Search Space Reduction via Clustering for Haplotype Reconstruction

Jinghua Hu, [1] **Weibo Gong,** [2] **Patrick A. Kelly** [3]

Keywords: Haplotype Reconstruction, Genotype, Space Reduction, Clustering, Algorithm

1 Introduction.

The problem of haplotype reconstruction from phase unknown genotypes has been formulated into two major categories. One is the Haplotype Identification(HI) problem that identifies the haplotype set to explain the genotypes. Related algorithms include Clark's parsimony algorithm [1] and Gusfield's Perfect Phylogeny [2]. The other is the Haplotype Frequency Estimation(HFE) problem that seeks the optimal estimation on haplotype frequencies for resolving the genotypes. Related algorithms include EM-based algorithm [3], Gibbs sampling based algorithm [4], Partition-Ligation [5], etc.

One of the challenges for large scale haplotype reconstruction is the problem size that grows exponentially with the number of heterozygous sites in genotype. Divide-and-Conquer strategies have been applied in [5] and [6] to address the problem. Genotypes are broken into short segments where haplotyping is easily performed, and then the partial solutions are combined together to recontruct the full sequences.

This poster introduces our study on a new approach to problem size reduction for large scale haplotyping. The goal of this approach is to systematically reduce the haplotype search space while preserving the most possible solutions in the reduced space. The reduced search space will then serve as the starting point for existing algorithms, such as EM, to complete the haplotype inference.

The motivation is to examine the inter-correlations within the sample population. We present the framework of applying clustering techniques on genotype data based on the compatibility rules and proximity measures derived from genotype patterns. We examine the quality of clustering results through the reduced haplotype pools constructed from the clusters. Comparison is carried out on several performance indices, such as set coverage rate, size reduction ratio, etc. Results from different clustering algorithms, as well as results integrated from multiple clustering runs are collected for analysis.

Through experiments on simulated data sets, we demonstrate that our approach to search space reduction, in combination with existing algorithms, would enable us to handle large scale haplotyping problems more efficiently.

2 Search Space Reduction via Clustering.

The input of the algorithm is the genotype data matrix of size $N \times L$, where N is the population size, and L the length of genotype. The output is the haplotype pool constructed from the clusters, i.e., the *reduced pool*. The framework consists of three stages.

[1]Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA. E-mail: jhu@ecs.umass.edu

[2]Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA. E-mail: gong@ecs.umass.edu

[3]Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA. E-mail: kelly@ecs.umass.edu

1. Build compatibility and proximity matrices from input data.
2. Apply clustering algorithms on genotypes. Genotypes are assigned into clusters based on compatibility and proximity.
3. Construct haplotype pools from clusters by keeping the haplotypes shared by all members in a cluster. All partial pools are merged together as the final pool.

We adopt two indices for evaluating the quality of clustering results. Here the *original pool* refers to the haplotype set created by enumerating all possible haplotypes from genotypes, and the *truth pool* refers to the correct haplotype set.

- Set Coverage Rate: Percentage of haplotypes in the *truth pool* correctly preserved in the *reduced pool*.
- Size Reduction Ratio: Ratio between the size of the *original pool* and the size of the *reduced pool*.

The choice of proximity measures, clustering criteria, and clustering algorithms should aim at the goal of constructing reduced haplotype pools of high coverage rates as well as large reduction ratios.

3 Experimental Results.

The experiments are carried out on simulated data sets with the following observations.

- Efficient space reduction: A basic implementation of the sequential clustering algorithm combined with randomized assignment of genotypes would yield an average set coverage rate of above 80%, with an average size reduction ratio of 60 on data sets $L = 32, N = 32$.
- Merged pools perform better: Merged pools from multiple runs of clustering yield better quality at limited extra cost. Clustering algorithms may produce complementary results that work better together.
- The quality of clustering results also depend on input data and the settings of clustering algorithms. Further study on the relationship between input data patterns and computational complexity is desired.

4 References and bibliography.

References

[1] Clark, A. G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7:111-122.

[2] Gusfield, D. 2002. Haplotyping as Perfect Phylogeny: Conceptual Framework and Efficient Solutions. *Proceedings of 6th ACM International Conference on Computational Biology (RECOMB)*, ACM Press, pp. 166-175.

[3] Excoffier, L. and M. Slatkin. 1995. Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Molecular Biology and Evolution*, 12(5):921-927.

[4] Stephens, M., J. J. Smith and P. Donnelly. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. *American Journal of Human Genetics*, 68:978-989.

[5] Niu, T., Z. S. Qin, X. Xu and J. S. Liu. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157-169.

[6] Eskin, E., E. Halperin and R. M. Karp. 2003. Large Scale Reconstruction of Haplotypes from Genotype Data. *Proceedings of 7th International Conference on Computational Biology (RECOMB)*, pp. 104-113.