BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 11 2010, pages 1446–1452
doi:10.1093/bioinformatics/btq158

Genetics and population analysis

Advance Access publication April 12, 2010

Multi-objective tag SNPs selection using evolutionary algorithms

Chuan-Kang Ting* , Wei-Ting Lin and Yao-Ting Huang*

Department of Computer Science and Information Engineering, National Chung Cheng University, Chia-Yi 621, Taiwan

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Integrated analysis of single nucleotide polymorphisms (SNPs) and structure variations showed that the extent of linkage disequilibrium is common across different types of genetic variants. A subset of SNPs (called tag SNPs) is sufficient for capturing alleles of bi-allelic and even multi-allelic variants. However, accuracy and power of tag SNPs are affected by several factors, including genotyping failure, errors and tagging bias of certain alleles. In addition, different sets of tag SNPs should be selected for fulfilling requirements of various genotyping platforms and projects.

Results: This study formulates the problem of selecting tag SNPs into a four-objective optimization problem that minimizes the total amount of tag SNPs, maximizes tolerance for missing data, enlarges and balances detection power of each allele class. To resolve this problem, we propose evolutionary algorithms incorporated with greedy initialization to find non-dominated solutions considering all objectives simultaneously. This method provides users with great flexibility to extract different sets of tag SNPs for different platforms and scenarios (e.g. up to 100 tags and 10% missing rate). Compared to conventional methods, our method explores larger search space and requires shorter convergence time. Experimental results revealed strong and weak conflicts among these objectives. In particular, a small number of additional tag SNPs can provide sufficient tolerance and balanced power given the low missing and error rates of today's genotyping platforms.

Availability: The software is freely available at *Bioinformatics* online and http://cilab.cs.ccu.edu.tw/service_dl.html

Contact: ckting@cs.ccu.edu.tw; ythuang@cs.ccu.edu.tw

Received on January 27, 2010; revised on April 6, 2010; accepted on April 7, 2010

1 INTRODUCTION

Genetic differences between two individuals range from single nucleotide polymorphisms (SNPs) to large structure variations (e.g. deletions and duplications). Based on analysis of linkage disequilibrium (LD) among SNPs, the entire human genome is shown to be composed of high-LD blocks interspersed by recombination hotspots (Altshuler *et al.*, 2005; Hinds *et al.*, 2005). A small subset of SNPs (termed *tag SNPs*) is capable of capturing haplotype information in a high-LD block (Carlson *et al.*, 2004; Chang *et al.*, 2006). Recently, large structure variations have been discovered by array comparative genomic hybridization (aCGH) or massively parallel sequencing platforms. Integrated analysis of SNPs and structure variations showed that the extent of LD is not limited to SNPs. For example, a number of copy number variations (CNVs) are found to exhibit high LD with flanking SNPs, implying the copy number of each individual is now inferable by tag SNPs (McCarroll *et al.*, 2008; Redon *et al.*, 2006). In an analysis of an inversion polymorphism at 8p23, the orientation of each haplotype can be predicted by 13 tag SNPs (Deng *et al.*, 2008). These results suggest that a single SNP genotyping platform and well-chosen tag SNPs are sufficient for capturing alleles of structure variations.

Owing to the limited capacity of a genotyping platform, the number of tag SNPs is often minimized to increase genome-wide coverage, which is ordinarily done by capturing extent of LD (e.g. r^2 and D') or haplotype diversity (Carlson *et al.*, 2004; Zhang *et al.*, 2004). However, alleles of tag SNPs may be missed or miscalled due to defective quality of signals during genotyping (Zhao *et al.*, 2002). These missing data and genotyping errors greatly reduce the accuracy of tag SNPs, because alleles of tagged variations may be wrongly inferred (Huang *et al.*, 2005; Liu *et al.*, 2006). Recently, variants of these methods have also been developed for tagging multi-allelic variations (e.g. CNV) (McCarroll *et al.*, 2008; Redon *et al.*, 2006). Nevertheless, the power of tag SNPs is further affected by tagging bias toward certain alleles. Some alleles may be well distinguished by many tag SNPs, whereas others are distinguished by only a few tags. Thus, subsequent association studies using the tag SNPs will fail to provide unbiased detection power for each allele.

These issues can be formulated into different objectives in tag SNPs selection. A small number of tag SNPs, on the one hand, is always desired given the limited capacity of a genotyping platform (Carlson *et al.*, 2004). On the other hand, the number of tag SNPs has to be increased in order to tolerate the influence of missing data and genotyping errors (Huang *et al.*, 2005). Additionally, the distance and diversity among haplotype backgrounds of different allele classes should be considered for providing sufficient and unbiased power in distinguishing each allele. However, some of these objectives are intrinsically conflict, e.g. genotyping cost and tolerance. Existing methods for tag SNPs selection fail to take multiple objectives into account. Moreover, different sets of tag SNPs are required to accommodate various platforms and scenarios (e.g. maximum 100 tags and tolerance for 10% missing data). Thus, a method that can simultaneously address these issues and generate multiple non-dominated solutions satisfying distinct constraints of each objective is highly demanded.

