Unsupervised haplotype reconstruction and LD blocks discovery in a hidden Markov framework

A. Perina[1], M. Cristani[1], G. Malerba[2], L. Xumerle[2], V.Murino[1], and P.F. Pignatti[2]

[1] Dipartimento di Informatica,
Università degli Studi di Verona,
Strada le Grazie 15, 37134 Verona, Italia.
{perina,cristanm,murino}@sci.univr.it
[2] Dipartimento Materno Infantile e di Biologia-Genetica,
Università degli Studi di Verona,
Strada le Grazie 8, 37134 Verona, Italia.
{luciano.xumerle,giovanni.malerba,pignatti}@medgen.univr.it

Abstract. In the last years *haplotype reconstruction* and *haplotype blocks discovery*, *i.e.*, the estimation of patterns of linkage disequilibrium (LD) in the haplotypes, riveted the attention of the computer scientists due to the involved strong computational aspects. Such tasks are usually faced separately; recently, statistical generative techniques permitted to solve them jointly. Following this trend, we propose a generative framework based on hidden Markov processes, equipped with two novel inference strategies. The first strategy estimates finely haplotypes, while the second provides a quantitative measure to estimate LD blocks boundaries. Comparative real data results validate the proposed framework.

1 Introduction

Estimating haplotype[3] frequencies becomes increasingly important in the mapping of complex disease genes, as large numbers of closely linked single nucleotide polymorphisms (SNPs) can be genotyped. SNPs are single base pair differences between individuals in a population. Association studies work on the premise that SNP genotypes are correlated with a disease phenotype. Numerous studies have shown that human genome contains regions of high *linkage disequilibrium* (LD) with low haplotype diversity[1]: these regions are called *haplotype blocks* or *LD blocks*, where LD is a non-random association of alleles between adjacent loci. It is worth noting that SNPs or haplotype in LD blocks may serve as proxy for causative alleles: therefore, an accurate study on the blocks diversity became a key factor in genome wide association studies [2]. Unfortunately, allele phase of multilocus genotype in unrelated individuals is unknown and haplotypes needs to be reconstructed [3], before the discovery of haplotype blocks [4].

In this paper, we propose a statistical framework aimed at the simultaneous haplotype reconstruction and block discovery. Simultaneous statistical strategies have been recently introduced [5]: the idea is to perform the two operations iteratively, providing temporary solutions (reconstructed haplotypes and blocks) which can be re-evaluated until a global data fitness criteria is met. Our framework is based on a hidden Markov setting similarly to [5], drawn here more

[3] *Haplotypes* are combinations of DNA marker alleles in a single chromosome.

correctly in terms of connection between fully non homogeneous hidden Markov models (FNH-HMM); differently to what carried out before, we do not add any a-priori knowledge (such as family data for reconstruction or block boundary hotspots) because this knowledge is not always recoverable. Most important, we introduce a simple way to reconstruct haplotypes and a novel inference to robustly individuate blocks. The idea is to first estimate from data relevant hidden "ancestral" patterns, i.e. allele patterns which represent high frequency haplotypes fragments. In this way, reconstructed haplotypes can be realized as the most probable path among these ancestral patterns, mimicking biological theories [3]. Frequent splits and joins among paths indicates block boundaries. The proposed strategy is compared with state-of-the-art methods and applied on real data; biological results attest the goodness of the strategy.
The paper continues as follows: Sec.2 gives preliminary notions; Sec.3 explains our framework and Sec.4 shows experimental results and draws some conclusions.

