Statistics and Its Interface Volume 4 (2011) 381–387

Stochastic deletion-insertion algorithm to construct dense linkage maps

Jixiang Wu*†, Xiang-Yang Lou‡ and Michael Gonda*

In this study, we proposed a stochastic deletion-insertion (SDI) algorithm for constructing large-scale linkage maps. This SDI algorithm was compared with three published approximation approaches, the seriation (SER), neighbor mapping (NM), and unidirectional growth (UG) approaches, on the basis of simulated F_2 data with different population sizes, missing genotype rates, and numbers of markers. Simulation results showed that the SDI method had a similar or higher percentage of correct linkage orders than the other three methods. This SDI algorithm was also applied to a real dataset and compared with the other three methods. The total linkage map distance (cM) obtained by the SDI method (148.08 cM) was smaller than the distance obtained by SER (225.52 cM) and two published distances (150.11 cM and 150.38 cM). Since this SDI algorithm is stochastic, a more accurate linkage order can be quickly obtained by repeating this algorithm. Thus, this SDI method, which combines the advantages of accuracy and speed, is an important addition to the current linkage mapping toolkit for constructing improved linkage maps.

AMS 2000 subject classifications: Primary 60H30; secondary 60E15.

Keywords and phrases: Linkage mapping, Stochastic deletion-insertion algorithm.

1. INTRODUCTION

Molecular markers are essential for quantitative trait locus (QTL) mapping and gathering molecular genetic information for the improvement of plant and animal species. With the advance of technology, the number of loci examined in different chromosome regions is continuously growing, which has quickly increased the computational burden for constructing high quality linkage maps. Thus, a linkage mapping method with high accuracy and less computational burden is needed for ordering a large number of loci.

*Supported in part by the Agricultural Experiment Station at the South Dakota State University.

†Corresponding author.

‡Supported in part by the National Institutes of Health Grant R01 DA025095.

Marker ordering was once considered a special case of the travelling salesman problem (TSP) [8], which is a classical non-deterministic polynomial-time (NP) complete problem [16, 11, 4]. There are two general types of approaches to tackling this problem. The first approach performs exhaustive searches to find the global optimal marker order solution. However, this approach is not practical when the number of linked loci is large because $n!/2$ distinct orders need to be evaluated for n linked loci. Even for just 10 loci on one chromosome, there are $1,814,400$ possible orders. For many experiments, n might vary from dozens to several hundreds. Thus, an exhaustive search is clearly computationally prohibited when the number of loci is greater than 30 [9, 10, 14]. The second approach uses approximation algorithms for determination of linkage order. These approximation algorithms for obtaining near-optimal marker order solutions are more practical for large-scale linkage mapping [8]. To date, several approximation algorithms for determination of linkage order have been proposed (i.e. [1, 13, 15, 7, 3, 2, 14, 12, 10]). All these linkage methods are based on the biological assumption that the true order of a set of linked loci has a minimum sum of adjacent recombination frequencies (SARF) or a maximum sum of adjacent LOD scores (SALOD).

Many approximation algorithms construct linkage maps sequentially; these algorithms start with only two markers and then add one of the remaining loci to the linkage map at a time until all loci are mapped. The advantage of the approximation mapping approaches is speed. However, since these algorithms are heuristic, the global optimal solution may not be reached if a two-point recombination frequency data set is not monotonic. Missing markers, crossover interference, and small population sizes can cause data non-monotonicity. An alternative approximation algorithm is the evolutionary strategy (ES) algorithm [10], which starts with an initial set of many random linkage orders and eventually reaches an optimal linkage order through crossover and mutant operators. Thus, the final linkage order and the speed to reach this final solution greatly depend on the size of the initial set of linkage orders and the operators chosen.

Though a linkage order achieved by approximation methods may not be the global optimal solution if a dataset is not monotonic, a local linkage order and the global solution would have similar linkage orders. It is unlikely that a local linkage order obtained by an approximation method is

Table 1. Four distinct linkage orders when a 4^{th} marker added to a partial linkage map with three markers

	Four possible linkage orders			
1:	**4** →	1 →	2 →	3
2:	1 →	**4** →	2 →	3
3:	1 →	2 →	**4** →	3
4:	1 →	2 →	3 →	**4**

completely different from the global optimal solution, implying a possibility that we can evolve a better/global order solution from some local optimal ones. Such a linkage mapping algorithm will improve accuracy compared to existing approximation methods while reducing the computational intensity relative to the exhaustive search algorithm. In this study, a stochastic deletion-insertion (SDI) algorithm, which minimizes SARF or maximizes SALOD values, is proposed. Various simulated F_2 data were generated to evaluate the performance of this approach with different population sizes, numbers of markers, and missing genotype rates. A real data set was also used to demonstrate the benefit of this SDI algorithm.