This study formulates the tag SNP selection problem into a four-objective optimization problem that minimizes the total amount of tag SNPs, maximizes tolerance for missing SNPs, enlarges

*To whom correspondence should be addressed.

and balances detection power of each allele class. To resolve this multi-objective optimization problem, we incorporate a greedy initialization into two well-established multi-objective evolutionary algorithms, viz. the non-dominated sorting genetic algorithm-II (NSGA-II; Deb *et al.*, 2002) and the multiple single objective Pareto sampling (MSOPS; Hughes, 2003, 2007). The proposed algorithms can simultaneously find different sets of tag SNPs considering all objectives, which allow users to extract different solutions for accommodating various genotyping platforms. The proposed method is shown to explore larger search space and reduce convergence time compared to conventional methods. Experimental results revealed strong and weak conflicts among these objectives. Given the low missing and error rates of today's genotyping platforms, a small number of additional tag SNPs is able to provide sufficient tolerance and balance detection power of each allele.

The rest of this paper is organized as follows: Section 2 formulates the multi-objective tag SNPs selection problem. Section 3 sheds light on the proposed methods. Section 4 summarizes the experimental results and Section 5 discusses these results. Finally, conclusions are drawn in Section 6.

2 PROBLEM FORMULATION

This section formulates the multi-objective tag SNPs selection problem. A set of tag SNPs is defined as a set of SNPs that can distinguish any two allele classes. The following gives a formal definition of tag SNPs.

DEFINITION. (*Tag SNPs*): *Given a set of N SNPs* $\{S_1,...,S_N\}$ *and M allele classes* $\{P_1,...,P_M\}$. *Let* $P_{i,k}$ *denote the k-th element of allele class* P_i. *A set of tag SNPs* S^T *is a subset of S that can distinguish any two allele classes. That is, for any two allele classes* P_i *and* P_j, *there exists at least one tag SNP* $S_k \in S^T$ *such that* $P_{i,k} \neq P_{j,k}$.

The multi-objective tag SNPs selection problem is to select a set of tag SNPs that minimizes the total amount of selected SNPs, maximizes their robustness against missing data, maximizes the pairwise distance among allele classes, and minimizes the variance of these pairwise distances. These four objectives are formally defined below.

2.1 Compactness

The first objective aims to achieve the greatest compactness by minimizing the total number of tag SNPs; formally,

$$\min \|S^{\mathrm{T}}\|,$$

where $\|S^{\mathrm{T}}\|$ denotes the cardinality of set S^{T}.

2.2 Tolerance

This objective is to maximize the tolerance of selected tag SNPs for missing data. Let $D_{ij}(S^{\mathrm{T}})$ denote the set of tag SNPs in S^{T} that can distinguish allele classes P_i and P_j. The minimum cardinality of $D_{ij}(S^{\mathrm{T}})$ among all pairs of alleles gives the number, i.e. $\min(\|D_{ij}(S^{\mathrm{T}})\|)-1$, of missing SNPs that the set S^{T} of tag SNPs can tolerate (Huang *et al.*, 2005).

The second objective is then defined by

$$\max\left(\min_{i,j}\|D_{ij}(S^{\mathrm{T}})\|\right).$$

2.3 Dissimilarity

The present study attempts to generate dissimilar haplotype backgrounds for distinct allele classes. The similarity of haplotype backgrounds is measured by the Hamming distance at the selected tag SNPs. Let K^{T} be the index set of S^{T}, i.e.

$$S^{\mathrm{T}} = \bigcup_{k\in K^{\mathrm{T}}} S_k.$$

For two allele classes P_i and P_j, their Hamming distance is defined by

$$H(P_i,P_j) = \sum_{k\in K^{\mathrm{T}}} |P_{i,k} - P_{j,k}|.$$

The third objective is to achieve the maximum average Hamming distance over all pairs of allele classes, i.e.

$$\max \bar{H}$$

with

$$\bar{H} = \frac{1}{\binom{M}{2}} \sum_{0\le i<j\le M} H(P_i,P_j)\ .$$

2.4 Balance

In addition to dissimilarity, this study attempts to balance the detection power of tag SNPs for each allele class. This phenomenon can be captured by minimizing the variance of Hamming distances between all pairs of haplotype background defined by one tagging solution. Consequently, the resulting tag SNPs is unbiased to distinguishing alleles with sufficient sampling.

