Clark's Genotype Phasing Algorithm

Elan Markowitz, 904485758

Overview

The attached genotype phasing software utilizes an implementation of Clark's Algorithm.

Clark's Algorithm

This algorithm is designed around the reasonable biologic assumption that only a relatively few haplotypes can explain the genetic variation in a population. It works as follows:
1. Let *H* be the set of candidate haplotypes for the population and *G* be the set of genotypes to be phased.

2. Find any genotypes that can be phased without ambiguity (One or fewer heterozygous sites) and add those haplotypes to *H*.

3. Go through *G* and find any genotypes that can be fully explained by *H*.

4. If any haplotype *h* in *H* is compatible with a genotype *g* in *G*, add the corresponding haplotype C(*h*, *g*) to H as the explanation for *h*. Continue until nothing more can be added.

5. If there are still remaining unexplained genotypes. Add a random pair of haplotypes that explain it to *H* and go back to step 4.

Implementation

There are a few problems with this algorithm that needed to be addressed in a full scale implementation. First, there are often no unambiguous genotypes in *G*. This is especially true as the number of SNP's increases. Second, as the number of SNP's increases, the algorithm becomes more likely to rely on randomly generated haplotypes causing accuracy to degrade.

To address the first problem, we find the genotype with the fewest heterozygous sites and generate a possible haplotype-pair explanation as the seed for *H*. We repeat this up to 2^k times and choose the result that uses the fewest haplotypes in its explanation. The hyperparameter *k* represents the number of heterozygous sites to check before randomly assigning the remainder.

To address the second problem, we must used a windowed approach. Rather than predicting the entire haplotype at once, we predict for a subset of the SNP's and then merge the results.

We choose successive windows of size *m* that overlap at *n* SNP's. We then run the predictions on each window. To merge the results, we look at the overlapping SNP's for a heterozygous site. If we find one, we use it to align the two windows so that that site is the same for both windows. Otherwise, we arbitrarily align them.

Hyperparameters

Our algorithm uses three hyperparameters: window size, overlap size, and seed depth. Tuning these parameters resulted in the following optimization. Window size of 35 and an overlap size of 5. The seed depth did not matter as long as it was sufficiently high. This is likely due to the fact that there will not be too many heterozygous sites when using a window size of 35.

Results

To measure results we used switch accuracy:

$Accuracy = \frac{n-1-sw}{n-1}$. Where *n* is the number of heterozygous sites and *sw* is the switching distance between the predicted results and the true results.

This implementation runs extremely quickly relative to more advance methods but still achieves an average accuracy of 84% on our training data.

Conclusion

While shy of the industry standard for accuracy, this is a very capable platform for rapid genotype phasing.