2. SDI ALGORITHM

If an optimal solution for a partial linkage map [12] with *three* loci has been determined, denoted $1 \rightarrow 2 \rightarrow 3$, then an optimal linkage order with a new locus 4 on the same chromosome can be determined through comparing the *four* possible orders (Table 1). Thus, the linkage order with the locus 4 that minimizes the SARF or maximizes the SALOD value should be the optimal solution conditioning on the previous linkage order. This process can continue until all unmapped loci n are added to the linkage map.

The SDI algorithm proceeds as follows:

- Step 1: Obtain the recombination frequency matrix for all possible pairs of n loci on the same chromosome. If multiple linkage groups are found, a criterion is required to separate the different loci into more than one linkage group [17]. A corresponding distance matrix can be converted from the recombination frequency matrix using a mapping function (i.e., [5, 8]). Since the distance matrix and recombination frequency matrix used for linkage mapping result in the same marker order, without loss of generality, the recombination frequency matrix will be used throughout this study.
- Step 2: Generate an initial linkage order L_B by using seriation (SER) [1], unidirectional growth (UG) [12], or neighbor mapping (NM) [3].
- Step 3: Deletion process:
  - 3a: Generate a random number s $(1 < s < n-1)$
  - 3b: Randomly select s loci from n loci in linkage order L_B to form a new sub linkage group L_s without changing the order of these s loci.
- Step 4: Insertion process: Randomly select one of unmapped $n-s$ loci and insert this locus into the linkage map L_s with $(s+1)$ possible positions. Thus a new optimal linkage order with $(s+1)$ loci can be determined by comparing SARF. Repeat this process until all $n-s$ markers are mapped into L_s. If L_1 is better than L_B, then set $L_B = L_1$.
- Step 5: Repeat Steps 3 to 4 for N times. As an alternative, one can use simulated annealing or other proper approaches in this step.
- Step 6: Repeat Step 5 until no better solution is obtained for a given number of consecutive times (we used 10 times in this study).

If a recombination fraction matrix is monotonic, then the global optimal solution could be obtained at step 2. If the recombination fraction matrix is not monotonic, the linkage order obtained by the SDI method at each iteration may be a local optimal solution. Since the process of this deletion-insertion procedure is stochastic, the SDI method has a higher probability of reaching the global optimal solution by repeating Step 5. It should be noted that a little larger N which depends on the number of markers, genotype missing rate, population size, etc., can be employed (see Simulation Results for details) in Step 5 for a large number of linked loci because a desirable solution may not be reached due to possible immature convergence.

3. SIMULATION SCENARIOS

Since in reality, few real data sets are available with known marker orders, computer simulations are an alternative method for evaluating the performance of this SDI algorithm and several widely used approximation algorithms: SER [1], NM [3], and UG [12]. In this study we simulated an F_2 population derived from a cross between two inbred lines because F_2 populations that have more segregated types are a commonly used experimental design for linkage mapping studies. To better demonstrate the performance of the SDI algorithm, we carefully considered several factors that may contribute to data non-monotonicity: distance among linked loci, mapping population size, missing genotype rate, and recombination interference. To increase the possibility of generating a non-monotonic dataset, probability distributions for distance between adjacent markers were set as follows: $P(1 < \text{d(cM)} < 5) = 0.70$, $P(5 < \text{d(cM)} < 10) = 0.15$, $P(10 < \text{d(cM)} < 30) = 0.10$, and $P(30 < \text{d(cM)} < 40) = 0.05$, with a uniform distribution within each of the four ranges. Data were simulated with six population sizes ($N = 50, 100, 150, 200, 250, \text{or } 300$), three numbers of co-dominant markers within the linkage group ($N = 10, 20, \text{or } 50$), and three missing genotype rates ($0\%, 10\%, \text{or } 15\%$). We also considered the presence of recombination interference. For simplicity of generating recombination interference, the recombination interference coefficient was randomly selected from the four possible values:

1.0 (without inference), 0.0 (complete positive interference), 0.5 (partial positive interference), and 2.0 (negative interference). The recombination frequency between any two markers was calculated by the Expectation-Maximization (EM) algorithm [8]. The mean percentage of correct order (PCO) value [8], a probability that a locus is ordered correctly, was used to compare different mapping methods. All simulations were repeated 200 times for each of these configurations with a computer program in C++ written by the authors of this paper. The significance of population size, marker number, missing genotype rate, and linkage mapping algorithm on PCO were estimated by one-way ANOVA by JMP 8.0 software (SAS Institute, Inc., Cary, NC).

4. SIMULATION RESULTS

4.1 The impact of repetitions on PCO values for the SDI method

Since an optimal linkage order obtained by the SDI approximation algorithm may be a local solution if a dataset is not monotonic, iterating the SDI algorithm is necessary to achieve a better solution. Therefore, it is helpful to determine the number of iterations sufficient to achieve a high PCO. Though we conducted various simulations for the impact of repetitions on PCO values, only the PCO values for each of six numbers of linked loci $(10, 20, 50, 100, 200,$ and $500)$ from 1 to 100 repetitions over 20 simulated F_2 data sets (10 with no missing rate and 10 with 10% missing rate, population size = 100) are reported in Figure 1. Our simulated results suggested that the PCO values remained stable after 20 repetitions (iterations) for most cases with 200 or fewer linked loci (Figure 1). The PCO values stabilized after only repeating 70 times for 500 linked loci. For a few cases (population sample size 50 with missing genotype rate 15%), over 1,000 repetitions were needed before the PCO values stabilized (data not shown). Thus, the simulation results implied that this SDI algorithm could reach a desirable solution with a light computational burden. In our following simulations, five times ($N = 5$) in step 5 were used and 10 consecutive times in step 6 were employed until no better solution was obtained, thus there were at least 55 iterations for each simulated data set.

4.2 Comparisons of PCO values among four mapping algorithm approaches

The empirical PCO values for four approximation algorithms (SER, NM, UG, and SDI) were calculated by comparing simulated data with results from each of the approximation algorithms (Tables 2 and 3). The PCO values for the SDI algorithm were consistently equal to or higher than the other three approximation algorithms (Table 2). In addition, the SDI algorithm was robust for small population sizes and maintained a high PCO (>90%) when the missing genotype data rate was high (15%) and the population size was ≥100.

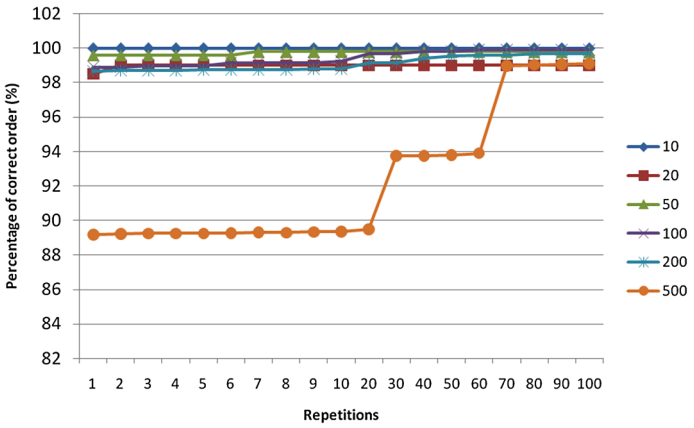

Figure 1. The impact of repetitions on percentage of correct order (PCO) over 20 F_2 *simulated data sets (*10 *data sets with no missing data rate and* 10 *with* 10% *missing rates) for each of six numbers of linked loci* $(10, 20, 50, 100, 200,$ and $500)$.

Therefore, our results suggest that the SDI method was the best among these four algorithms. On average, the NM algorithm was the poorest among the four methods and was sensitive to small population sizes, large marker numbers, and high missing genotype rates. The other two algorithms (SER and UG) performed nearly equally well in our simulation study: PCOs were often higher than the NM but lower than the SDI.

