Constrained Hidden Markov Models for Population-based Haplotyping

Extended abstract

Niels Landwehr[1], Taneli Mielikäinen[2], Lauri Eronen[2], Hannu Toivonen[1,2], and Heikki Mannila[2]

[1] Machine Learning Lab, Dept. of Comp. Science, University of Freiburg, Germany
[2] HIIT Basic Research Unit, Dept. of Comp. Science, University of Helsinki, Finland

1 Introduction

Analysis of genetic variation in human populations is critical to the understanding of the genetic basis for complex diseases. Although genomes of several species have been sequenced, it is still too expensive to sequence genomes of several individuals to analyze genetic variation. Furthermore, most of the genome is invariant among individuals. A sequence representing (some of) the variant regions is called a *haplotype*. The positions in the sequence are called *markers* and the different possible values *alleles*. Most studied differences in DNA are at single-nucleotide locations. Such differences and locations are called *single nucleotide polymorphisms* (SNPs). In most of the cases only two alternative nucleotides (alleles) occur in the population.

In diploid organisms such as humans there are two *homologous* (i.e., almost identical) copies of each chromosome. The current measurement techniques produce a *genotype*—a sequence of unordered pairs of alleles—for each individual, instead of the two actual haplotypes. Two alternative approaches exist for inferring haplotypes from genotypes: If family trios are available, most of the ambiguity in the haplotype pair can be resolved analytically. If not, population-based computational methods have to be used to estimate the haplotype pair. Because trios are more difficult to recruit and more expensive to genotype, the population-based approach is often the only cost-effective method for large-scale studies. In this paper, we propose and evaluate a constrained Hidden Markov Model for population-based haplotyping.

2 Population-based Haplotype Reconstruction

A haplotype h is a sequence of alleles $h[i]$ in markers $i = 1, \ldots, m$. For brevity, we assume that $h \in \{0,1\}^m$, i.e., that the markers are SNP markers. A genotype g is a sequence of unordered pairs $g[i] = \{h_g^1[i], h_g^2[i]\}$ of alleles in markers $i = 1, \ldots, m$. Hence, $g \in \{\{0,0\},\{1,1\},\{0,1\}\}^m$. A marker with alleles $\{0,0\}$ or $\{1,1\}$ is *homozygous* whereas a marker with alleles $\{0,1\}$ is *heterozygous*.

Problem 1 (haplotype reconstruction). Given a multiset $\mathcal{G}$ of genotypes, find for each $g \in \mathcal{G}$ the most likely haplotypes h_g^1 and h_g^2 such that h_g^1 and h_g^2 are a *consistent* reconstruction of g, i.e., $g[i] = \{h_g^1[i], h_g^2[i]\}$ for each $i = 1, \ldots, m$.

If $\mathcal{H}$ denotes a mapping $\mathcal{G} \to \{0,1\}^m \times \{0,1\}^m$, associating each genotype $g \in \mathcal{G}$ with a pair (h_g^1, h_g^2) of haplotypes, the goal is to find the $\mathcal{H}$ that maximizes $\mathbb{P}(\mathcal{H} \mid \mathcal{G})$. It is usually assumed that the sample $\mathcal{G}$ is in Hardy-Weinberg equilibrium, i.e., that $\mathbb{P}((h_g^1, h_g^2)) = \mathbb{P}(h_g^1)\,\mathbb{P}(h_g^2)$ for all $g \in \mathcal{G}$, and that genotypes are independent of each other. With such assumptions, the likelihood $\mathbb{P}(\mathcal{H} \mid \mathcal{G})$ of the reconstruction $\mathcal{H}$ given $\mathcal{G}$ is proportional to $\prod_{g \in \mathcal{G}} \mathbb{P}(h_g^1)\,\mathbb{P}(h_g^2)$ if the reconstruction is consistent for all $g \in \mathcal{G}$, and zero otherwise. In population-based haplotyping, a probabilistic model λ for the distribution over haplotypes is estimated from the available genotype information $\mathcal{G}$, and $\mathbb{P}(h \mid \lambda)$ is then used to find the most likely reconstruction $\mathcal{H}$ for $\mathcal{G}$. Here, the relationship between haplotypes and the available genotype data is defined by Hardy-Weinberg equilibrium.

