Multiobjective Genetic Marker Selection

Robert Hubley[1], **Eckart Zitzler**[2], **Andrew F. Siegel**[1,3], **Jared Roach**[1]

1 Introduction

A genetic mapping project, typically implemented during a search for genes responsible for a disease, requires the acquisition of a set of data from each of a large number of individuals. This data set includes the values of multiple genetic markers. These genetic markers occur at discrete positions along the genome, which is a collection of one or more linear chromosomes. Typing the value of a marker in an individual carries a cost; one seeks to minimize the number of markers typed without excessively jeopardizing the probability of detecting an association between a marker and a disease phenotype.

The probability of detecting an association between a marker and a disease phenotype decreases with distance between the marker and the actual position of the gene responsible for the phenotype. Thus, one can maximize the probability of detecting disease linkage by choosing markers as closely spaced as possible.

In general, the decrease in probability of detecting association is not linear with distance; this probability tends to be relatively constant across patches of the genome known as "haplotype blocks". Thus one can save considerably on the cost of a mapping project by choosing no more than one marker from each haplotype block. Generally, the exact boundaries of haplotype blocks are not known prior to project execution, but it is often possible to assume that all haplotype blocks are of the same constant length s.

One typically searches for a marker-disease linkage within a given locus, or possibly set of locuses, of the genome. For purposes of this paper, a "locus" is any linear segment of the genome; in practice, a locus is typically fifty kilobases (kb) to several megabases (Mb). Each locus can be considered independently.

The locations of genetic markers are known prior to project initiation. In general, the number of known genetic markers exceeds the number necessary and/or affordable for a project. Thus, prior to project initiation, one is faced with the task of selecting a subset of markers from this initial library of markers. This paper presents algorithms for the solution of this selection task.

Markers in the library will have been previously characterized to a lesser or greater extent. A marker may be listed in error such that there is not in reality a marker at that position in the genome. A marker may be present in some genomes within a population of individuals, but so infrequently that it is of limited use for linkage studies. A marker may be at a position biochemically difficult to assay. A marker may be intellectual property, and have a licensing fee. All of these factors, which are either known or can be estimated, need to be considered during the selection of a set of markers from the library. We consider these factors weighted together as the "quality" of a marker. A "single nucleotide polymorphism" (SNP) is a particular kind of marker. SNPs are currently the most popular marker [1]. This paper assumes markers are SNPs, but extensions would permit inclusion of other markers.

One thus wishes to choose a set of SNPs from the library that meets the following criteria: (1) there is at least one SNP per haplotype block, to maximize the probability of detecting disease linkage, (2) there are as few SNPs as possible, to minimize cost, (3) the SNPs are uniformly distributed over the locus, and (4) the SNPs are of as high quality as possible, to minimize the need to choose between replacing a useless SNP or proceeding with a gap in the selected set.

2 Problem Formulation

The optimization problem described above can be stated formally as follows. Let n be the length of the locus under consideration and m be the number of available SNPs where each SNP i is described by two attributes:

$$\begin{aligned} p_i &= \text{position } (p_i \in \mathbb{N}, 1 \leq p_i \leq n) \\ q_i &= \text{quality } (q_i \in \mathbb{R}, q_i > 0) \end{aligned}$$

The attribute p_i denotes the position of the corresponding SNP. We here assume that the SNPs are ordered and unique according to the position, i.e., $p_1 < p_2 < \ldots < p_m$. The quality of a SNP is represented by a positive real number q_i; a larger value stands for higher quality. Furthermore, the length of the haplotype blocks is denoted as s. It can be considered as the optimal distance between two consecutive SNPs within a SNP selection.

A solution to the problem, a non-empty subset of the available SNPs, can be expressed in terms of m decision variables $x_i \in \{0, 1\}$ with $x_i = 1$ if and only if SNP i is in the subset. For convenience, we introduce in addition the variables x_0 and x_{m+1} which are by definition set to 1 and refer to two fictive SNPs that mark the left (position 0) and the right end (position $n + 1$) of the locus. Now, consider the deviation d_{ij} of the distance

[1]Institute for Systems Biology, Seattle, WA 98103-8904, USA

[2]Computer Engineering Laboratory, Swiss Federal Institute of Technology, Zurich, Switzerland

[3]Departments of Management Science, Finance, and Statistics, University of Washington, Seattle, WA 98195, USA

between two SNPs from the optimal distance s:

$$d_{ij} = (s - |p_j - p_i|) \cdot x_i \cdot x_j \cdot \prod_{k=i+1}^{j-1} (1 - x_k)$$

