BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 20 2005, pages 3912–3917
doi:10.1093/bioinformatics/bti642

Genetics and population analysis

Modified SIMPSON $O(n^3)$ algorithm for the full sibship reconstruction problem

Dmitry A. Konovalov*, Nigel Bajema and Bruce Litow

School of Information Technology, James Cook University, Townsville, QLD 4811, Australia

Received on June 20, 2005; revised on August 11, 2005; accepted on August 22, 2005
Advance Access publication August 23, 2005

ABSTRACT

Motivation: The problem of reconstructing full sibling groups from DNA marker data remains a significant challenge for computational biology. A recently published heuristic algorithm based on Mendelian exclusion rules and the Simpson index was successfully applied to the full sibship reconstruction (FSR) problem. However, the so-called SIMPSON algorithm has an unknown complexity measure, questioning its applicability range.

Results: We present a modified version of the SIMPSON (MS) algorithm that behaves as $O(n^3)$ and achieves the same or better accuracy when compared with the original algorithm. Performance of the MS algorithm was tested on a variety of simulated diploid population samples to verify its complexity measure and the significant improvement in efficiency (e.g. 100 times faster than SIMPSON in some cases). It has been shown that, in theory, the SIMPSON algorithm runs in non-polynomial time, significantly limiting its usefulness. It has been also verified via simulation experiments that SIMPSON could run in $O(n^a)$, where $a > 3$.

Availability: Computer code written in Java is available upon request from the first author.

Contact: Dmitry.Konovalov@jcu.edu.au

1 INTRODUCTION

A number of genetics areas, e.g. conservation and behavioral genetics, routinely require knowledge of the pedigree structure for a given population sample. However, obtaining such structural knowledge is not always easy (or even possible) in practice. Another approach is to infer the structure from DNA markers. The markers, especially microsatellite markers (Blouin, 2003), then provide the necessary genotypes that could be used to infer the population structure. One such reconstruction problem is the reconstruction of all full sibling groups within the sample without the availability of parental information.

Currently there are a number of full sibship reconstruction (FSR) algorithms and, in some cases, corresponding readily available software programs. They are the AF (Almudevar and Field, 1999), descending ratio (DR) (Konovalov *et al*., 2004), GRAPH (Beyer and May, 2003), JW (Wang, 2004), SC and FJL (Smith *et al*., 2001), SIMPSON (Butler *et al*., 2004) and TH (Thomas and Hill, 2000) algorithms. Each FSR algorithm varies in how the space of all possible partitions is searched and how the visited partitions are ranked to determine the best according to each algorithm's scoring function. For example, the full joint likelihood (FJL) algorithm [which was referred to as Likelihood in Butler *et al*. (2004)] uses a Markov chain Monte Carlo (MCMC) method to move through the partition space and the FJL as a partition scoring function, whilst AF enumerates all possible maximal feasible sibling groups (MSG) and then uses likelihood scoring function based on multinomial coefficients. Another significant difference between algorithms lies in the application of the Mendelian exclusion rules, where each sib group must strictly obey the Mendelian rules of inheritance, e.g. in the AF, SC and SIMPSON algorithms.

Butler *et al*. (2004) compared the AF, SC, FJL and SIMPSON algorithms and concluded that none of the four algorithms emerged as an overall winner in all respects when accuracy, efficiency and robustness to genotype errors are considered. However their results indicated that in the absence of genotype errors the exclusion based algorithms (AF, SC and SIMPSON) could be significantly more accurate than the likelihood based FJL, especially when a small number of loci are considered. This indication is consistent with the results of Thomas and Hill (2002) who demonstrated that the likelihood scores underestimated the degree of relatedness in the FSR problem. This is understandable since the Mendelian exclusion works exactly even with one locus. Furthermore the accuracy of exclusion methods will also increase rapidly with the number of loci. On the other hand, the likelihood methods work gradually, yielding an accurate prediction only when sufficient genotype information was supplied (large enough number of loci and alleles). Given the comparable accuracy of the three (AF, SC and SIMPSON) algorithms, the AF and SIMPSON algorithms have the further advantage of not requiring population allele frequencies.

Whilst accuracy is a key factor, from the bioinformatics point of view an algorithm should have a known complexity measure (e.g. the big O) and be relatively simple to implement (or the corresponding software should be readily available). The lack of some or all of the above factors arguably contributed to an observation that the FSR (as opposed to parentage inference) algorithms are not used as much as they could be (Pemberton, 2004). With the notable exception of Butler *et al*. (2004) there has been very little reporting of efficiency, and even then only qualitative comparisons for the four (AF, FJL, SC and SIMPSON) algorithms were presented.

In this paper we start addressing the gap in efficiency knowledge for the FSR algorithms on a quantitative rather than qualitative basis, beginning with the accuracy and complexity of the SIMPSON algorithm. The algorithm was chosen because it has a number of above mentioned advantages (simple implementation, comparable or better accuracy and does not require allele frequencies) over the

*To whom correspondence should be addressed.