The objective is defined by

$$\min \mathrm{Var}(H)$$

with

$$\mathrm{Var}(H) = \frac{1}{\binom{M}{2}} \sum_{0\le i<j\le M} (H(P_i,P_j)-\bar{H})^2\ .$$

3 THE PROPOSED METHOD

The tag SNPs selection problem of satisfying a single objective is known to be NP-hard (Huang *et al.*, 2005). The increase of objectives makes tag SNPs selection even more intractable. To address this multi-objective optimization problem, this work presents a novel greedy initialization and integrates it into multi-objective genetic algorithms (GAs) based on NSGA-II (Deb *et al.*, 2002) and MSOPS (Hughes, 2003, 2007), respectively. GA has dealt successfully with various optimization problems. The basic idea of GA is mimicking the process of natural evolution—selection, reproduction and mutation—to manipulate candidate solutions (Holland, 1975). Based on the principle 'survival of the fittest', GA is expected to evolve candidate solutions toward the optima.

The algorithm NSGA-II is known for its effectiveness in dealing with optimization problems with two or three objectives. However, as the number of objectives increases, the utility of Pareto-ranking methods, like NSGA-II, deteriorates due to the augmented likelihood of non-dominance (Hughes, 2005; Ishibuchi *et al.*, 2008b; Kukkonen and Lampinen, 2007; Purshouse and Fleming, 2003; Wagner *et al.*, 2007). Hughes (2003, 2007) developed MSOPS to resolve this issue by assessing solutions with multiple ranks. In this article, we propose

a greedy initialization to deal with the issue at NSGA-II and to improve the performance of MSOPS.

The components and operators of the proposed multi-objective GAs are described in the following subsections.

3.1 Representation and fitness function

A candidate solution, viz. a set of selected tag SNPs, is encoded into a chromosome $\mathbf{c}=(c_1,\ldots,c_N)$, where gene $c_k \in \{0,1\}$ denotes whether SNP S_k is selected (value 1) or not (value 0).

The multi-objective tag SNPs selection problem considers compactness, tolerance, dissimilarity and balance. Given multiple objectives, this study adopts the notion of dominance for fitness evaluation. A chromosome **a** is said to *dominate* chromosome **b** if **a** is better than **b** in one objective and not worse than **b** in all other objectives. In this case, **a** is assigned a superior rank. If neither **a** nor **b** are dominated, the two chromosomes are said to *non-dominate* each other and are given the same rank. The Pareto front represents the set of solutions that are not dominated by any solutions.

This paper presents multi-objective evolutionary algorithms based on NSGA-II and MSOPS to address the tag SNPs selection problem. The NSGA-II algorithm determines whether a chromosome survives or dies according to two-level evaluation, namely, dominance rank and crowding distance. The former ranks chromosomes based on the dominance relation. For non-dominated chromosomes, NSGA-II further determines their ranks according to the distance (i.e. crowding distance) of the chromosome from its neighbors.

The MSOPS algorithm uses multiple target vectors to collect solutions. The target vectors divide the objective space as even as possible. Each chromosome is evaluated based on its closest vector. MSOPS gathers good chromosomes in each small space and accordingly approximates the Pareto front. This study utilizes the weighted min–max function in Hughes (2003) and ranks chromosomes using the scores from the aggregate function.

3.2 Initialization

The evolutionary process in GA begins with initialization of a set (population) of chromosomes. The initialization ordinarily generates chromosomes at random. Additionally, an ideal GA for multi-objective optimization problems should not require prior information about the Pareto front. However, in solving the four-objective tag SNPs selection problem, we found that conventional multi-objective GAs result in candidate solutions gathering around the center of objective space, caused by the easiness of being non-dominated for more than three objectives. An initialization based on greedy heuristic is, therefore, proposed to address this issue.

For the greedy initialization, this study defines the distinguishability of an SNP S_k as the number of allele pairs that can be distinguished by this SNP (see Fig. 1). For instance, distinguishability of S_1 is 15 in that this SNP can distinguish 15 pairs of alleles. Notably, the tag SNP selection problem can be solved by reduction to a variant of set cover problem (Huang *et al.*, 2005), which requires each element must be covered by a specified number of sets (termed coverage). For each specified coverage, a chromosome is constructed by the greedy initialization, which iteratively selects SNPs for maximum distinguishability until all elements are covered with the required coverage. Finally, the initial population is composed of chromosomes with respect to different coverage. This approach can achieve an initial population with diverse chromosomes promoting exploration of solution space.

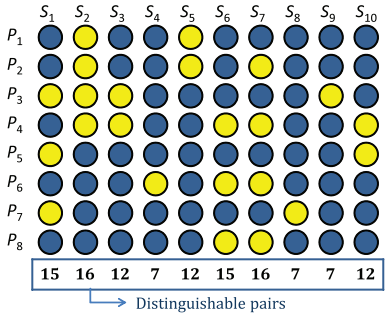

Fig. 1. An instance of calculating SNPs distinguishability.