If the SNPs i and j are direct neighbors regarding the selected SNP subset, then d_{ij} gives the number of positions enclosed by them; otherwise, d_{ij} equals 0. Given this notation, the goal is to minimize the average deviation from the ideal gap length s

$$f_1(x_1, x_2, \ldots, x_m) = \frac{1}{1 + \sum_{i=1}^{m} x_i} \sum_{i=0}^{m} \sum_{j=i+1}^{m+1} d_{ij}$$

while maximizing the average quality

$$f_2(x_1, x_2, \ldots, x_m) = \frac{1}{\sum_{i=1}^{m} x_i} \sum_{j=1}^{m} q_i \cdot x_i$$

under the constraint that all gaps are less than or equal to s, i.e.,

$$\forall\, 0 \leq i \leq m\ \forall\, i < j \leq m+1\ :\ d_{ij} \leq s$$

For certain problems, it may not be possible to fulfill the constraint, i.e., there is a pair of SNPs i and j with $d_{ij} > s$ even if all m SNPs are selected. In this case, the problem can be divided into separate subproblems which can be solved independently. The algorithms presented in the following automatically take this into account.

3 Implementation

In this study, we use SPEA2 [2], a state-of-the-art multiobjective evolutionary algorithm, to approximate the *Pareto-optimal set* of this problem. The core implementation is based on ECJ8, a Java-based Evolutionary Computation and Genetic Programming Research System [3]. It has been substantially extended to implement SPEA2, which includes multiobjective specific operations such archiving, breeding selection, etc.

Each individual is encoded as a bitstring of length m, where the ith bit corresponds to the decision variable x_i. Individuals are recombined on the basis of the uniform crossover operator (the probability of recombination is set to 0.8) and modified according to a simple bitflip mutation using a mutation rate of $4/m$, following recommendations in [4]. Furthermore, a repair mechanism is incorporated that ensures each individual to be feasible. Whenever there are two SNPs i, j with $i < j$ such that $d_{i,j} > s$, then a SNP k with $d_{i,k} \leq s$ is inserted such that $d_{i,k}$ is maximum. The repair mechanism is only needed for calculating the objective values; the post-optimized bitstrings do not replace the original ones. Finally, the individuals in the initial population are created randomly such that in average the number of selected SNPs equals the number of haplotype blocks per locus ($\frac{n}{s}$). Both population and archive size are set to 250.

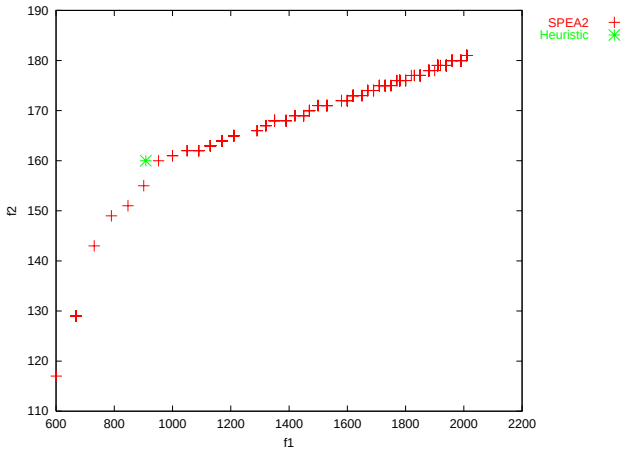

Figure 1: Alternative solutions generated by the evolutionary algorithm (represented by the plus symbol). The asterisk stands for the solution found by the heuristic.

As an alternative to compare with SPEA2, we developed a heuristic that generates a single solution to the problem by putting more emphasis on the first over the second objective. It first tries to choose SNPs of the highest quality such that the number of gaps that are greater than s is minimized. Afterwards, a local optimization step aims at improving the distribution of the chosen SNPs. If the resulting solution does not meet the constraint, then the above procedure is repeated for the SNPs of second highest quality, and so forth.

4 Results

The algorithms were tested for a target sequence that is a 90kb segment of the human major histocompatibility locus. The library contained 626 SNPs (cf. Fig. 2).

The trade-off front generated by the evolutionary algorithm after 200 generations is depicted in Fig. 1. The density of solutions increases as the first objective increases. This illustrates the structure of the solution space for this particular problem. The heuristic solution represents a trade-off that neither dominates nor is dominated by any SPEA2 solution, and is located in the middle of the front rather than on one of its extremes.

Fig. 2 shows some selected solutions. The one best in the first objective contains only 35 SNPs with an average quality of 117. The other extreme solution includes 85 SNPs and achieves an average quality of 181; basically all high quality SNPs are chosen and the large gaps are filled by SNPs of lower quality.