2 Preliminaries

2.1 Fully non homogeneous hidden Markov model

Let us suppose to have a set $\mathbf{O}$ of J mono-dimensional observation sequences $\{\mathbf{O}_j\}, j=1,\ldots,J$, of length N, formed by symbols from a finite vocabulary V. Formally, a FNH-HMM (depicted in Fig.1a) is a set $\mathbf{\Theta_h}$ of (hidden) parameters $(\{\mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\pi}\})$, *i.e.*, a site-dependent transition matrix $\mathbf{A}_k = \{a_k^{mn}\}$ with $a_k^{mn} = P(S_{k+1} = n | S_k = m)$, $1 \leq m, n \leq L$ and $k = 1, \ldots N$; a site-dependent emission matrix $\mathbf{B}_k = \{b_k^m(v)\}$ where $b_k^m(v) = P(v|S_k = m), v \in V$ and an initial state distribution $\boldsymbol{\pi} = \{\pi_n\}$.
Assuming the learning of a HMM as known, we propose the learning of a FNH-HMM as a modified version of the Baum-Welch algorithm (BW) [6] considered here as specialization of the Expectation-Maximization (EM) iterative procedure [7]. In the FNH-HMM learning, the E-step consists in first calculating the standard forward and backward variables, paying attention that all the transition and emission probabilities involved are site dependent (*i.e.*, dependent on k). From these variables key quantities can be obtained, such as the conditional probability of two consecutive hidden states in an observation sequence at site k, *i.e.*, $P(S_k = m, S_{k+1} = n | \mathbf{O}_j) = \xi_{k,j}(m,n)$ and the conditional $P(S_k = m | \mathbf{O}_j) = \sum_{n=1}^{L} \xi_{k,j}(m,n) = \gamma_{k,j}(m)$. In the M-step the parameters are update using these quantities. For our interest, the transition $\mathbf{A}_k$ and the emission $\mathbf{B}_k$ matrices are updated as follows:

$$a_k^{mn} = \frac{\sum_{j=1}^{J} \xi_{k,j}(m,n)}{\sum_{j=1}^{J} \sum_{n=1}^{L} \xi_{k,j}(m,n)} \qquad b_k^m(v) = \frac{\sum_{j=1}^{J} \gamma_{k,j}(m)}{\sum_{j=1}^{J} \sum_{n=1}^{L} \xi_{k,j}(m,n)} \qquad (1)$$

Differences with respect to the HMM framework are that here the statistics are collected for each site k individually, i.e. no summation over k is present.

3 The proposed model: connection between FNH-HMMs

In our framework, $\mathbf{O}$ is formed by J observation *samples*. Each sample represents the genotype of the j-th human subject, *i.e.*, a sequence of N allele pairs; each

k-th pair, $k = 1, \ldots, N$, is formed by unordered variables $\{x_k, y_k\}$ taking values from $\{A, C, G, T\}$.
Assuming that in the samples every k-th SNP takes two symbols from two *hidden ancestral patterns*, we instantiate two independent state variables s_k and t_k that represent the k-th sites of such patterns [4]. These variables take pattern indexes values $1, \ldots, L$ by considering a first-order Markov property, i.e. considering the states s_{k-1} and t_{k-1} (Fig.1c, step 1, *Pattern choice*)[5]. Then, to each state is associated a probability of emission of a particular nucleotide symbol x_k and y_k (Fig.1c, step 1, *Symbols emission*). Now, in order to simulate the allele phasing

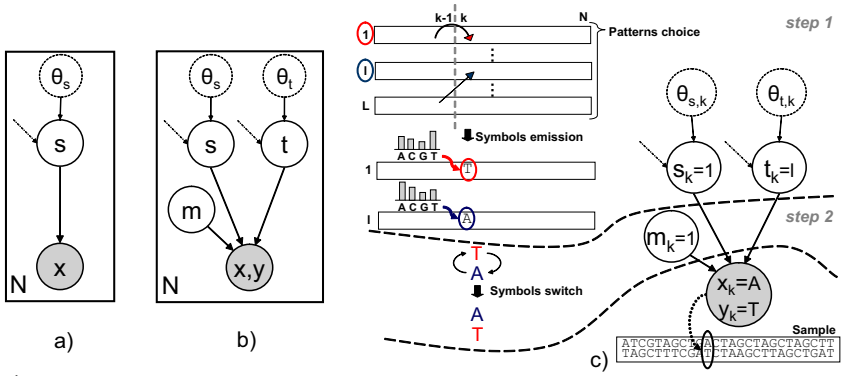

Fig. 1. a) FNH-HMM and b) FNH-HMM *double net*: nodes in a solid box indicate that they are replicated the number of times indicated in the bottom left corner; point-dashed arrows mean 1-st order Markov dependency. Filled (unfilled) circles mean observed (unobserved) random variables; dotted circle indicate the parameter set of the variable linked by the arrow; c) SNPs generative process: the picture is divided in two steps. Each step shows a portion of the process, drawn in an intuitive fashion (left) and in a formal graphical way (right).