3 (Hidden) Markov Models for Haplotyping

We model the probability distribution on haplotypes by a left-right Markov Model λ, as shown in Figure 1. This is motivated by the observation that, due to recombination, dependency between markers (so-called linkage disequilibrium) decreases with increasing marker distance. Unfortunately, this model is not directly applicable in haplotyping, because in reality only genotypes are observed whereas the phasing information (the order of the allele pair) is hidden. The unobserved phase can be modeled by a Hidden Markov Model λ' as shown in Figure 2. A path through the model corresponds to sampling a pair of haplotypes, while the corresponding genotype is emitted. To reflect the Hardy-Weinberg equilibrium assumption, constraints have to be placed on transition probabilities. A transition in this model corresponds to independently sampling two new markers $h^1[t+1]$ and $h^2[t+1]$ for both modeled haplotypes based on their respective histories $h^1[t]$ and $h^2[t]$. Therefore, its probability is actually a product $\mathbb{P}(h^1[t+1] \mid h^1[t], \lambda)\,\mathbb{P}(h^2[t+1] \mid h^2[t], \lambda)$ where λ is the model on haplotypes outlined above. The advantage of this approach is that the model can be learned directly from genotype data using standard Baum-Welsh training, and the most likely reconstruction of a genotype can be obtained by the Viterbi algorithm [1]. For a different but related Hidden Markov modeling approach for haplotype reconstruction based on "founder" haplotypes see [2].

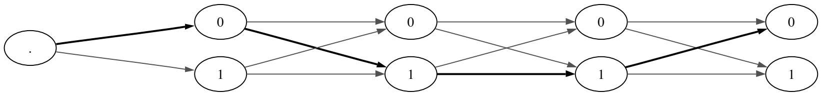

Fig. 1. A Markov Model over haplotypes. The highlighted path encodes the haplotype "0110".

The expressivity of the model can be increased by using a Markov Model of order $k > 1$ for the underlying haplotype distribution [3], but the number of

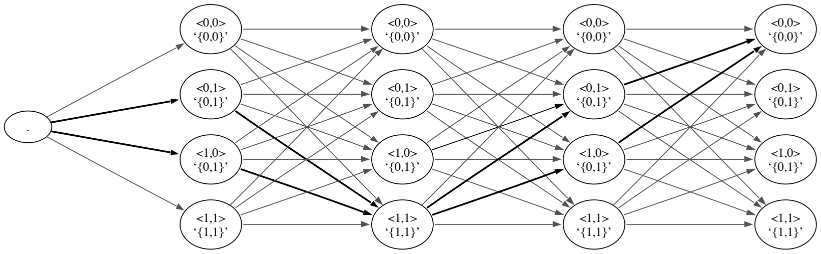

Fig. 2. A Hidden Markov Model over genotypes. Possible paths for genotype observation '$\{0,1\}$','$\{1,1\}$','$\{0,1\}$','$\{0,0\}$' are highlighted. The corresponding haplotype pairs are $\{(0100, 1110), (0110, 1100), (1100, 0110), (1110, 0100)\}$.

parameters increases exponentially in the history length k. Observations on real-world data (e.g., [4]) show that long histories are needed, but only few conserved fragments from the set of 2^k possible fragments actually occur. This can be exploited by modeling sparse distributions, where fragment probabilities which are estimated to be very low are set to zero. We propose a level-wise learning algorithm that constructs a sparse order-k Hidden Markov Model by iteratively refining models of order $i = 1, \ldots, k$. After training a model of order i, it is extended to a model of order $i + 1$, and at the same time transition probabilities $p < \epsilon$ are set to zero (and the corresponding transition is removed from the model). Conceptually, this is related to pattern-mining techniques such as Apriori [5], which identify long frequent patterns by extending shorter ones that are already known to be frequent. Also, the HaploRec system [3] uses frequent fragments to model haplotype distributions, although they are computed in a different way. At an abstract level, the learning algorithm can be illustrated using pseudocode as follows:

```
Initialize $i := 1$
$\lambda_1$ := initial-model()
$\lambda_1$ := em-training($\lambda_1$)
repeat
  $i := i + 1$
  $\lambda_i$ := extend-and-regularize($\lambda_{i-1}$)
  $\lambda_i$ := em-training($\lambda_i$)
until $i = k$
```

