

HaploBlock Version 1.2

LD Mapping Supplement

Gideon Greenspan
Computer Science Department
Technion – Israel Institute of Technology

gdg@cs.technion.ac.il

April 28, 2004

1 Mapping using models

A high density LD mapping study is based on a list $\mathcal{H} = \{h^1, \dots, h^n\}$ of n phased haplotypes or a list $\mathcal{G} = \{g^1, \dots, g^n\}$ of n unphased genotypes over the entire region of interest. We use the symbol \mathcal{D} to refer to input \mathcal{H} or \mathcal{G} as appropriate. The other inputs are a list $\mathcal{P} = \{p^1, \dots, p^n\}$ of phenotypes associated with each haplotype or genotype and the distances d_j in base pairs between adjacent SNPs j and $j+1$ over $j = 1 \dots l-1$. For haplotype mapping, each haplotype h^i is a string of l symbols from the set B of SNP alleles, where l is the number of loci examined. For genotype mapping, each genotype g^i is a string of l elements from the set D of unordered SNP allele pairs. Each p^i is in the range $1 \dots p_{max}$ where p_{max} is the total number of phenotypes observed. In a simple case-control study, $p_{max} = 2$.

We are searching for an unobserved genetic locus within the candidate region that affects the phenotypes observed. Let L_j denote the hypothesis that this locus is situated in the interval between SNPs j and $j+1$, so that we consider the set of hypotheses $\{L_1, \dots, L_{l-1}\}$. We express the output of a mapping study as a posterior distribution $Pr(L_j|\mathcal{P}, \mathcal{D})$ over these alternatives, normalized so that $\sum_{j=1}^{l-1} Pr(L_j|\mathcal{P}, \mathcal{D}) = 1$. This distribution is calculated in the following four stages.

First, we infer an ensemble \mathcal{M} of statistical models which are locally optimal in terms of the MDL criterion, i.e. those which provide a compact explanation of the observed data \mathcal{D} . We ignore the phenotypes \mathcal{P} during this process, since they barely affect the data likelihood. We explore the search space of models using Gibbs-style iterations, in which the existence and location of each block divider constitute the variable for resampling. The initial model has dividers distributed evenly over the region. During a sampling iteration, each of the dividers in the current model is removed in turn to create a larger block, into which we attempt to add up to 3 new dividers at optimal locations, so long as this improves the MDL score. All model parameters are optimized at each stage of this process, using the local search and modified EM algorithms described previously [2].

Second, for each model M in the ensemble \mathcal{M} , we calculate the posterior probability that each block contains the phenotypic locus. Let U_k denote the hypothesis that the locus is in block k of M . The posterior distribution $Pr(U_k|\mathcal{P}, \mathcal{D}, M)$ is calculated using the method described in section 1.1 or 1.2 as appropriate. Note that at this stage the phenotype data is used to assess hypotheses relating to blocks, rather than SNP intervals, since each model inferred assumes that the alleles within each block segregate together.

Third, the posterior distribution $Pr(U_k|\mathcal{P}, \mathcal{D}, M)$ over the blocks in model M is converted into a posterior $Pr(L_j|\mathcal{P}, \mathcal{D}, M)$ over SNP intervals. For an interval $(j, j+1)$ in block k , for which $s_k \leq j < e_k$, we allocate the posterior in proportion to the length d_j of the interval, setting $Pr(L_j|\mathcal{P}, \mathcal{D}, M) = \frac{d_j}{V_k} Pr(U_k|\mathcal{P}, \mathcal{D}, M)$, where V_k is the total length of block k . For an interval $(j, j+1)$ on the boundary between blocks k and $k+1$, for which $j = s_{k+1} - 1 = e_k$, we assume that half of the interval lies

in each block, setting $Pr(L_j|\mathcal{P}, \mathcal{D}, M) = \frac{d_j}{2V_k}Pr(U_k|\mathcal{P}, \mathcal{D}, M) + \frac{d_j}{2V_{k+1}}Pr(U_{k+1}|\mathcal{P}, \mathcal{D}, M)$. The block length V_k is obtained by summing the interlocus distances d_j within the block and half of those at either end, i.e. $V_k = \sum_{j=s_k}^{e_k-1} d_j + \frac{1}{2}(d_{s_k-1} + d_{e_k})$. Note that V_1 and V_b lose elements d_{s_k-1} and d_{e_k} respectively from this sum, where b is the number of blocks in the model.

In the fourth and final stage, the individual posterior distributions $Pr(L_j|\mathcal{P}, \mathcal{D}, M)$ obtained from each model M in the ensemble \mathcal{M} are combined into a single statistic by uniform model averaging, so that $Pr(L_j|\mathcal{P}, \mathcal{D}) = \frac{1}{|\mathcal{M}|} \sum_{M \in \mathcal{M}} Pr(L_j|\mathcal{P}, \mathcal{D}, M)$. We use a uniform prior for the averaging since the sampling process has already introduced a strong bias towards models with a low MDL score.