As expected, PCO increased with larger population sizes, smaller number of markers, and lower missing genotype rates (Table 2). When the missing genotype rate was 0%, a desirable mapping power (≥95%) could be achieved by SER, UG, and SDI algorithms when population size was ≥100. When the missing genotype rate was 10% or 15%, a desirable mapping power (≥95%) could be achieved for SER, UG, and SDI when population size ≥200. The NM algorithm required a population size ≥200 to consistently achieve a PCO ≥95% when the missing genotype rate was 0%. When the missing genotype rate was 15%, the NM algorithm did not achieve a PCO ≥95% under all simulated situations.

Population size, linkage mapping algorithm, and missing genotype rate were significantly associated with PCO ($P < 0.05$) (Table 3). Population size explained the largest amount of variation in PCO ($R^2 = 0.478$) when estimating marker order with our simulated dataset. The PCO was >95% when the population size was 150 or greater. Number of markers was not significantly associated with PCO ($P = 0.128$), although the PCO decreased as the number of markers increased (in general, as the number of markers increases, both the chance to cause non-monotonicity of the recombination frequency matrix and the solution space will increase, resulting in a higher probability to be trapped by a local solution and thus a reduced PCO).

Table 2. The percentage of correct orders (PCO, %) obtained by four algorithms for six population sizes (PS), three missing genotype rates (MS, %) and different numbers of markers (MN) on one chromosome

MN	PS	SER[1]			UG			NM			SDI		
		0%	10%	15%	0%	10%	15%	0%	10%	15%	0%	10%	15%
10	50	94.50	87.90	83.05	93.65	87.90	81.60	82.10	70.55	63.90	95.25	91.30	87.35
	100	99.05	95.10	95.70	98.30	95.95	94.30	89.95	81.25	79.00	99.25	96.90	97.80
	150	98.60	97.70	97.00	98.40	97.05	97.10	94.05	88.05	83.00	98.80	98.00	97.35
	200	99.90	99.55	99.00	99.90	99.50	99.30	96.70	90.35	89.45	99.90	99.75	99.50
	250	99.50	99.35	99.45	99.40	99.55	99.00	98.00	93.00	90.70	99.70	99.55	99.50
	300	100.00	98.90	100.00	99.20	99.15	99.10	98.30	92.55	92.35	100.00	99.75	99.65
20	50	92.77	83.25	78.57	93.53	81.40	81.33	84.95	66.85	62.05	95.10	89.00	86.30
	100	94.35	95.38	94.33	94.23	94.85	94.20	92.80	84.67	84.17	96.58	97.42	97.97
	150	97.78	97.02	97.15	97.08	95.67	96.53	96.03	90.95	88.80	97.85	98.30	98.20
	200	97.80	98.17	96.58	96.78	95.58	95.53	98.30	95.17	89.42	98.85	99.10	97.32
	250	99.20	98.28	96.05	98.35	97.42	94.53	98.63	96.10	91.95	99.25	98.42	96.63
	300	99.30	98.90	97.97	97.90	98.08	97.65	98.72	96.15	94.63	99.30	98.90	98.58
50	50	89.49	76.52	67.09	88.50	78.20	71.06	85.22	61.18	51.31	92.81	87.62	79.15
	100	96.08	93.28	89.46	95.63	91.91	85.47	95.20	81.17	72.44	96.45	96.66	93.65
	150	98.57	97.52	94.77	95.17	96.60	90.50	98.30	92.38	83.70	98.75	99.21	96.17
	200	98.88	97.03	98.14	97.00	92.29	95.27	98.95	94.19	91.68	99.16	97.78	98.51
	250	96.34	97.20	95.98	94.98	92.98	92.93	95.91	95.30	91.96	96.84	97.21	96.34
	300	98.90	98.63	97.29	96.83	96.66	94.13	98.94	98.07	93.65	99.02	98.67	97.79

[1] SER = seriation, UG = unidirectional growth, NM = neighbor mapping, and SDI = stochastic deletion-insertion.