AF, FJL and SC algorithms. Furthermore, even though the AF and SIMPSON algorithms are comparable in accuracy, the SIMPSON algorithm appeared to be more stable and faster under all circumstances (Butler *et al*., 2004). Another factor that influenced the choice of SIMPSON over AF for this study was that the speed of the AF algorithm was found to depend primarily on the family structure of the population sample and in some cases AF failed to produce answers within a reasonable time, e.g. for 200 unrelated individuals (Butler *et al*., 2004). We note that this conflicts with the study by Almudevar (2001), who successfully applied the AF algorithm to a sample of 781 individuals.

The SIMPSON algorithm uses the Simpson index to compare possible partitions. Given a partition, the index S is defined as

$$S = \frac{1}{n(n-1)} \sum_{j=1}^{r} n_j(n_j - 1), \quad (1)$$

where a population sample of size n is partitioned into r sib groups, with group j containing n_j individuals. In this paper we demonstrate that the SIMPSON algorithm randomly searches a non-polynomial size partition space and therefore its running time could be non-polynomial when maximizing the accuracy. Retaining the Simpson index as the scoring function together with the Mendelian exclusion method we applied the partition space search technique from Konovalov *et al*. (2004) to devise the modified SIMPSON (MS) algorithm, obtaining a running time of $O(n^3)$. Rigorous tests are presented confirming the significant improvement in efficiency while achieving equal or better accuracy of the sibship reconstruction. The MS algorithm was compared with the SIMPSON, GRAPH and DR algorithms in terms of accuracy and found to outperform all others with the only exception of the dataset containing only unrelated individuals, where DR was the best. Unfortunately quantitative comparison with (and between) other FSR algorithms (AF, FJL, JW, SC and TH) was not possible without further study due to the lack of the corresponding data in tabular form.

2 METHODS

2.1 Accuracy

The FSR problem is essentially a partitioning problem in which a given population sample is partitioned into a set of clusters, where each cluster represents a group of full siblings (groups of size one are also permitted). A population sample with a known family structure may be generated by simulation and then given to an algorithm for reconstruction. The structure may consist of various known sib groups and corresponds to the partition of the sample. The accuracy of an algorithm could then be estimated by the distance between the original partition A and the reconstructed partition B, i.e. 100% accuracy corresponds to zero distance,

$$\text{accuracy} = (1 - D(A,B)/n), \quad (2)$$

where $D(A, B)$ is the distance between the partitions A and B, and n is the sample size defined as the number of individuals in the sample. The accuracy is then averaged over a number of trials where each trial involves the reconstruction of a freshly generated population sample with the known structure. The partition-distance $D(A, B)$ equals to the minimum number of individuals in the reconstructed partition B that must be moved to different groups in order to transform B into the original partition A (Almudevar and Field, 1999). The partition-distance can be calculated in $O(r^3)$ via the maximum (Gusfield, 2002) or minimum (Konovalov *et al*., 2005) assignment problem for the bipartite graphs, where r is the maximum number of groups either in A or B.

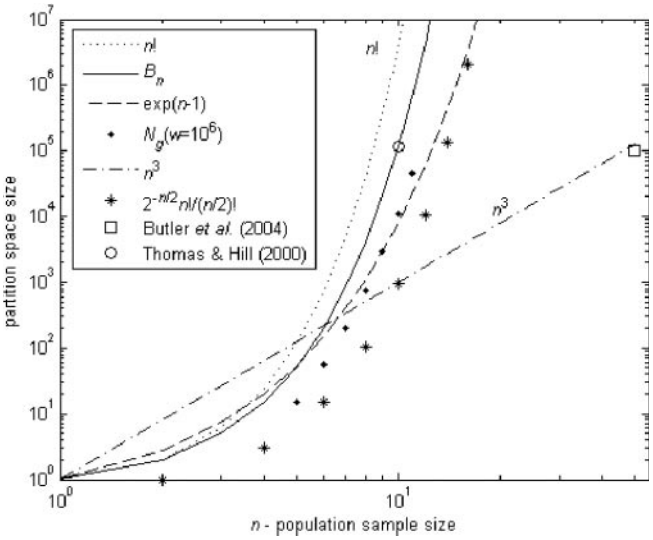

Fig. 1. Partition space size. The Bell numbers (B_n) are plotted using the solid line. The number of iterations ($T_{50} = 10^5$) used with the SIMPSON algorithm (Butler *et al*., 2004) is denoted by a square. Estimation (115 975) of Thomas and Hill (2000) for 10 individuals is denoted by a circle. The N_g line (solid dots) denotes the actual number of unique groups (three or more) that were checked for sibship while reconstructing a population of unrelated individuals ($N_L = 5, N_A = 10$) by setting w to a non-restricting value ($w = 10^6$) for the considered sample sizes.