1.1 Haplotypes posterior

Recall that hypothesis U_k states that the phenotypic locus is located in block k of a model. Under Bayes' Rule, the posterior probability of hypothesis U_k is given by $Pr(U_k|\mathcal{P}, \mathcal{H}, M) = \frac{Pr(\mathcal{P}|U_k, \mathcal{H}, M)Pr(U_k|\mathcal{H}, M)}{Pr(\mathcal{P}|\mathcal{H}, M)}$. Since $Pr(\mathcal{P}|\mathcal{H}, M)$ is the same for all k and we assume that the prior $Pr(U_k|\mathcal{H}, M)$ does not depend on \mathcal{H} , this can be rewritten as:

$$Pr(U_k|\mathcal{P}, \mathcal{H}, M) \propto Pr(\mathcal{P}|U_k, \mathcal{H}, M)Pr(U_k|M) \quad (1)$$

In this equation, $Pr(U_k|M)$ is the prior probability that block k of model M contains a locus which affects the observed phenotypes, while $Pr(\mathcal{P}|U_k, \mathcal{H}, M)$ is the posterior probability of phenotypes \mathcal{P} given haplotypes \mathcal{H} under that assumption.

Phenotype information is expressed as the variable P in our model. Under hypothesis U_k , P is directly dependent only on variable C_k , as depicted in Figure 1. This simple dependency is sufficient because the differences in ancestry reflected by variable C_k capture the ancestral variation at all loci within block k , including those which are not observed.

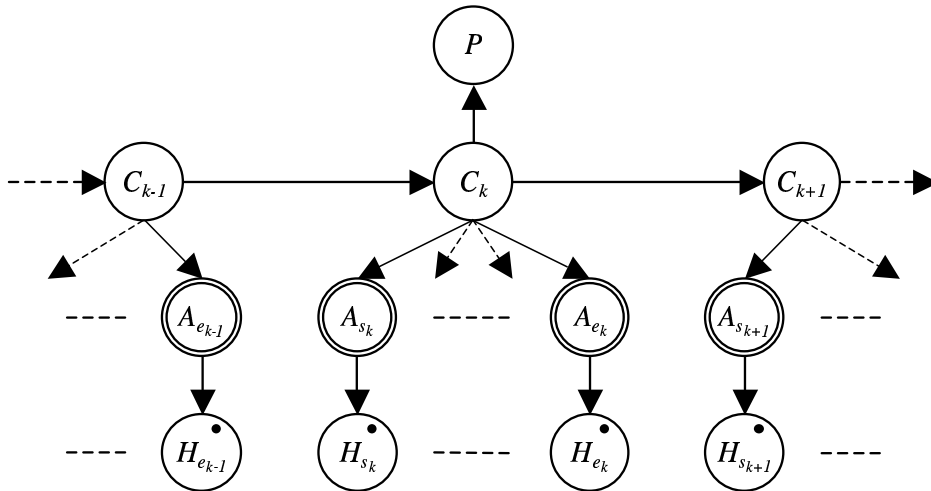


Figure 1: Bayesian Network for modeling haplotype data.

We approximate the term $Pr(\mathcal{P}|U_k, \mathcal{H}, M)$ of Equation 1 by assuming sample independence and inferring maximum likelihood parameters for $Pr(P|C_k, M)$. These parameters are obtained using the EM algorithm with the haplotypes \mathcal{H} and phenotypes \mathcal{P} as evidence [4]. The subsequence of each haplotype for block k is usually compatible with only one value of C_k , so the EM algorithm converges uniquely and quickly.

The prior probability $Pr(U_k|M)$ of Equation 1 is based on two elements. The first element assigns probability in proportion to V_k , the length of block k . The second element adjusts for the fact that blocks with more ancestors have more parameters for maximizing the likelihood

$Pr(\mathcal{P}|U_k, \mathcal{H}, M)$. We compensate by considering the optimal number of bits W_k required to represent $Pr(\mathcal{P}|C_k, M)$. Using a standard encoding, $W_k = \frac{q_k}{2}(p_{max} - 1) \log_2 n$, where q_k is the number of ancestors for block k , p_{max} is the number of phenotypes and n is the number of samples observed [7]. Applying the MDL schema, elements V_k and W_k are combined to obtain $Pr(U_k|M) \propto V_k \cdot 2^{-W_k}$ [6].

1.2 Genotypes posterior

For genotype data, the posterior distribution $Pr(U_k|\mathcal{P}, \mathcal{G}, M)$ is obtained in a similar manner as for haplotypes. Equation 1 is trivially rewritten as:

$$Pr(U_k|\mathcal{P}, \mathcal{G}, M) \propto Pr(\mathcal{P}|U_k, \mathcal{G}, M)Pr(U_k|M) \quad (2)$$

As before, we represent phenotype information as the variable P in our model. For dominant, recessive and codominant disease models, the phenotype is affected by genetic variation in both chromosomes. Therefore, under hypothesis U_k , P depends on both variables C_k and C'_k , as depicted in Figure 2. The differences between haplotype and genotype posterior calculations stem only from this more complex dependency.