that produces the final SNP pair, we add a switch variable m_k that decides the order of the alleles (Fig.1c, step 2, *Symbols switch*). We call this model FNH-HMM *double net*.
The joint distribution of the model over the samples is $\prod_{j=1}^{J} P(\{x_k, y_k, m_k, s_k, t_k\})$, $k = 1, \ldots, N$. Its factorization mirrors formally the above mentioned generative process, with a simplification done to make the learning process tractable. First of all, it is reasonable to consider samples as i.i.d. generated; so, in the following, we consider only the joint distribution P over a single sample, which can be written as:

$$P = P(x_1, y_1 | m_1, s_1, t_1) P(m_1 | s_1, t_1) P(s_1) P(t_1) \cdot \prod_{k=2}^{N} P(x_k, y_k | m_k, s_k, t_k) P(m_k | s_k, t_k) P(s_k | s_{k-1}) P(t_k | t_{k-1})$$

Here we note that each sample is considered as formed by two independent fully non homogeneous hidden Markov processes of states s_k and t_k, coupled at the level of the emission probability $P(x_k, y_k | m_k, s_k, t_k)$ plus the presence of the phasing distribution $P(m_k | s_k, t_k)$. The emission distribution can be further factorized, making clear the meaning of the switching variable $m_k \in \{0, 1\}$, which

[4] In the rest of the paper, we use indistinctively the terms *states* or *patterns*.
[5] Choosing the "right" L is an unsolved issue in this context; driven by biological issues, and setting $7 \leq L \leq 15$, very similar results have been achieved.

determines the phase of the chromosome pair $\{x_k, y_k\}$. If $m_k = 1$, the state s_k (t_k) generates symbol x_k (y_k), viceversa if $m_k = 0$; in formulae this becomes

$$P(x_k, y_k | m_k, s_k, t_k) = (P(x_k|s_k)P(y_k|t_k))^{m_k}(P(y_k|s_k)P(s_k|t_k))^{1-m_k} \quad (2)$$

Finally, in order to ease the learning step, $P(m_k|s_k, t_k) = P(m_k)$, *i.e.*, we assume the switching variable only dependent on the sample site k.

The learning step, performed by a generalized EM, consists in iteratively evaluating for each k the parameters of 1) the switch distributions $P(m_k)$, which permit to estimate haplotypes; 2) the emission distributions $P(\cdot|s_k)$, $P(\cdot|t_k)$, and the transition distributions, which are useful to estimate haplotype blocks. In Eq.1 we show how to find the transition and the emission parameters; in [5], Sec.3.1, the update of $P(m_k)$ is shown.

After the model learning, the haplotype reconstruction strategy consists in evaluating for each sample $\mathbf{O}_j$ the related probability values of the masks $P(\{m_k\})$. If at site k $P(m_k) < 0.5$, then the input order of the allele couple is $< x_k, y_k >$, otherwise it is switched. This provides two haplotypes for each genotype $\mathbf{O}_j$.

As written above, the hidden patterns $1, \ldots, L$ model ancestral haplotype sequences which are fragmented and blocks-recombined in the human history, producing all the observed haplotypes. As first step toward the block discovery, we estimate, for each reconstructed haplotype sequence, the most probable pathway through these hidden patterns. This is done with a non-homogeneous version of the Viterbi algorithm applied on the learned model, designed with the same intuition used in the learning step (*i.e.*, paying attention to use site dependent transition and emission parameters). All the Viterbi paths are then

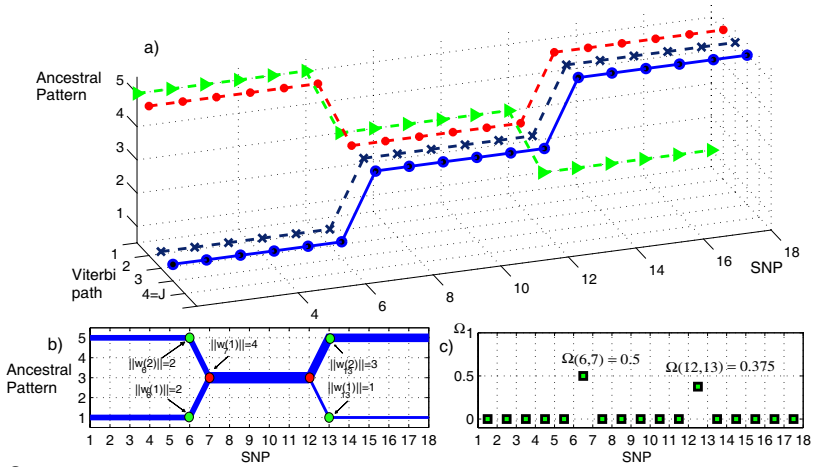

Fig. 2. Toy example (J=4 haplotypes): a) Viterbi paths b) paths over the lattice structure; note that b) is projection of a) over the SNPs-"Pattern Number" plane. c) Ω plot: between sites 6 and 7 there is a boundary "stronger" than the one present between sites 12 and 13.