2.2 Partition space size

Let $P = \{1, 2, \ldots, n\}$ represent a population sample with n individuals (genotypes). The sample can be partitioned $(n;\ a_1,\ a_2, \ldots,\ a_n)' = n!/((1!)^{a_1}a_1!(2!)^{a_2}a_2!\ldots(n!)^{a_n}a_n!)$ number of different ways, where a_j is the number of groups, each having j elements [using notations of Abramowitz and Stegun (1972)]. The total number of partitions each containing m groups is given by the Stirling numbers of the second kind $S(n, m) = \sum(n;\ a_1,\ a_2, \ldots,\ a_n)'$, where summation is done over all possible values of $\{a_1,\ a_2, \ldots,\ a_n\}$ such that $a_1 + 2a_2 + \cdots + na_n = n$ and $a_1 + a_2 + \cdots + a_n = m$. The total number of all possible partitions is then given by the Bell number (Bell, 1934; Weisstein, E.W., from MathWorld—Wolfram Web Resource available at: http://mathworld.wolfram.com/BellNumber.html.) $B_n = \sum_{m=1}^{n} S(n, m)$ obeying the recursive relationship $B_{n+1} = \sum_{k}^{n} = O_n C_k B_k$, where ${}_nC_k = n!/(k!(n-k)!)$ is a binomial coefficient. Given the first $B_0 = 1$ number, B_n can then be calculated, obtaining $B_1 = 1$, $B_2 = 2$, $B_3 = 5$, $\ldots$, $B_{10} = 115\,975$.

To illustrate the non-polynomial growth of the partition space in relation to the sample size, B_n is compared with the cubic, exponential and factorial functions (see Fig. 1). The Bell number increases faster than the exponent but slower than the factorial. Lovasz (1993) showed that B_n has the asymptotic limit $B_n \sim n^{-1/2}[\lambda(n)]^{n+1/2}e^{\lambda(n)-n-1}$, where $\lambda(n)$ is defined implicitly by $\lambda(n)\ \ln[\lambda(n)] = n$.

2.3 SIMPSON algorithm

The SIMPSON algorithm (Butler *et al*., 2004) starts with a partition where each individual is in a group by itself, and then repeats the following loop for a large but fixed number of iterations T_n [$T_{50} = 10^5$ is used in Butler *et al*. (2004), displayed in Fig. 1].

(1) Choose two individuals i and j at random.

(2) Abandon the iteration and start a new one if the two individuals are already in the same group.

(3) Test whether the multilocus genotype of the individual i is compatible with the genotypes of the sib group containing the individual j. If not, abandon iteration and start a new one.

(4) Add individual i to the sib group containing individual j, and calculate the partition's Simpson index.

(5) If the new partition has the highest Simpson index seen so far, store it as the best.

After the prescribed number of iterations, the solution is given by the best partition corresponding to the highest Simpson index.

Essentially the algorithm performs a 'random walk' from one partition to the next by moving one individual at a time and remembering the partition with the highest Simpson index so far. This approach is inefficient since the same partition may be evaluated more than once. In addition the algorithm is a brute force algorithm requiring, at least in theory, a visit to every unique non-prohibited partition. In practice a somewhat arbitrary T_n is used to terminate the search since, at present, it is not known how to choose an optimal T_n. It will be shown in the results section that if an insufficient T_n is used then the reconstruction accuracy deteriorates dramatically. As n grows, T_n must also grow in order to achieve suitable accuracy, and even then the accuracy may plateau without reaching 100%, e.g. see Figures 3a and 3b.

The actual number of all non-prohibited configurations can be calculated numerically using the technique described in the next section, where by setting the w parameter of the MS algorithm to plus infinity the algorithm was transformed into the true brute force algorithm for this problem. Figure 1 shows such actual numbers (denoted N_g) for a population of unrelated individuals and verifies the non-polynomial behavior of the SIMPSON partition space size (and hence the SIMPSON algorithm). It is interesting to note that even a relatively small population size of $n = 11$ requires $N_g \sim 10^5$ unique (non-repeated) sibship checks, while the comparable number ($T_n = 10^5$) of iteration could still be used at $n = 50$ (Butler *et al*., 2004) since N_g represents the statistically improbable scenario when the best partition happened to be the very last being checked.

The lower bound of a brute-force FRS algorithm could be estimated by observing that any two individuals can always form a sib group when parental information is unknown. This means that as a minimum the algorithm must visit all partitions containing at least one group with three individuals while the rest of the groups are of size one and two. The number of such partitions is given by $n!/((1!)^{a_1}a_1!(2!)^{a_2}a_2!(3!)^{a_3}a_3!)$, where $a_1 = \{0, 1\}$ and $a_1 + 2a_2 + 3 = n$. Since $a_2 \sim n/2$ the lower bound of the algorithm increases as $O(2^{-n/2}n!/(n/2)!)$, which is faster than any polynomial power but slower than N_g (Figure 1).