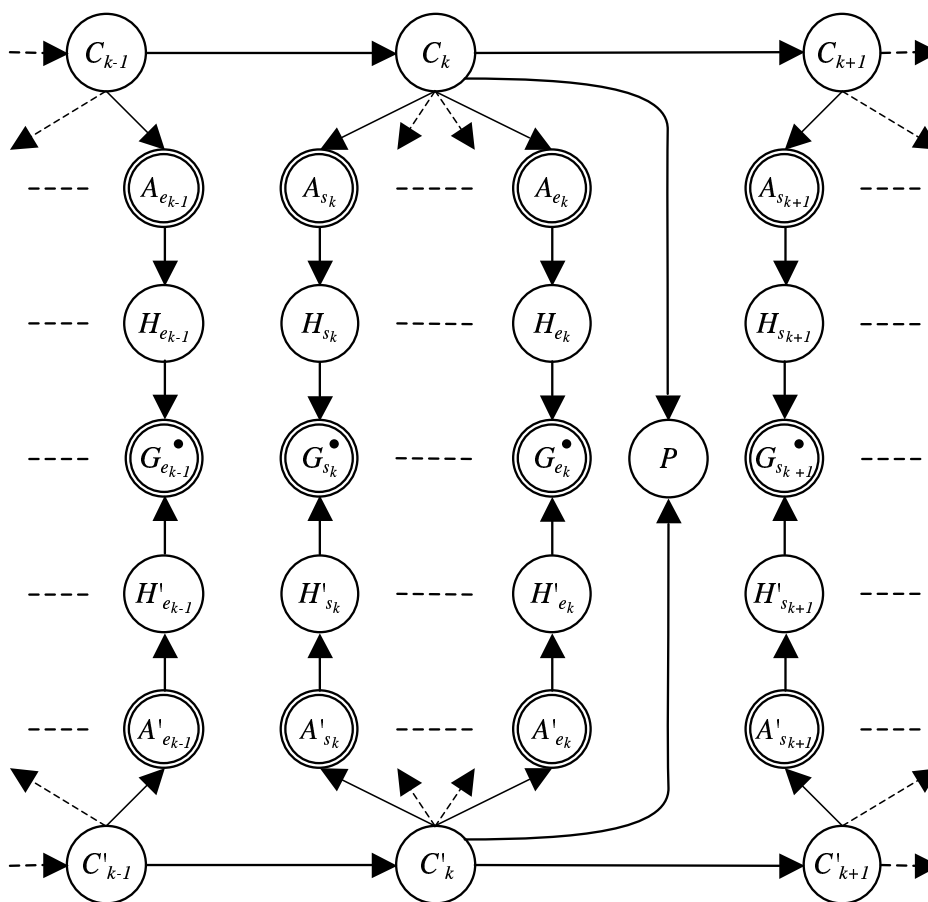


Figure 2: Bayesian Network for modeling genotype data

Element $Pr(\mathcal{P}|U_k, \mathcal{G}, M)$ of Equation 2 is calculated as before by assuming sample independence and inferring the parameters of $Pr(\mathcal{P}|C_k, C'_k, M)$ by EM. This distribution is symmetrical for the two variables C_k and C'_k , reflecting the functional symmetry between the maternal and paternal chromosomes in a cell.

The prior probability $Pr(U_k|M)$ of Equation 2 is also calculated as before, based on the length V_k and the number of bits W_k required to represent $Pr(\mathcal{P}|C_k, C'_k, M)$. Since the distribution

$Pr(P|C_k, C'_k, M)$ is symmetrical, we set $W_k = \frac{q_k \cdot (q_k + 1)}{4} (p_{max} - 1) \log_2 n$. The two elements are combined as before so that $Pr(U_k|M) \propto V_k \cdot 2^{-W_k}$.

References

- [1] Dechter,R. (1996) Bucket elimination: A unifying framework for probabilistic inference. In *Proc. Twelfth Conf. on Uncertainty in Artificial Intelligence (UAI-96)* pp. 211–219.
- [2] Greenspan,G. and Geiger,D. (2003) Model-based inference of haplotype block variation. In *Proc. Seventh Annual Inter. Conf. on Computational Molecular Biology (RECOMB 2003)* pp. 131–7.
- [3] Jensen,F. (1996) *An Introduction to Bayesian Networks*. Springer Verlag, New York, New York.
- [4] Lauritzen,S. (1995) The EM algorithm for graphical association models with missing data. *Comp. Stat. Data Analysis*, **19**, 191–201.
- [5] Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, California.
- [6] Rissanen,J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–471.
- [7] Rissanen,J. (1983) A universal prior for integers and estimation by minimum description length. *Ann. Stat.*, **11**, 416–431.