4 Experiments

The proposed method was implemented in a haplotyping system called SpaMM (for **Spa**rse **M**arkov **M**odeling) [3]. We compared its accuracy and computational

[3] The implementation is available from `http://www.informatik.uni-freiburg.de/~landwehr/haplotyping.html`

Table 1. Normalized switch distance on the Daly dataset, and average normalized switch distance for the datasets in the Yoruba-20, Yoruba-100 and Yoruba-500 dataset collections.

Method	**Yoruba-20**	**Yoruba-100**	**Yoruba-500**	**Daly**
PHASE	0.027	0.025	*n.a.*	0.038
fastPHASE-AVG	0.033	0.031	0.034	0.027
SpaMM	0.034	0.037	0.040	0.033
HaploRec	0.036	0.038	0.046	0.034
fastPHASE	0.041	0.060	0.069	0.045
HIT	0.042	0.050	0.055	0.031
GERBIL	0.044	0.051	*n.a*	0.034

performance to several other state-of-the art haplotype reconstruction systems: PHASE version 2.1.1. [6], fastPHASE version 1.1. [7], GERBIL as included in GEVALT version 1.0. [8], HIT [2] and HaploRec version 2.0. [9]. All methods were run using their default parameters. The fastPHASE system, which also employs EM for learning a probabilistic model, uses a strategy of averaging results over several random restarts of EM from different initial parameter values. This reduces the variance component of the reconstruction error and alleviates the problem of local minima in EM search. As this is a general technique applicable also to our method, we list results for fastPHASE with averaging (fastPHASE-AVG) and without averaging (fastPHASE).

The methods were compared using real data with known haplotypes inferred from family trios. Nontransmitted parental chromosomes of each trio were combined to form additional artificial haplotype pairs. Markers with minor allele frequency of less than 5% and genotypes with more than 15% missing values were removed.

We used a collection of datasets from the Yoruba population in Ibadan, Nigeria [10], and the well-known dataset of **Daly** et al. [4], which contains data from a European-derived population. For the Yoruba population, information on 3.8 million SNPs spread over the whole genome is available. We sampled 100 sets of 500 markers each from distinct regions on chromosome 1 (**Yoruba-500**), and from these we subsampled smaller datasets by taking the first 20 (**Yoruba-20**) or 100 (**Yoruba-100**) markers only for every individual. There are 60 individuals in the dataset after preprocessing, with an average fraction of missing values of 3.6%. For the Daly dataset, there is information on 103 markers and 174 individuals available after data preprocessing, and the average fraction of missing values is 8%. Although results on a single dataset are not very meaningful, the Daly dataset was included because is has been used often in the literature.

The accuracy of the reconstructed haplotypes produced by the different methods is measured by normalized switch distance. The switch distance is the minimum number of recombinations needed to transform the reconstructed haplotype pair into the original pair. For the datasets Yoruba-20, Yoruba-100 and Yoruba-500 the normalized switch distance is averaged over the individual datasets.

Table 2. Runtime (in seconds) of different methods averaged over 10 datasets each in the Yoruba-20, Yoruba-100 and Yoruba-500 dataset collection.

Method	**Yoruba-20**	**Yoruba-100**	**Yoruba-500**
PHASE	137	5088	∞
fastPHASE-AVG	47	242	1420
SpaMM	14	141	670
HaploRec	2	10	62
fastPHASE	22	111	548
HIT	2	13	143
GERBIL	2	122	∞

Table 3. Average error for reconstructing masked genotypes on Yoruba-100. From 10% to 40% of all genotypes were masked randomly.