2.4 Genotype distance

In order to find the 'direction' towards the partition with the highest Simpson index we propose a heuristic distance between two individuals X and Y at a locus l as the number of alleles that are not shared,

$$D_l(X,Y) = \max\left(\begin{array}{c}\sum\limits_{\alpha=\{m,p\}}\left(1-\delta_{x_\alpha y_m}\right)\left(1-\delta_{x_\alpha y_p}\right),\\ \sum\limits_{\alpha=\{m,p\}}\left(1-\delta_{y_\alpha x_m}\right)\left(1-\delta_{y_\alpha x_p}\right)\end{array}\right), \quad (3)$$

where the two available alleles of each individual are marked as m and p for maternal and paternal alleles (their order is arbitrary but fixed), δ_{xy} is the Kronecker delta, $\delta_{xy} = 1$ when $x = y$ and zero otherwise. The locus index l is omitted from the definition to simplify the notations, e.g. x_p is used instead of x_p^l. $D_l(X,Y)$ will be referred to as the locus distance. Each summation term in the above equation is zero if the considered allele (x_a or y_a) is present in the second individual.

For the multilocus distance between two genotypes we define

$$D_{XY} = \min\left(D_l(X,Y)\right) \quad (4)$$

and will be referred to as the genotype distance. The minimum rather than the maximum value is used because we are trying to identify the most probable full-sibling pairs without the knowledge of how often a given allele appears in the wider population. Even though the genotype distance varies over the same range of values $\{0, 1, 2\}$ for sibs and unrelated individuals, $\{0, 1\}$ values are more probable for sibs. Other forms of the genotype distance are also considered in a later subsection.

Table 1. Sample workings of the Modified SIMPSON (MS) algorithm with $w = 2$

Step no.	Partitions
	...
	{abc}, {{ab}{c}}
4	+d[a]
5–8[b]	{abcd}[c], {{abc}{d}}, {{abd}{c}}, {{ab}{cd}}, {{ab}{c}{d}}
9	{{abc}{d}}, {{abd}{c}}
	...

[a]Assuming the order of individuals is already determined as per steps (2–4) of the MS algorithm (see text) and it is a,b,c,d, ...
[b]The partitions are displayed in descending order of their Simpson index.
[c]Assuming that this partition cannot be a sib group.

2.5 Modified SIMPSON (MS) algorithm

The algorithm was inspired by the descending ratio (DR) and the exhaustive descent (ED) algorithms (Konovalov *et al*., 2004). While ED builds every possible partition by adding one individual at a time, DR does the same but discards all but one best partition for the next iteration (addition of the next individual). Both DR and ED use overall likelihood to rank the goodness of each partition. DR and ED could be generalized as two extremes of a method that is controlled by a window size w. DR corresponds to $w = 1$ while ED corresponds to $w = \infty$. ED is a brute-force algorithm and is not practical for the FSR problem that has the exponential partition search space.

The new algorithm was constructed by replacing the overall likelihood with the Simpson index as the scoring function for ranking the partitions in DR and ED, the following MS algorithm is proposed:

(1) Calculate pairwise matrix of genotype distances. This is a $O(n^2)$ step since there are $n(n-1)/2$ unique pairs.

(2) Create a list of pairs sorted in the ascending order of their genotype distances—$O(n^2 \log n)$.

(3) Create a pool of all individuals and mark them as unassigned—$O(n)$.

(4) Repeat this and the following steps until all individuals are assigned. Select a next unassigned individual from the pair with the lowest genotype distance from the list of sorted pairs—$O(n^2)$.

(5) For every existing partition create a set of new partitions by placing the individual to every existing group as well as by creating a new group containing just that individual—$O(wn^2)$. See Table 1 for an example.

(6) Discard any partition if its newly created group consists of individuals that cannot be full siblings simultaneously—$O(wn^3)$.

(7) Calculate the Simpson index for each of the new partitions—$O(wn^2 \log n)$ where $\log n$ is the estimated arithmetic cost.

(8) Sort the partitions in the descending order of their Simpson index—$O(wn^2 \log n)$.

(9) Discard the partitions with the lowest index to keep the total number of the retained partitions not exceeding w—$O(wn)$.

Steps (1) and (6) are done in $O(N_L)$, where N_L is the number of loci. In step (6) the extra $O(n)$ comes from the sibship consistency checking. Even though each check can be done in $O(1)$ (Almudevar and Field, 1999), the actual number of sib groups could still be proportional to n (Fig. 4a) yielding $O(n)$.