3.3 Genetic operators

GA selects chromosomes as parents from the population and then performs crossover and mutation operations to generate their offspring. This study adopts the binary tournament selection in view of its accepted good performance. The binary tournament selection chooses the better of two random chromosomes as a parent. Performing this selection twice yields a pair of parents for reproduction, i.e. crossover and mutation.

The crossover operation exchanges and recombines the genetic information of parents, while the mutation operation slightly changes the composition of offspring. In this study, we apply the widely used uniform crossover (Syswerda, 1989) and bit-flip mutation for the proposed GA. The uniform crossover determines each offspring gene from either parent at random. The bit-flip mutation flips (i.e. $0\rightarrow 1, 1\rightarrow 0$) genes with a predefined probability called mutation rate p_m. That is, each gene has a probability of p_m to be flipped. By using crossover and mutation, candidate solutions can be recombined and changed.

After generating a set of offspring, GA applies the principal of 'survival of the fittest'. Restated, only the fittest chromosomes are selected to survive into the next generation. This article makes use of the above-stated strategies of NSGA-II and MSOPS for survival selection concerning multiple objectives.

4 EXPERIMENTAL RESULTS

The present study conducts a series of experiments to evaluate the proposed methods on the multi-objective tag SNPs selection problem. All experiments use the data of population haplotypes from Hinds *et al.* (2005), where the haplotypes are partitioned into blocks with limited diversity by (Halperin and Eskin, 2004). Each haplotype pattern in a block with >1000 SNPs is simulated as a haplotype background of one allele class. This study uses a block of 1032 SNPs in the experiments. Incidentally, genotype data can be processed by reconstructing their haplotypes through phasing programs such as PHASE (Scheet and Stephens, 2006; Stephens and Donnelly, 2003).

Table 1 summarizes the parameter setting for the adopted multi-objective evolutionary algorithms based on NSGA-II and MSOPS. The MSOPS uses 200 vectors dividing the objective space. To demonstrate the experimental results for four objectives, the solutions obtained are projected onto six 2D plains (see Figs 2 and 3), of which each axis represents one objective and arrows

indicate the optimization directions. Additionally, each figure plots the regression line of power model $y=ax^b+c$ to indicate the distribution of solutions found by our algorithm.

Figure 2 illustrates the experimental results of original NSGA-II and MSOPS. The solutions obtained from NSGA-II gather around the center; on the other hand, those from MSOPS spread over the objective space. Specifically, NSGA-II fails to identify solutions with fewer than 150 tag SNPs, whereas MSOPS can discover solutions that require as few tag SNPs as about 20 (see Fig. 2a and e). Similar trends also exist in other objectives. This is because four objectives cause solutions easy to be non-dominated and then hinder NSGA-II from exploring the objective space. This outcome implies that the distribution of the initial population significantly influences subsequent offspring in NSGA-II.

Next, we examine the effect of greedy initialization on NSGA-II and MSOPS. Figure 3 shows that greedy initialization can lead to broader distribution of solutions than that in Figure 2. Notably, the broadness (or diversity) of solutions is of paramount importance for multi-objective optimization problems in that it directly affects the flexibility for users to choose and adopt the solutions. In addition, owing to its preference for SNPs with high distinguishability, greedy initialization tends to increase the difference between haplotypes and drives solutions toward the optima for the third objective. These outcomes validate the advantages of the proposed greedy initialization in improving NSGA-II and MSOPS on the multi-objective tag SNPs selection problem.

Comparing improvement level, greedy initialization has a more significant influence on NSGA-II than on MSOPS. This effect indicates that a well-constructed initial population can direct subsequent exploration over the objective space and then overcome the issue of NSGA-II for more than three objectives. As for MSOPS, greedy initialization can also lead to better distribution of solutions. In addition, we found that some non-dominated solutions may be discarded by MSOPS since these solutions are out of their closest subspace and assigned lower ranks. To address this issue, this work maintains an archive for MSOPS to store the non-dominated solutions obtained.

Table 1. Parameter setting for the multi-objective evolutionary algorithms used in experiments

Parameter	Value
Representation	Binary string
Population size	200
Initialization	Random/greedy
Selection	Binary tournament selection
Crossover	Uniform crossover with $p_c=0.7$
Mutation	Bit-flip mutation with $p_m=1/l$
Termination	500 generations

5 DISCUSSION

5.1 General relation between objectives

This subsection discusses our major findings about the conflicts between the four objectives from experimental results.