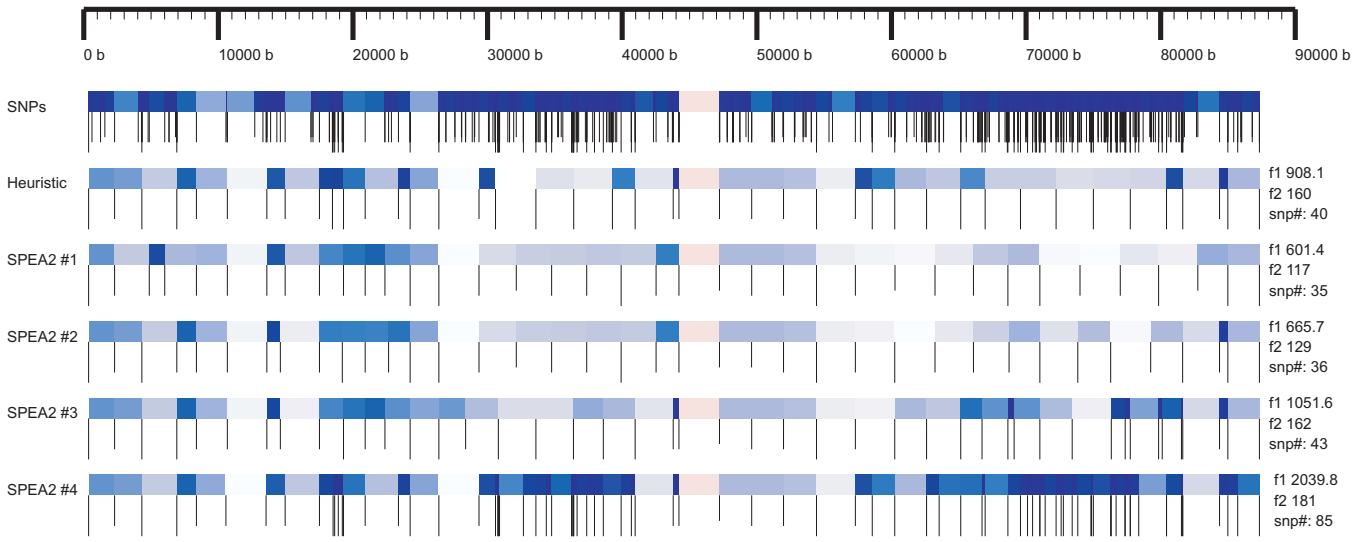

Figure 2: Comparison of heuristic solution with selected trade-offs generated by the evolutionary algorithm. The vertical lines represent the chosen SNPs with the height reflecting their quality. The top most tiling path on the figure is the total set of SNPs available. The shaded bar above each tiling path indicates varying degrees of spacing fitness. The darker the shading the farther the distance between the two SNPs is from the optimal distance s.

5 Conclusions

This new application demonstrates the usefulness of evolutionary algorithms in the presence of multiple optimization criteria. Firstly, an evolutionary-based approach allows generation of a set of trade-off solutions which provide additional information about the problem, such as the magnitude of the conflict between objectives, whether there are many or few potential solutions, and the structure of the search space. Knowing which alternatives are available can strengthen the confidence in the choice of a particular solution. Secondly, an evolutionary algorithm provides flexibility. Additional objectives and constraints can be incorporated with only little programming effort. For instance, in the future we may split the one quality objective into several, as stated in the introduction, such that a more accurate model is possible.

Finally, deterministic heuristics tailored to the application at hand often produce reasonably good results if sufficient problem knowledge is available. With the SNP selection problem, the solution found by the heuristic neither is dominated nor dominates any of the trade-offs generated by SPEA2. However, the design of these heuristics gets more difficult as more objectives are involved, and the single solution produced does not provide information about alternative solutions. In general, a promising way to tackle complex multiobjective optimization problems is to combine both approaches into a single algorithm.

References

[1] C. S. Carlson, T. L. Newman, D. A. Nickerson (2001): *SNPing in the human genome.* Curr Opin Chem Biol 5(1):78–85.

[2] E. Zitzler, M. Laumanns, and L. Thiele (2002): *SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization.* Evolutionary Methods for Design, Optimisation and Control with Application to Industrial Problems. Proceedings of the EUROGEN2001 Conference, Eds. K.C. Giannakoglou et al., pages 95–100, International Center for Numerical Methos in Engineering (CIMNE), Barcelona, Spain.

[3] S. Luke (2001): *An Evolutionary Computation and Genetic Programming System.* http://www.cs.umd.edu/projects/plus/ec/ecj/

[4] M. Laumanns, E. Zitzler, and L. Thiele (2001): *On The Effects of Archiving, Elitism, and Density Based Selection in Evolutionary Multi-Objective Optimization.* Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization (EMO 2001), Eds. E. Zitzler et al., pages 181–196, Springer-Verlag, Berlin, Germany.