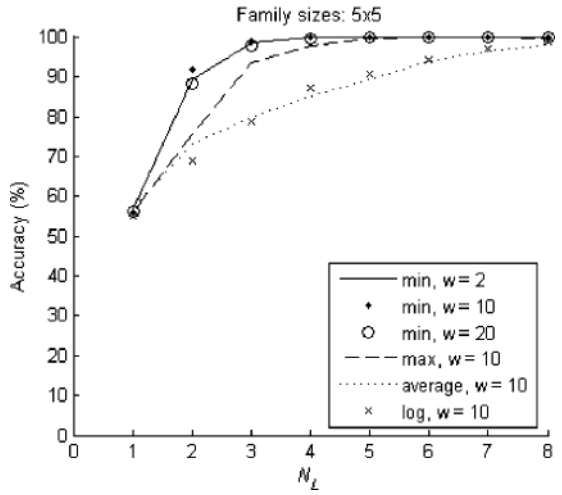

Fig. 2. Effect of the genotype distance definition on the reconstruction accuracy. Minimum (D_{XY}), maximum ($\hat{D}_{XY}$), average ($\bar{D}_{XY}$) and logarithmic (d_{XY}) locus distances are denoted by 'min', 'max', 'average' and 'log', respectively. Each reconstruction was done on a randomly generated population sample of five families each containing five full siblings, giving the total sample size $n = 25$. Each of the N_L loci was simulated with 10 equifrequent alleles ($N_A = 10$).

Since the steps (1–3) are at maximum $O(n^2 \log n)$ the whole algorithm is $O(n^3)$. However if the number of groups in the population sample is fixed or relatively stable, the overall theoretical complexity of the algorithm would be reduced.

2.6 Other genotype distances

Three other distance measures were also considered:

$$\hat{D}_{XY} = \max(D_l(X,Y)), \tag{5}$$

$$\bar{D}_{XY} = \frac{1}{N_L}\sum_{l=1}^{N_L} D_l(X,Y), \tag{6}$$

$$d_{XY} = -\ln((2-\bar{D}_{XY})/2), \tag{7}$$

where N_L is the number of loci. The average genotype distance $\bar{D}_{XY}$ relates directly to the logarithmic distance $d_{XY} = -\ln(M_{XY}/2)$ used by Blouin *et al.* (1996) as the measure of bandsharing between individuals, where $M_{XY} = 2-\bar{D}_{XY}$ is defined as the average number of allelic matches per locus between individuals X and Y. Figure 2 examines the effect each of the distances has on the accuracy of the FSR. The minimum based distance outperforms the other three in all the cases. In addition, the distance is resilient to the window size parameter w [compare $w = 2$ (solid line), $w = 10$ (solid dots) and $w = 20$ (circles) in Fig. 2]. Figure 2 demonstrates that the genotype distance allows for an accurate (better than 95%) reconstruction without the knowledge of the allelic frequencies in the biologically tractable region of 3–10 loci.

3 RESULTS AND DISCUSSION

Unless stated otherwise all results are averaged over 100 trials and summarized in Tables 2 and 3.

3.1 Accuracy

First of all the accuracy of the MS algorithm is verified against the original. Figure 3 displays the MS results for 50 individuals with the following family sizes: (a) 50×1 (50 unrelated), (b) 25×2 (25 families with 2 sibs each), (c) 5×10, (d) (20, 10, 10, 5, 5), (e) (30, 5, 5, 5, 5) and (f) (40, 5, 2, 2, 1). It is safe to conclude from Figure 3 that the new algorithm is more accurate than the original in all the considered cases.

Since any two individuals can always form a sib group, the theoretical limit of the MS and SIMPSON algorithms in Figure 3a is 50% (the reconstruction partition-distance cannot fall below 25). The failure of a SIMPSON index based algorithm to deal with the presence of unrelated individuals is known (Butler *et al.*, 2004) and is outside the scope of this paper.

Figures 3b and 3f demonstrates the advantage of the 'directed' search within the MS algorithm. In Figure 3f each trial sample contained a full sibling pair (fs1 and fs2) plus one unrelated individual (u1). Neither the Simpson index nor the Mendelian exclusion can differentiate the {{fs1, fs2}, {u1}}, {{fs1,u1}, {fs2}} and {{u1,fs2}, {fs1}} partitions, and therefore the SIMPSON algorithm returns one of the three partitions with equal probability (assuming that the rest of individuals are assigned correctly). On the other hand, MS is most likely to start with (and keep) the fs1–fs2 pair rather than fs1-u1 or fs2-u1 delivering the correct partitioning more often. While in the case of Figure 3f the effect is somewhat negligible, MS significantly outperforms SIMPSON in Figure 3b (25×2).

We also verified that the order in which the unassigned individuals were selected was critical (steps 1–4 of the MS algorithm). This was studied by creating a version of the MS algorithm where the ascending distance order was replaced by a random order. We found that the accuracy deteriorated significantly in the uniformly distributed families of 5 and 10 (results are not shown) while, as expected, the effect is negligible when a sample lacks any structure (e.g. 50×1) or when most individuals are of the same type (e.g. 1×50).