5.1.1 Compactness and tolerance The first objective (compactness) is in conflict with the second one (tolerance) as expected. Experimental results show that the increase of tolerance requires additional tag SNPs. More precisely, the growth in the number of tag SNPs for tolerance is linear rather than quadratic or exponential (see Fig. 3a and e). With the advance of genotyping technologies, the number of missing SNPs in most

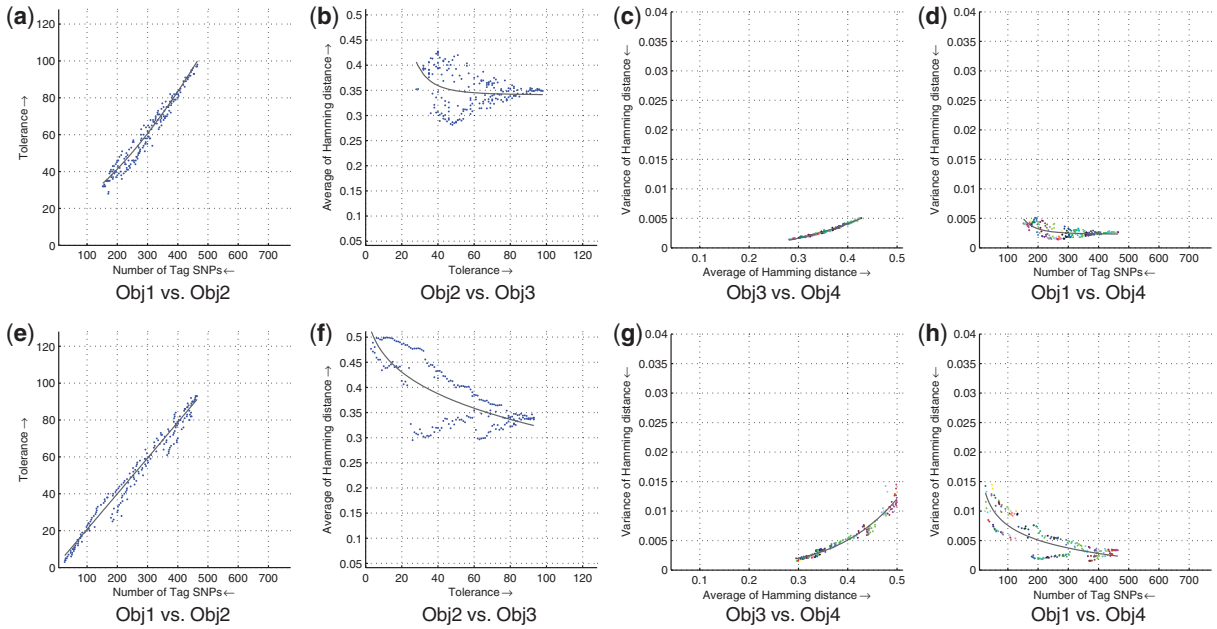

Fig. 2. Experimental results of NSGA-II (**a**–**d**) and MSOPS (**e**–**h**) for the multi-objective tag SNPs selection problem. Obj1 stands for the number of tag SNPs, Obj2 is the tolerance for missing data, Obj3 measures the average Hamming distance between alleles and Obj4 controls the variance of detection power in each allele.