Table 3. Effects of population size (PS), marker number (MN), missing genotype rates (MS, %), and linkage mapping algorithm (LMA) on percent correct order (PCO, %)[1]

Effect	Level	Mean PCO, %	F-ratio	P-value	Adj R^2
PS			40.35	<0.0001	0.478
	50	81.73			
	100	92.53			
	150	95.50			
	200	95.90			
	250	96.82			
	300	97.88			
MN			2.07	0.128	0.010
	10	94.65			
	20	93.98			
	50	92.06			
MS, %			9.61	0.0001	0.074
	0	96.52			
	10	93.22			
	15	90.96			
LMA[2]			13.98	<0.0001	0.153
	NM	88.21			
	SDI	96.74			
	SER	95.15			
	UG	94.17			

[1]F-ratios, P-values, and adjusted R^2 calculated individually for each effect by one-way ANOVA.
[2]SER = seriation, UG = unidirectional growth, NM = neighbor mapping, and SDI = stochastic deletion-insertion.

The NM algorithm on average resulted in the lowest PCO (88.21%). The SDI algorithm resulted in the highest PCO (96.74%), although both SER and UG algorithms had PCOs (95.15% and 94.17%, respectively) similar to the SDI.

5. A REAL DATASET

To evaluate the performance of the SDI algorithm, we applied SDI to a real dataset of 26 loci on barley chromosome I from the North American Barley Genome Mapping project (see Liu 1998, [8], p. 288). Though the number of linked loci was 26, there are two reasons we chose this dataset: (1) the recombination matrix was not monotonic because different approximation methods generated different linkage orders and (2) two different linkage orders have been published using this dataset [8, 12]. The linkage orders obtained by the SER and SDI methods in this paper were compared to the two published orders [8, 12]. The genetic distances presented in this paper were calculated based on Haldane's mapping function [5], so the total distances listed in Table 4 for the two published linkage orders are slightly different from the published distances [8, 12]. Even though the different mapping functions may slightly result in different genetic distances, the marker orders from the different mapping functions still remained the same. Results showed that the total distances obtained by the UG algorithm [12] and the combination of Simulated Annealing (SA) and Branch and Bound (BB) algorithms used by Liu [8] were very close to each other (150.38 cM vs 150.11 cM, respectively) (Table 4). The SER algorithm generated the largest genetic distance (225.52 cM), while the SDI algorithm generated the shortest

Table 4. Marker orders and their genetic distances (cM) obtained from different mapping algorithms

Liu[1]		Tan & Fu		SER		SDI	
Order	Dist	Order	Dist	Order	Dist	Order	Dist
[2] 1	0.00	1	0.00	1	0.00	1	0.00
2	12.65	2	12.65	2	12.65	2	12.65
3	4.80	3	4.80	3	4.80	3	4.80
4	15.52	4	15.52	5	16.47	5	16.47
5	3.94	5	3.94	4	3.94	4	3.94
6	11.55	7	13.03	7	11.61	7	11.61
7	0.76	6	0.76	6	0.76	6	0.76
8	8.68	9	5.05	8	7.67	9	5.05
9	0.77	8	0.77	9	0.77	8	0.77
10	1.52	10	2.11	10	1.52	10	2.11
11	2.93	11	2.93	11	2.93	11	2.93
12	2.25	12	2.25	12	2.25	12	2.25
13	5.22	14	2.27	14	2.27	13	5.22
14	4.51	13	4.51	15	3.10	14	4.51
15	3.10	15	7.89	16	8.42	15	3.10
16	8.42	16	8.42	17	13.29	16	8.42
17	13.29	17	13.29	18	6.99	17	13.29
18	6.99	18	6.99	19	21.89	18	6.99
19	21.89	19	21.89	20	3.54	19	21.89
20	3.54	20	3.54	21	2.21	20	3.54
21	2.21	21	2.21	22	0.00	21	2.21
22	0.00	22	0.00	24	0.73	22	0.00
23	2.17	23	2.17	23	1.48	23	2.17
24	1.48	24	1.48	25	8.34	24	1.48
25	3.70	25	3.70	26	8.21	25	3.70
26	8.21	26	8.21	13	79.68	26	8.21
[3]	150.11		150.38		225.52		148.08

[1]Liu = the combination of Simulated Annealing (SA) and Branch and Bound (BB) algorithms, Tan & Fu = unidirectional growth (UG), SER = seriation, and SDI = stochastic deletion-insertion.
[2]Marker codes for $1 \sim 26$ and the linkage order were reported by Liu (1998); the linkage order obtained by the UG method was reported by Tan and Fu (2006); and linkage orders by the other two methods were reported by this paper.
[3]The values in the bottom line represent the genetic distance obtained by Haldane's function.

distance (148.08 cM) by updating the linkage order obtained by the SER method with a small number of iterations.