The DR results (Fig. 3) were obtained with the null and primary hypotheses being the unrelated and diploid full-sibling relationships, respectively. Figures 3a–f show that DR is more accurate than the GRAPH algorithm in all cases especially for highly skewed family distributions (Fig. 3f). As expected, the MS and SIMPSON algorithms were consistently more accurate than the likelihood based DR and GRAPH if non-small size families were actually present in the samples (families of size five or larger were used in Figs 3c–e). Note that the presented GRAPH results were re-calculated according to Equation (2) from the original results of Beyer and May (2003) for the average number of individuals who had to be moved to another family to create a correct classification, e.g. the original 8.09 (for 5×10, $N_L = 4$ and $N_A = 8$) is plotted as $(1 - 8.09/n) = 0.8382$ or 83.82%, where $n = 50$.

The presented MS accuracy results were obtained with $w = 2$ (Figs 3 and 4c) and $w = 10$ (Fig. 4c) confirming the stability of MS with respect to variations in the window size w.

3.2 Efficiency

There are a number of ways by which the efficiency of the algorithm could be measured. The most straightforward is the total running time per reconstruction. Figure 4a shows such a measure in arbitrary units of time. In practical terms one reconstruction of 100 families of 5 sibs each ($n = 500$) took an order of seconds on a 3 GHz PC, making the MS algorithm convenient for empirical FSRs and further study. Figure 4a also verifies the asymptotic $O(n^3)$ behavior of the MS algorithm.

Table 2. Accuracy (the percentage of correctly classified individuals) of the MS, SIMPSON and DR algorithms for 50 individuals in different family sizes

N_L	1	2	3	4	5	6	7	8	9	10	11	12
MS(50 × 1)	22.6	36.7	44.1	48.5	49.9	50	50	50	50	50	50	50
MS(25 × 2)	32.1	50.1	58.4	64.1	69.1	72.1	74.5	81.5	83.2	91.3	94.1	96.9
MS(5 × 10)	72.6	97.7	99.6	99.9	*	*	*	*	*	*	*	*
MS(20,10,10,5,5)	81.3	98	99.5	*	*	*	*	*	*	*	*	*
MS(30,5,5,5,5)	81.2	96.5	99.6	99.9	*	*	*	*	*	*	*	*
MS(40,5,2,2,1)	93	96.3	97.8	98.4	99.1	99.5	99.5	99.8	99.8	99.9	*	*
SIMPSON(50 × 1)	20.3	32.9	40.9	46.5	49.5	50	50	50	50	50	50	50
SIMPSON(25 × 2)	29.3	43.8	51.5	54.4	55	53.3	52.8	51.9	51.5	51	51	51.1
SIMPSON(5 × 10)	61.3	94.5	99.5	99.7	99.9	*	*	*	*	*	*	*
SIMPSON(20,10,10,5,5)	77	96.3	99.5	99.9	99.7	*	*	99.8	99.8	99.8	*	*
SIMPSON(30,5,5,5,5)	82.9	95.8	98.5	99.4	*	99.8	99.8	99.5	99.5	99.2	99.4	*
SIMPSON(40,5,2,2,1)	92.1	95.6	96.5	96.7	96.5	96.9	96.6	96.8	96.4	96.4	96.1	96.4
DR(50 × 1)	15.4	28.2	36.6	43.5	49.8	55.3	62.6	64.7	68.5	75.4	77.3	78.7
DR(25 × 2)	24.7	41.1	49.7	59	67	71	77.5	83.6	86.7	89.5	91.9	93.8
DR(5 × 10)	57.8	69.3	82	89.2	91.6	95.5	96.9	96.1	98	98.5	99.5	99.2
DR(20,10,10,5,5)	60.3	74.4	85.8	89.6	94	95.7	96.9	98.4	99.3	99.2	99.4	99.4
DR(30,5,5,5,5)	63.5	79.3	84.9	91.8	93.5	97.2	97.6	99.2	99.1	99.5	99.6	99.7
DR(40,5,2,2,1)	71.6	84.3	90.4	95.6	97.5	98.4	98.5	98.8	99.4	99.3	99.7	99.7

The sizes are displayed in parentheses. The MS results were obtained with the window parameter w =2. The SIMPSON results were obtained with $T_{50} = 10^5$. Each of the N_L loci was simulated with 8 equifrequent alleles (N_A =8). Asterisk denotes 100% accuracy.

Table 3. Accuracy (%) of the MS algorithm with two values for the window parameter w: MS ($w = 10$) and MS ($w = 2$)

r	10	20	40	60	80	100
MS ($w = 10$)	99.9	99.8				
MS ($w = 2$)	99.9	99.8	99.6	99.5	99.5	99.3

Each simulated population sample is generated using 5 loci ($N_L = 5$) and 10 equifrequent alleles ($N_A = 10$) consisting of equal size groups with 5 full siblings each. The total number of individuals in the sample is $n = 5 \times r$, where r is the number of groups.

