Minimal Marker Sets to Discriminate Among Seedlines

Thomas C. Hudson, Ann E. Stapleton, Amy M. Curley
University of North Carolina at Wilmington
Departments of Computer Science and Biological Sciences
{hudsont, stapletona}@uncw.edu

Abstract

Raising seeds for biological experiments is prone to error; a careful experimenter will test in the lab to verify that plants are of the intended strain. Choosing a minimal set of tests that will discriminate between all known seedlines is an instance of Minimal Test Set, a NP-complete problem. Similar biological problems, such as minimizing the number of haplotype tag SNPs, require complex nondeterministic heuristics to solve in reasonable timeframes over modest datasets. However, selecting the minimal marker set to discriminate among seedlines is less complicated than other problems considered in the literature; we show that a simple heuristic approach works well in practice. Finding all minimal sets of tests to identify 91 Zea mays *recombinant inbred lines would require months of CPU time; our heuristic gives a result less than twice the minimal possible size in under five seconds, with similar performance on* Arabidopsis thaliana *recombinant inbred lines.*

1. Introduction

When a plant geneticist wants to conduct an experiment, she needs samples of a plant. Frequently, she will grow the plant herself from seeds kept in her laboratory. However, raising these plants is a labor-intensive, error-prone procedure: seeds can be wrongly sown, fields wrongly marked, natural pollination occur unintentionally, collected seeds mislabelled in the field or stored incorrectly in the lab. A cautious scientist will perform tests on the plants she takes her experimental samples from to confirm that they are from the intended seedline.

To verify the genotype of the sample, the scientist selects markers, extracts DNA from the sample plants, and amplifies each test region; these regions have known detectable differences in length. In the case of recombinant inbred lines, there are only two possibilities for each marker, conventionally referred to as size "A" and "B".

Our poster reports on heuristic algorithms developed to help minimize the expense of testing. Finding the optimum set of markers to use is a problem that can take months or years of CPU time; this software produces near-optimum answers in under a minute.

The algorithms discussed in our poster have been implemented in Java and are available under an open-source license at http://www.uncw.edu/csc/bioinformatics/.

2. Heuristic Solution

A randomized greedy algorithm gives a reasonable first answer for the problem of finding minimal marker sets to distinguish among the seedlines:

1. Shuffle the markers into a random order
2. Examine each marker in order
   (a) Remove it from the set of markers if the resultant set is still able to discriminate among all the seedlines

In our experiments on *Zea mays* (134 markers, 91 seedlines) and *Arabidopsis thaliana* (99 markers, 162 seedlines), this random greedy approach produces answers no more than twice the size of the theoretical optimum; repeated trials show that the results are roughly normally distributed (see our poster). If there are N seedlines and M markers, the theoretical complexity is $O(M^2N^2)$; the algorithm runs in seconds on those datasets.

These distributions imply that random sampling of the search space could yield reasonable results. The quality of the result of random sampling is very sensitive to the input: some subsets of the full data have many minimal-length answers, making random discovery likely, while others have only one. However, in practice they seem to have a large number of solutions requiring one marker more than minimal, which are reasonably likely to be found by random search. As problems grow larger – more seedlines are developed and more markers are identified – larger and larger

samples of the search space will be necessary to have a reasonable likelihood of finding a good solution.

Sorting according to simple metrics does not yield any improvement on random ordering, but provides consistency. Assigning a large negative value to a marker for every seedline about which the marker returns an inconclusive result gives a coarse ordering. If A and B appear with dissimilar frequency, adding a small positive value to the marker's rating for every seedline on which it returns the less-common result gives a finer ordering. Neither of these metrics outperforms random ordering; both typically give a result comparable to the median result returned in one thousand trials of random ordering. However, they do so in a single trial (under five seconds for both *Zea mays* and *Arabidopsis thaliana*), which gives us good input for the second stage of our algorithm.

We then filter the data. If the initial greedy heuristic returns a solution S containing K markers, we run the greedy algorithm K additional times. Let S_i be the i^{th} marker in S; on the i^{th} additional execution in this filtering pass, we remove S_i from the set of possible markers.

Whether we start with a random or sorted list of markers, running the basic greedy algorithm and then one pass of filtering gives us an answer of the same size as the best answers ever returned by the randomized algorithm. Additional passes of the filtering algorithm do not yield further improvement. For both *Zea mays* and *Arabidopsis thaliana*, this is roughly one point five times the length of the smallest possible answer.

In essence, this algorithm performs a heuristic search of the M-dimensional space of possible answers to find a candidate answer, and then exhaustively explores its K-dimensional immediate neighborhood looking for a local minimum. We find that the solution initially reported by the greedy heuristic is rarely a local minimum, but that it consistently has an adjacent local minimum. Over the data currently available, the two-stage approach gives reliably good results about a minute.

3. Exact Solution

A heuristic solution to the problem is not strictly necessary. The minimal discriminating set of markers can be found by examining all potentially discriminating sets. However, this requires an exhaustive search over a large search space.

For N seedlines and M markers, there are $\binom{M}{J}$ subsets of markers of size J. For each subset that we examine, a straightforward determination of whether the subset distinguishes between each pair of seedlines takes O(JN^2) time. The total predicted time is O($M^K N^2 K$), where K is the size of the minimal discriminating marker set; $K >= \log_2(N)$.

To verify this O() characterization, we implemented an exact solver for the minimal discriminating marker set problem and ran it over subsets of the *Zea mays* data. Graphs of the time performance of the exact solver can be found on the poster. A trial run of the exact solver on a dedicated 2.4 GHz Xeon CPU examined only 1.33% of the possible size-7 solutions for *Zea mays* in 17.5 CPU hours; if there is a size 7 answer, it would take us 54 days to find.

4. Theory and Context

Finding the minimal discriminating set of markers is an instance of a well-known NP-complete problem, Minimal Test Set [1]. In Garey and Johnson's formulation, the associated decision problem is:

> INSTANCE: A collection C of subsets of a finite set S, a positive integer $K \leq ||C||$.
> QUESTION: Is there a subcollection $C' \subseteq C$ with $||C'|| < K$ such that for each pair of distinct elements $u, v \in S$, there exists some set $c \in C'$ that contains exactly one of u and v?

[3] is a comprehensive survey of approaches to the Minimal Test Set problem.

This problem looks similar to another question intensely studied in bioinformatics, Haplotype Tag Selection. Although the decision problems are only subtly different, this difference significantly increases the complexity of algorithms that solve Haplotype Tag Selection. Approaches like ours to Minimal Test Set are not sufficient to solve Haplotype Tag Selection.

[2] is a survey of current work on Haplotype Tag Selection. The authors fit 21 published Haplotype Tag Selection algorithms into a three-stage framework: evaluating each SNP for how well it describes other nearby SNPs, evaluating a candidate set of SNPs for how well they classify the entire set of data, and constructing a final minimal set of SNPs. Our algorithm performs three analogous activities, albeit in a different order: filtering to minimize the set of results, sorting metrics, and a greedy minimization phase.

References

[1] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, San Francisco, 1979.

[2] B. V. Halldorsson, S. Istrail, and F. M. De La Vega. Optimal selection of snp markers for disease association studies. *Human Heredity*, 58:190 – 202, 2004.

[3] B. Moret and H. Shapiro. On minimizing a set of tests. *SIAM Journal of Scientific Computing*, 6(4):983 – 1003, 1985.