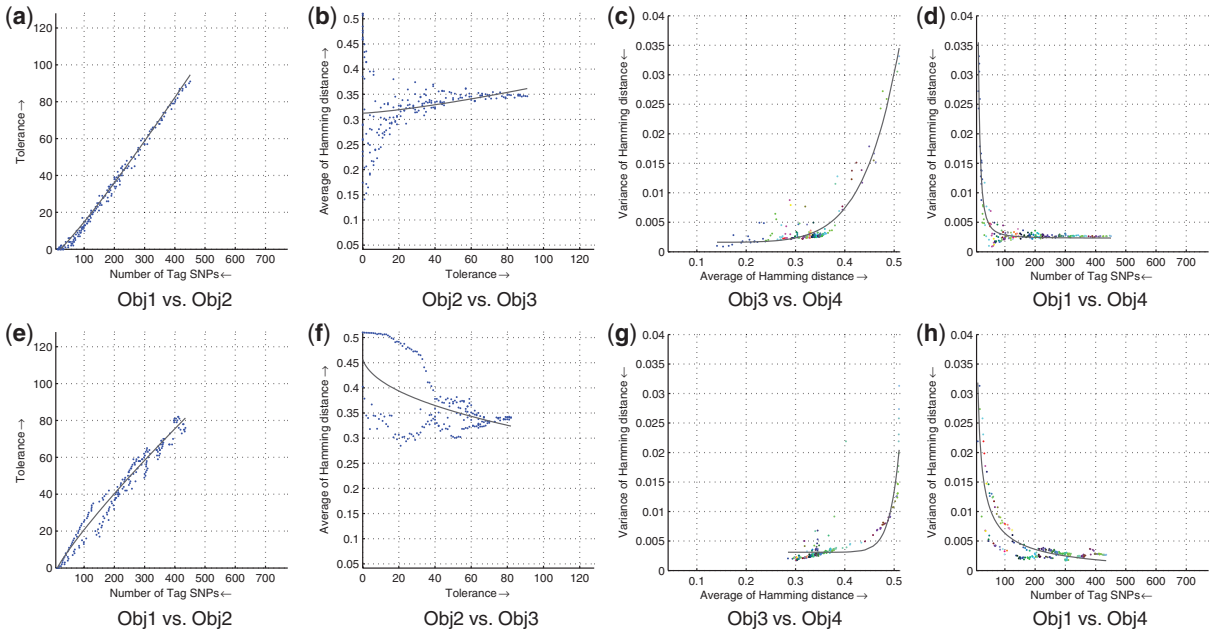

Fig. 3. Experimental results of NSGA-II (**a**–**d**) and MSOPS (**e**–**h**) using greedy initialization for the multi-objective tag SNPs selection problem.

platforms is limited. Therefore, the consideration of tolerance provides sufficient power for subsequent association studies at a low cost in the additional number of tag SNPs. For the platforms yielding high missing or error rates, however, it requires a relatively large number of tag SNPs that is directly proportional to the number of missing data to be tolerated.

5.1.2 Compactness and balance The first objective is in conflict with the fourth one (balance). That is, balancing the power of distinguishing each allele class requires more tag SNPs to be selected. The experimental results in Figure 3d and h indicate that increase of the number of tag SNPs greatly reduces the variance of hamming distances between each allele class. After having sufficient number of tag SNPs, the improvement of balancing becomes less significant. Therefore, a small increased amount of selected tag SNPs is able to maintain balance of detection power in each allele class.

5.1.3 Dissimilarity and balance The third objective (dissimilarity) and the fourth objective (balance) are in conflict (see Fig. 3c and g). A small average distance between allele classes comes with a low variance; on the other hand, as a high average distance is required, the variance between detection power is augmented as well. Notably, the variance grows exponentially with the increase of average distance.

5.1.4 Tolerance and dissimilarity The second (tolerance) and third (dissimilarity) objectives are in conflict (see Fig. 3b and f). This is because high tolerance genuinely requires more tag SNPs, but the average Hamming distance does not necessarily increase with the number of tag SNPs selected. Additionally, the non-dominated solutions with respect to these two objectives are quite diverse.

In summary, these results showed that some of these objectives are in weak conflict. Given that the missing and error rates of most genotyping platforms nowadays are quite low (<5%), the increase of tolerance and balance of detection power only require a small additional amount of tag SNPs. In addition, the proposed evolutionary algorithms provide a collection of candidate sets of tag SNPs for the users to select according to their requirements. This flexibility is helpful in selecting tag SNPs with respect to the constraints of various genotyping projects and platforms. Restated, the users can sort out and reduce the number of candidate sets by fixing the values or confining the bounds for certain objectives, e.g. the number of tag SNPs and missing rate of the platform. The decrease in number of candidate sets can be significant, especially when the spread of Pareto solutions is narrow.

5.2 Performance comparison

This study further adopts the performance measures proposed by Ishibuchi *et al.* (2008a, b) to assess performance of the proposed algorithms for the multi-objective tag SNPs selection problem. To this end, the second and fourth objectives are transformed into minimization objectives, and all objectives are normalized by the maximum value. All algorithms were tested five times for the average results.

5.2.1 Range (sum of the ranges of the objective values) This measure examines the *diversity* of solutions, where the range value means the difference between the highest and the lowest values of an objective. Figure 4a validates that the greedy initialization can improve MSOPS and, especially, NSGA-II in spreading the solution range.

5.2.2 SumMin (sum of the minimum objective values) This measure is for observing the convergence of solutions toward the *marginal region* of the Pareto front. It sums up the minimum values obtained for each objective. Figure 4b displays the SumMin