Method	**10%**	**20%**	**30%**	**40%**
fastPHASE-AVG	0.045	0.052	0.062	0.075
SpaMM	0.058	0.066	0.078	0.096
fastPHASE	0.067	0.075	0.089	0.126
HIT	0.070	0.079	0.087	0.098
GERBIL	0.073	0.091	0.110	0.136

Table 1 shows the accuracy of the reconstructed haplotypes for the different methods and datasets. PHASE and Gerbil did not complete on the Yoruba-500 collection in two weeks[4]. Overall, the PHASE system achieves highest reconstruction accuracies. After PHASE, fastPHASE with averaging is most accurate, then SpaMM, and then HaploRec.

Table 2 shows average runtime of the methods for marker maps of different length. The most accurate method PHASE is also clearly the slowest, especially for long marker maps. fastPHASE and SpaMM are substancially faster, and HaploRec and HIT very fast. Gerbil is fast for small marker maps but slow for larger ones.

Finally, we ran experiments to test how well the different methods can infer missing genotypes based on the observed data. Table 3 shows the accuracy of inferring artificially masked genotypes, for different fractions of masked data.PHASE was too slow to run in this task, and HaploRec does not impute missing genotype values. Evidence from the literature [7] suggests that for this task, fastPHASE outperforms PHASE and is indeed the best method available. In our experiments, fastPHASE-AVG is most accurate. SpaMM is slightly less accurate than fastPHASE-AVG, but more accurate than any other method (including fastPHASE without averaging).

[4] All experiments were run on standard PC hardware with a 3.2GHz processor and 2GB of main memory.

5 Conclusions

We proposed a simple haplotype reconstruction method that is based on iterative refinement and regularization of constrained Hidden Markov Models. In our experimental study, PHASE was the most accurate haplotype reconstruction method, but also very slow. fastPHASE and SpaMM are slightly less accurate but much faster. In terms of both haplotype reconstruction accuracy and accuracy of missing genotype imputation (and also runtime), SpaMM lies in between fastPHASE with averaging and fastPHASE without averaging. The SpaMM method presented here is quite basic in that it does not use variance reduction techniques like averaging, nor is the model fine-tuned (e.g. using priors) to the haplotyping problem. Nevertheless, it offers a competitive trade-off between accuracy and computational complexity compared to other state-of-the-art systems developed for this task.

Acknowledgements The authors would like to thank Luc De Raedt, Kristian Kersting and Pasi Rastas for helpful discussions. The research was supported by the European Union IST programme, contract no. FP6-508861, *Application of Probabilistic Inductive Logic Programming II.*

References

1. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE **77** (1989) 257–286
2. Rastas, P., Koivisto, M., Mannila, H., Ukkonen, E.: A Hidden Markov Technique for Haplotype Reconstruction. In Casadio, R., Myers, G., eds.: WABI. Volume 3692 of Lecture Notes in Computer Science. Springer (2005) 140–151
3. Eronen, L., Geerts, F., Toivonen, H.: A Markov Chain Approach to Reconstruction of Long Haplotypes. In Altman, R.B., Dunker, A.K., Hunter, L., Jung, T.A., Klein, T.E., eds.: Pacific Symposium on Biocomputing. World Scientific (2004) 104–115
4. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander, E.S.: High-Resolution Haplotype Structure in the Human Genome. Nature Genetics **29** (2001) 229–232
5. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Databases. (1994) 487–499
6. Stephens, M., Scheet, P.: Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation. The American Journal of Human Genetics **76** (2005) 449–462
7. Scheet, P., Stephens, M.: A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics **78** (2006) 629–644
8. Kimmel, G., Shamir, R.: A Block-Free Hidden Markov Model for Genotypes and Its Applications to Disease Association. Journal of Computational Biology **12** (2005) 1243–1259
9. Eronen, L., Geerts, F., Toivonen, H.: HaploRec: Efficient and Accurate Large-Scale Reconstruction of Haplotypes. Manuscript (2006)
10. The International HapMap Consortium: A Haplotype Map of the Human Genome. Nature **437** (2005) 1299–1320