6. DISCUSSION

As pointed out earlier, the exhaustive search algorithm can be applied to find the global optimal solution only when the number of linked loci is small. The evolutionary strategy (ES) algorithm can have the potential to find the correct linkage order but it is computationally intensive and the final solution is dependent on many factors such as the cross operator, mutation operator, and population size. On the other hand, many approximation algorithms like NM, SER, and UG have the advantage of speed but may not find the global optimal linkage order when the data is non-monotonic. If a two-point dataset for linkage mapping is monotonic, all these approximation algorithms can reach the global optimal linkage order. However, the recombination frequency matrix is calculated from a sample, which could be influenced by many factors like sample size, linkage distance among markers, presence of crossover interference, and missing genotypes [9, 10].

Through partially updating linkage orders obtained by approximation methods, the SDI algorithm successfully improved mapping accuracy. The first advantage of this SDI algorithm is the high power in recovering the true map. Our simulations showed that the SDI method was the best among four methods to resolve true linkage orders. The application to an experimental data set [8] also showed that the SDI method could find the linkage order with the shortest linkage distance. The second advantage of the SDI method is to balance the accuracy and computing speed well. Compared to several approximate algorithms such as SER, UG and UM, the SDI algorithm only takes a few additional iterations and minor extra computational time to reach an improved solution; however, it greatly reduces computational intensity as compared to the exhaustive search algorithm.

Two significant differences exist between the SDI method and other approximation methods. The first difference is

that many approximation methods add one remaining locus onto one of two ends of a partial linkage order at each step. Therefore, these methods are hill-climbing approaches. The SDI method, however, allows an unmapped locus to be added onto any position of a partial linkage order at each step, increasing the chance of obtaining a more accurate linkage map. The second difference is that previous approximation methods can obtain only one solution; whereas, the SDI algorithm generates more than one optimal solution, leading to improved power for finding a more accurate solution. In practice, researchers can choose the best linkage order from several approximation methods as an initial order. Then, the SDI algorithm can be employed to update and improve this linkage order as more data are collected. Therefore, this SDI algorithm should be an important supplement to the current tools to improve linkage mapping power.

Linkage power depends on estimates of recombination frequencies. Many factors such as linkage distance, population size, missing genotype rate, and crossover interference can lead to biased estimation and thus cause non-monotonicity of the recombination frequency matrix [9]. These four factors were considered in our simulation study to increase the possibility of non-monotonic data. It appeared that the SDI algorithm was robust to non-monotonic data in recovering true linkage maps after a few iterations; however, we observed that a large number of iterations (more than 1,000) were needed to resolve true linkage orders for a few cases (small population sizes and high missing genotype rates). This observation suggests that a large N in step 5 will be needed to achieve a more accurate linkage order for a large number of linked loci in a small population with a high missing genotype rate. We also observed that resolving powers decreased when the number of iterations for the SDI method increased in several cases where the populations size was small (50) with high missing genotype rates (10% and above). This result suggests that small population sizes and high missing genotype rates should not be recommended when constructing large-scale linkage maps.

The types of markers in an F_2 population are related to the mapping power as well [6, 8, 17, 9]. Dominant markers act differently under conditions of coupling-phase and repulsion-phase. When all dominant markers were in coupling-phase, the proportion of dominant and co-dominant markers had little impact on mapping power [9, 17]. However, the repulsion-phase dominant markers often caused biased estimations and thus low mapping power [8, 17]. Hence, the direct use of minimum SARF or maximum SALOD values may not work efficiently when markers are in repulsion-phase. Knapp et al. [6, 9] suggested that two complimentary linkage orders, each including co-dominant markers, can be constructed. An alternative methodology is to use the SDI method to merge two complimentary maps into one. Further investigation is needed to improve the accuracy of marker order and estimating linkage distance.

