linear form for $x_{i,2}$ in $x_{i,3}, \ldots, x_{i,p}$ and some variables belonging to phenotypes $i' \neq i$. In the same way we can eliminate all of the variables through $x_{i,p}$. The stage for each phenotype introduces a new λ variable, so after all $x_{i,j}$ have been eliminated, we still have H of the λ variables. We re-impose the nonnegativity constraints for all $x_{i,j}$ in terms of their forms in the λ variables. Notice that these forms are rational functions and may have terms of degree roughly $\sum_i b_i \leq M$. Thus we obtain a TA formula of size $O(M) + \log n)$ in H variables, with $O(1)$ alternations of quantifiers. By algorithm $\mathcal{A}$, we can obtain a relaxed admissible partition within $1/2^k$ of y_+ in time $(k + M + \log n)^{H^{O(1)}}$.

References

[1] A. Almudevar, A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theor. Popul. Biol.* **63** (2003) 63–75.

[2] A. Almudevar and C. Field, Estimation of single-generation sibling relationships based on dna markers. *J. Agr. Biol. Envir. St.* **4** (1999) 136–165.

[3] S. Basu, R. Pollack and M-F. Roy, *Algorithms in Real Algebraic Geometry*. Springer (2005).

[4] T.Y. Berger-Wolf, B. DasGupta, W. Chaovalitwongse and M. Ashley, Combinatorial reconstructions of sibling relationships. In *6th International Symposium on Computational Biology and Genome Informatics (CBGI)*, Salt Lake City, Utah (2005) 1252–1255.

[5] G. Collins, Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In *Automata theory and formal languages*. Springer (1975) 134–183.

[6] L. Csanky, Fast parallel matrix inversion algorithms. In *16th IEEE FOCS* (1975) 11–12.

[7] H. Ebbinghaus, J. Flum and W. Thomas, *Mathematical Logic*. Springer (1984).

[8] K.F. Goodnight and D.C. Queller. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Mol. Ecol.* **8** (1999) 1231–1234.

[9] D.Yu. Grigoriev, Complexity of deciding Tarski algebra. *J. Symb. Comput.* **5** (1988) 65–108.

[10] N. Jacobson, *Basic Algebra, Vol. I*. Freeman, 2nd edn. (1985).

[11] A.G. Jones and W.R. Arden, Methods of parentage analysis in natural populations. *Mol. Ecol.* **12** (2003) 2511–2523.

[12] D.A. Konovalov, Accuracy of four heuristics for the full sibship reconstruction problem in the presence of genotype errors. In *The Fourth Asia Pacific Bioinformatics Conference*, 13-16 Feb, 2006, Taiwan (2006) 7-16.
[13] D.A. Konovalov, C. Manning and M.T. Henshaw, Kingroup: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Mol. Ecol. Notes* **4** (2004) 779–782.
[14] D.A. Konovalov, N. Bajema and B. Litow, Modified simpson o(n3) algorithm for the full sibship reconstruction problem. *Bioinformatics* **21** (2005) 3912–3917.
[15] D.A. Konovalov, B. Litow and N. Bajema, Partition-distance via the assignment problem. *Bioinformatics* **21** (2005) 2463–2468.
[16] B. Mishra, *Algorithmic Algebra*. Springer (1993).
[17] P.T. O'Reilly, C. Herbinger and J.M. Wright, Analysis of parentage determination in atlantic salmon (salmo salar) using microsatellites. *Anim. Genet.* **29** (1998) 363–370.
[18] A. Tarski, Sur les ensembles définissables de nombres réels. *Fundamenta Mathematicae* **17** (1931) 210–239.
[19] A. Tarski, *A decision method for elementary algebra and geometry.* Technical report, Rand Corp. (1948).
[20] J.L. Wang, Sibship reconstruction from genetic data with typing errors. *Genetics* **166** (2004) 1963–1979.

Communicated by C. Choffrut.
Received October 6, 2006. Accepted August 3, 2007.