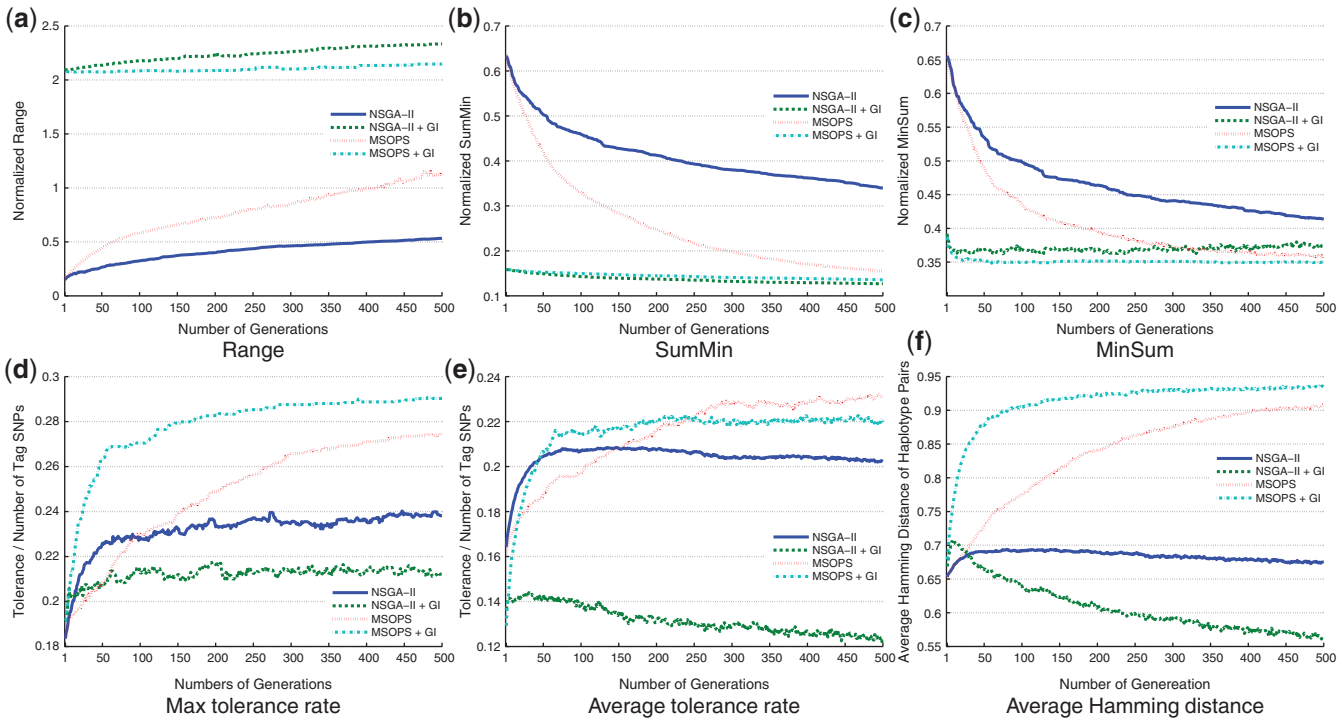

Fig. 4. Performance comparison of four test algorithms in terms of (**a**) range of objective values, (**b**) sum of minimum value of each objective, (**c**) minimum sum of each objective values, (**d**) maximum tolerance rate, (**e**) average tolerance rate and (**f**) average Hamming distance. GI: greedy initialization.

values for each algorithm. As aforementioned, greedy initialization generates chromosomes with possible minimum objective values. Using greedy initialization, both NSGA-II and MSOPS can then keep evolving these chromosomes to get refined solutions, whereas those without greedy initialization spend extra time in searching for promising regions and the optima.

5.2.3 MinSum (minimum sum of the objective values) This measure calculates the minimum sum of the four-objective values to assess the intensity that an algorithm converges toward the *central* region of the Pareto front. According to the experimental results in Figure 4c, the original MSOPS gradually exploits the solutions with minimum sum and achieves better results than NSGA-II does. On the other hand, NSGA-II using greedy initialization holds good solutions in the beginning but fails to further improve these solutions because the excessive number of non-dominated solutions hinders and deteriorates the development of the population toward the optima. In addition to solution quality, greedy initialization enhances MSOPS and NSGA-II in exploitation in the early phase of evolution and then reduces time consumption on the whole.

5.2.4 Tolerance rate The tolerance rate is defined as the tolerance divided by the number of selected tag SNPs. Its goal is to check the ability of a test algorithm in finding the tag SNPs set with high tolerance rate. The maximum and average tolerance rates are considered here. Figure 4d and e demonstrate that, regardless use of greedy initialization, MSOPS yields higher tolerance rate than NSGA-II does. Particularly, the greedy initialization helps MSOPS achieve high tolerance rate in the very beginning. As for NSGA-II, the greedy initialization decreases the tolerance rate in that the original NSGA-II tends to focus on the central region and then keeps the tolerance rate at the cost of broadness of solutions.

5.2.5 Average Hamming distance This measure considers the third objective only. It simply reflects how the average Hamming distance of haplotype pairs grows in each test algorithm, for which an increase of hamming distance in the population is desired. According to Figure 4f, the average Hamming distance for original NSGA-II does not increase during evolution. NSGA-II using greedy initialization even decreases the average Hamming distance, which is caused by its effort on extension of solution distribution. Notably, these broad non-dominated solutions may be very poor in terms of average Hamming distance but good for others such as tolerance and balance. By contrast, MSOPS performs well in this measure; the greedy initialization further benefits MSOPS in yielding a large average Hamming distance.