The important part of the new algorithm is the building of every possible partition from a group limited by w set of partitions (retained from a previous iteration). Figure 4b verifies that the total number of examined groups increases as expected in $O(n^2)$.

Figures 4a and 4c compare the efficiency and accuracy of the MS and SIMPSON algorithms. The big O complexity measure of the SIMPSON algorithm obviously depends on the optimal value of T_n, which is an unknown function of n. To study the SIMPSON efficiency we used two simple estimations $T_n = n^3$ and $\hat{T}_n = n^3/5$. Comparing the running time for SIMPSON using $T_n = n^3$ (solid line in Fig. 4a) with MS ($w = 2$-crosses and $w = 10$-solid dots) we conclude that MS is significantly more efficient. For example, at $n = 50$ the original SIMPSON algorithm takes >100 times longer to achieve the same accuracy as MS ($w = 2$), where $T_{50} = 10^5$ was taken from Butler *et al*. (2004). The values of $T_n = n^3$ were limited by the maximum $T_{\max} = 10^5$ to demonstrate that the SIMPSON accuracy deteriorates very rapidly if a required number of iterations was not 'guessed' correctly.

Figure 4c indicates that $\hat{T}_n = n^3/5$ is probably the smallest (optimal) number of iterations before the accuracy of the SIMPSON algorithm deteriorates below 95%. However, $\hat{T}_n = n^3/5$ is insufficient at larger n to maintain the accuracy achieved as the small n,

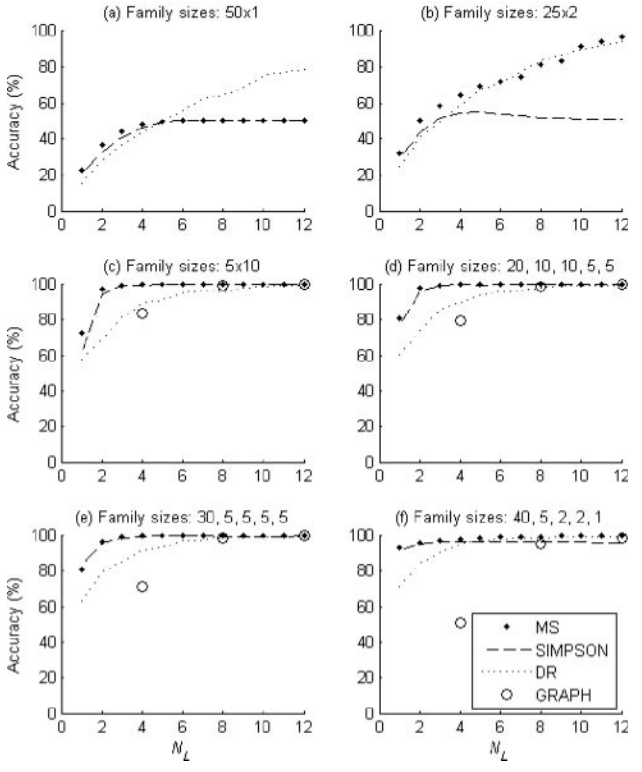

Fig. 3. Accuracy comparison of the MS algorithm with the SIMPSON (Butler *et al*., 2004), descending ratio (DR) (Konovalov *et al*., 2004) and GRAPH (Beyer and May, 2003) algorithms for 50 individuals in different family sizes. The MS results were obtained with the window parameter $w = 2$ and are denoted by solid dots. The DR and GRAPH results are denoted by dotted lines and circles, respectively. The SIMPSON results were obtained with $T_{50} = 10^5$, as per Butler *et al*. (2004), and are denoted by dashed lines. Each of the N_L loci was simulated with 8 equifrequent alleles ($N_A = 8$).