7. CONCLUSIONS

A stochastic deletion-insertion (SDI) algorithm for constructing large-scale linkage maps was proposed in this study. The mapping power of this SDI algorithm was evaluated along with three published approximation approaches: SER, NM, UG, on the basis of simulated F_2 data with different population sizes, missing genotype rates, and numbers of linked loci. Simulation results showed that the SDI method had similar or higher PCO values than the other three methods. The NM algorithm on average resulted in the lowest PCO (88.21%). The SDI algorithm resulted in the highest PCO (96.74%). Both SER and UG algorithms had PCOs (95.15% and 94.17%, respectively) less than the SDI. The SDI algorithm was also applied to a real dataset and compared with the other three methods. The total linkage map distance obtained by the SDI method was smaller than the distance obtained by SER and two published distances. Since this SDI algorithm is stochastic and progressive-iterative, it increases the possibility to obtain a more accurate linkage order. Thus, this SDI method, which combines the advantages of accuracy and speed, is an important addition to the current linkage mapping toolkit for constructing dense linkage maps.

ACKNOWLEDGEMENTS

The authors would like to thank the two reviewers for their helpful comments that have helped improve this paper. We are also grateful to Krishna Bondalapati for her great help reformatting this paper.

Received 1 June 2010

REFERENCES

[1] Buetow, K. H. and Chakravarti, A. (1987). Multipoint gene mapping using seriation. I. General methods. *American Journal of Human Genetics* **41** 180–188.

[2] Crane, C. F. and Crane, Y. M. (2005). A nearest-neighboring-end algorithm for genetic mapping. *Bioinformatics* **21** 1579–1591.

[3] Ellis, T. H. N. (1997). Neighbour mapping as a method for ordering genetic markers. *Genetical Research* **69** 35–43.

[4] Falk, C. T. (1992). Preliminary ordering of multiple linked loci using pairwise linkage data. *Genetic Epidemiology* **9** 367–375.

[5] Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8** 299–309.

[6] Knapp, S. J., Holloway, J. L., Bridges, W. C. and Liu, B. H. (1995). Mapping dominant markers using F2 mating. *Theoretical and Applied Genetics* **91** 74–81.

[7] Lathrop, G. M., Lalouel, J. M., Julier, C. and Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics* **37** 482–498.

[8] Liu, B. (1998). *Statistical Genomics: Linkage, Mapping, and QTL Analysis*. CRC Press LLC.

[9] Mester, D. I., Ronin, Y., Hu, Y., Peng, J., Nevo, E. and Korol, A. (2003a). Efficient multipoint mapping: making use of dominant repulsion-phase markers. *Theoretical Applied Genetics* **107** 1102–1112.

[10] Mester, D. I., Ronin, Y., Minkov, D., Nevo, E. and Korol, A. (2003b). Constructing large scale genetic maps using an evolutionary strategy algorithm. *Genetics* **165** 2269–2282.

[11] Olson, J. M. and Boehnke, M. (1990). Monte Carlo comparison of preliminary methods for ordering multiple genetic loci. *American Journal of Human Genetics* **47** 470–482.

[12] Tan, Y. and Fu, Y. (2006). A novel method for estimating linkage maps. *Genetics* **173** 2383–2390.

[13] Thompson, E. (1984). Information gain in joint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **1** 31–49.

[14] Van Os, H., Stam, P., Visser, R. G. F. and Van Eck, H. J. (2005). RECORD: a novel method for ordering loci on a genetic linkage map. *Theoretical and Applied Genetics* **112** 30–40.

[15] Weeks, D. and Lange, K. (1987). Preliminary ranking procedures for multilocus ordering. *Genomics* **1** 236–242.

[16] Wilson Sr, M. (1988). A major simplification in the preliminary ordering of linked loci. *Genetic Epidemiology* **5** 75–80.

[17] Wu, J., Jenkins, J., Zhu, J., McCarty, J. and Watson, C. (2003). Monte Carlo simulations on marker grouping and ordering. *Theoretical and Applied Genetics* **107** 568–573.

Jixiang Wu
Plant Science Department
South Dakota State University
Box 2140C, Brookings
SD 57007
USA
E-mail address: jixiang.wu@sdstate.edu

Xiang-Yang Lou
Section on Statistical Genetics
Department of Biostatistics
University of Alabama at Birmingham
Royals Public Health Building, Suite 414
1665 University Boulevard
Birmingham
Alabama 35294
USA
E-mail address: XLou@ms.soph.uab.edu

Michael Gonda
Animal and Range Science Department
South Dakota State University
Box 2170, Brookings
SD 57007
USA
E-mail address: michael.gonda@sdstate.edu