6 CONCLUSIONS

This study formulates the tag SNPs selection problem as a four-objective optimization problem concerning the number of tag SNPs, tolerance for missing data, dissimilarity among allele classes, and balance of detection power. To resolve this multi-objective optimization problem, we incorporate greedy initialization with multi-objective evolutionary algorithms based on NSGA-II and MSOPS. The proposed greedy initialization provides promising feasible solutions to impel the evolution of population and to reduce the convergence time.

Experimental results indicate that MSOPS outperforms NSGA-II, where the latter suffers from the excessive non-dominated solutions

for a four-objective optimization problem. The proposed greedy initialization substantially improves NSGA-II and MSOPS in terms of solution quality and efficiency. By using the greedy initialization, both NSGA-II and MSOPS can approach the optimal values for all objectives. In particular, MSOPS with greedy initialization achieves satisfactory results in several performance measures for the multi-objective optimization problem.

In summary, the formulation of multi-objective tag SNPs selection considers more than one aspect, which is more pertinent to real-world applications. The multiple non-dominated solutions obtained from the proposed evolutionary algorithms provide users with a great flexibility in selecting different sets of tag SNPs for different genotyping platforms and scenarios. Given the low missing and error rates of today's genotyping platforms, we found that a small number of additional tag SNPs can provide sufficient tolerance and balanced power for most genotyping projects.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers and the Associate Editor for their valuable comments and suggestions.

Funding: This work was supported by the National Science Council of Taiwan, under contract NSC98-2621-B-194-001.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al*. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Carlson,C. *et al*. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Chang,C.J. *et al*. (2006) A greedier approach for finding tag SNPs. *Bioinformatics*, **22**, 685–691.

Deb,K. *et al*. (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, **6**, 182–197.

Deng,L. *et al*. (2008) An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. *Hum. Mutat.*, **29**, 1209–1216.

Halperin,E. and Eskin,E. (2004) Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, **20**, 1842–1849.

Hinds,D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.

Holland,J.H. (1975) *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, Michigan.

Huang,Y.T. *et al*. (2005) Selecting additional tag SNPs for tolerating missing data in genotyping. *BMC Bioinformatics*, **6**, 263.

Hughes,E.J. (2003) Multiple single objective Pareto sampling. In *Proceeding of 2003 Congress on Evolutionary Computation*, Vol. 4, pp. 2678–2684.

Hughes,E.J. (2005) Evolutionary many-objective optimization: many once or one many? In *Proceedings of 2005 IEEE Congress on Evolutionary Computation*, IEEE Press, pp. 222–227.

Hughes,E.J. (2007) MSOPS-II: a general-purpose many-objective optimiser. In *Proceedings of 2007 IEEE Congress on Evolutionary Computation*, IEEE Press, pp. 3944–3951.

Ishibuchi,H. *et al*. (2008a) Effectiveness of scalability improvement attempts on the performance of NSGA-II for many-objective problems. In *Proceedings of 10th Genetic and Evolutionary Computation Conference*, ACM Press, pp. 649–656.

Ishibuchi,H. *et al*. (2008b) Evolutionary many-objective optimization: a short review. In *Proceedings of 2008 IEEE Congress on Evolutionary Computation*, IEEE Press, pp. 2419–2426.

Kukkonen,S. and Lampinen,J. (2007) Ranking-dominance and many-objective optimization. In *Proceedings of 2007 IEEE Congress on Evolutionary Computation*, IEEE Press, pp. 3983–3990.

Liu,W. *et al*. (2006) The impact of missing and erroneous genotypes on tagging SNP selection and power of subsequent association tests. *Hum. Hered*., **61**, 31–44.

McCarroll,S.A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.

Purshouse,R. and Fleming,P. (2003) Evolutionary many-objective optimisation: an exploratory analysis. In *Proceedings of 2003 IEEE Congress on Evolutionary Computation*, IEEE Press, pp. 2066–2073.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.

Syswerda,G. (1989) Uniform crossover in genetic algorithms. In *Proceedings of the 3rd International Conference on Genetic Algorithms*. Morgan Kaufmann, San Mateo, CA, pp. 2–9.

Wagner,T. *et al*. (2007) Pareto-, aggregation-, and indicator-based methods in many-objective optimization. In *Proceedings of the 4th International Conference on Evolutionary Multi-Criterion Optimization*, Vol. 4403 of LNCS, Springer, pp. 742–756.

Zhang,K. *et al*. (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.*, **14**, 908–916.

Zhao,J.H. *et al*. (2002) GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics*, **18**, 1694–1695.