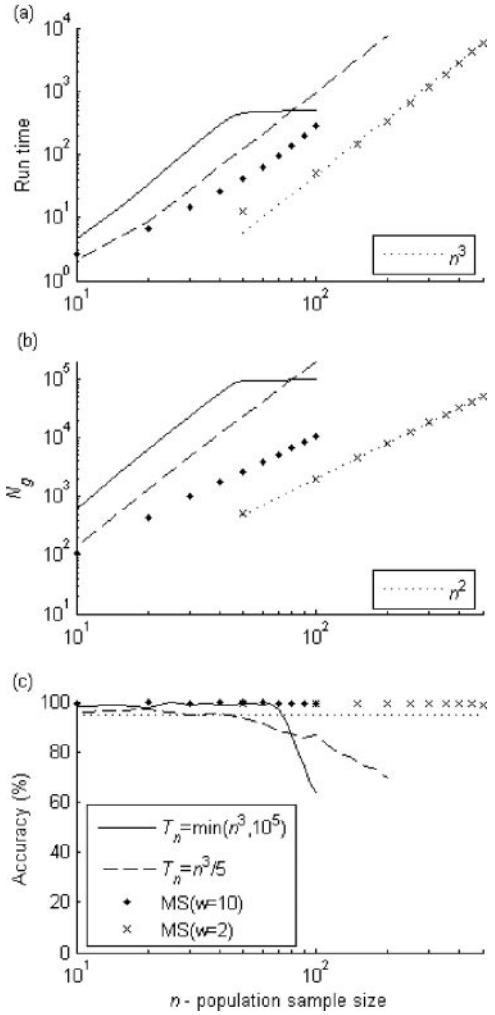

Fig. 4. Efficiency and accuracy (%) of the MS and SIMPSON algorithms (solid and dashed lines). N_g is the number of unique groups checked for sibship. Each simulated population sample is generated using 5 loci ($N_L = 5$) and 10 equifrequent alleles ($N_A = 10$) consisting of equal size groups with 5 full siblings each. The dotted line in subfigure (c) is the 95% level. The last calculated point is for 100 families of 5 sibs each, $n = 500$.

which shows that the SIMPSON algorithm could run in $O(n^\alpha)$, where $\alpha > 3$. We have also verified both theoretically and experimentally (results are not shown) that the DR algorithm runs in $O(n^3)$ or faster regardless of the family distributions.

4 CONCLUSION

We presented a polynomial-time algorithm based on the Simpson index that significantly outperforms the original SIMPSON algorithm of Butler *et al*. (2004) in any combination of accuracy and efficiency for the FSR problem. The new algorithm is governed by only one heuristic parameter (typically $w \leqslant 10$) and is stable to a wide range of variations of that parameter ($w = 2$, 10, and 20 were considered). More theoretical work could be carried out to discover a relationship between the optimal value for w (if one exists in terms of accuracy–efficiency trade-off) and the population parameters: size n, number of loci N_L and alleles N_A. In fact, we conjecture that w is mainly sensitive to N_L and N_A but relatively insensitive to n. It was shown in this paper that even $w(n = 50) = 1$ performs well in the case of the DR algorithm when applied to the FSR problem. This indicates strongly that even if the optimal value $w(N_L, N_A, n)$ does exist, the variation in accuracy should not be significant.

The MS algorithm was also shown to outperform the likelihood based DR and GRAPH algorithms for uniformly distributed families (Fig. 3c) as well as for skewed family distributions (Figs 3d–f). Only in the presence of unrelated individuals the MS algorithm was found to be less accurate than DR (Fig. 3a).

Even though the MS algorithm arguably makes the SIMPSON algorithm redundant, the question of what is the optimal number of iterations, T_n, remains an intriguing bioinformatics problem. Also it is unknown whether the Simpson index based formulation of the FSR problem could be solved exactly in polynomial time or can be shown to be NP-hard.

ACKNOWLEDGEMENTS

The authors thank David Browning, Christophe Herbinger and Tuan Pham for helpful discussions and assistance, as well as two anonymous referees for their constructive comments.

Conflict of Interest: none declared.

REFERENCES

Abramowitz,M. and Stegun,I.A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th printing edn. Dover, New York.

Almudevar,A. (2001) A bootstrap assessment of variability in pedigree reconstruction based on genetic markers. *Biometrics*, **57**, 757–763.

Almudevar,A. and Field,C. (1999) Estimation of single-generation sibling relationships based on DNA markers. *J. Agri. Biol. Environ. Stat*., **4**, 136–165.

Bell,E.T. (1934) Exponential numbers. *Amer. Math. Monthly*, **41**, 411–419.

Beyer,J. and May,B. (2003) A graph-theoretic approach to the partition of individuals into full-sib families. *Mol. Ecol*., **12**, 2243–2250.

Blouin,M.S. (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol. Evol*., **18**, 503–511.

Blouin,M.S. *et al*. (1996) Use of microsatellite loci to classify individuals by relatedness. *Mol. Ecol*., **5**, 393–401.

Butler,K. *et al*. (2004) Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Mol. Ecol*., **13**, 1589–1600.

Gusfield,D. (2002) Partition-distance: a problem and class of perfect graphs arising in clustering. *Inform. Proc. Lett*., **82**, 159–164.

Konovalov,D.A. *et al*. (2005) Partition-distance via the assignment problem. *Bioinformatics*, **21**, 2463–2468.

Konovalov,D.A. *et al*. (2004) KinGroup: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Mol. Ecol. Notes*, **4**, 779–782.

Lovasz,L. (1993) *Combinatorial Problems and Exercises*. 2nd edn. Netherlands: North-Holland, Amsterdam.

Pemberton,J. (2004) Measuring inbreeding depression in the wild: the old ways are the best. *Trends Ecol. Evol*., **19**, 613–615.

Smith,B.R. *et al*. (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.

Thomas,S.C. and Hill,W.G. (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, **155**, 1961–1972.

Thomas,S.C. and Hill,W.G. (2002) Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genet. Res*., **79**, 227–234.

Wang,J.L. (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.