disposed on a lattice $L \times N$ (Fig.2a). In this way, at each allele site k we can distinguish W_k distinct paths, each one of them indicated with $w_k(i), i = 1, \ldots, W_k$; $\| w_k(i) \|$ indicates the number of haplotypes traversing $w_k(i)$ (see Fig.2b).

We are now able to perform blocks discovery. The idea is that if two paths $w_k(i)$

and $w_k(i')$ do join, they represent two sets of haplotypes which have highly different haplotype fragments until k, becoming similar after site k; therefore, a block boundary exists between k and $k+1$. Similar reasoning holds for a split site (see Fig.2b). We translate this intuition with the *boundary presence strength* measure $\Omega(k, k+1) \in [0, 1)$, which models the existence of a block boundary between sites k and $k+1$, which is

$$\Omega(k, k+1) = \mathbf{1}_{\text{Join}}(k)G(k) + \mathbf{1}_{\text{Split}}(k+1)G(k+1) \tag{3}$$

where $\mathbf{1}_{\text{Join}}(k)$ ($\mathbf{1}_{\text{Split}}(k+1)$) equals 1 when a join (split) is present at time k ($k+1$), and $G(\cdot)$ is the *Gini* index [8]

$$G(k) = 1 - \sum_{i=1\ldots W_k} \left(\frac{\parallel w_k(i) \parallel}{W_k} \right)^2 \tag{4}$$

Gini index can be used to describe whether a graph join or split is well balanced or not. For example, a split at site k is well balanced if the cardinalities $\{\parallel w_k \parallel\}$ of the child paths $\{w_k\}$ are similar; the idea is that the higher is $\Omega(k)$, the more likely is the presence of a block boundary between site k and $k+1$; viceversa, a low $\Omega(k)$ means that in the join (split) site k, a dominant path (*i.e.*, with a high number of haplotypes associated) merges (splits) with one ore more irrelevant paths (see Fig.2b). Given a threshold τ_Ω we can assign a block boundary to the site k when $\Omega(k) > \tau_\Omega$; in all the experiments we set $\tau_\Omega = 0.2$.

4 Experimental results

Our framework has been tested on different data sets; here we report two explicative tests. For what concerns the initialization, no a-priori knowledge has been used, *i.e.*, for every site k, transition matrices $\{\mathbf{A}_k\}$ have been initialized to favor staying in the same state($a_k^{ii} > 0.5$) while mask distributions have been initialized uniformly to 0.5.

The first data set is taken from the HAPMAP project (`www.hapmap.org`) on chromosome 7 from SNP marker *rs323917* to SNP *rs324375*. In table 1 we show haplotype reconstruction results. Please note that our approach obtains results comparable with `Phase` [3], which is the best algorithm for haplotype reconstruction. Its computational complexity ranges from $O(N^2)$ to $O(N^3)$, while our method is $O(L^2N)$. Moreover, Phase is built on a (visible) Markov model of variable order, and does not perform blocks discovery, while our method is built on a (hidden) first-order Markov model and it performs blocks discovery. In this sense, we believe that augmenting the order of the (hidden) Markov process can improve the overall performances.

Blocks discovery results are shown in Fig.3; in Fig.3b the Pairwise LD table [4]

Table 1. Haplotype frequencies obtained with a training set composed by 60 genotypes of 25 SNPs. FNH-HMM *double net* reports the mean values over more than 100 experiments.

Haplotypes	Haplotype Frequencies		
	Ground truth	Phase	FNH-HMM *double net*
C A C G C C C T A T G T T A G A C T C A G G T T A	0.475000	0.4760	0.475000
C G T A T G C T A T G T C G G A C T C T A A C A A	0.183333	0.1833	0.176655
C G T A T G C T A T G T C G G A C T C A A G C T A	0.116667	0.1167	0.109999
G A C G T C C T A T G T C A G A C T C A A G C A A	0.066667	0.0667	0.066667
C A C G C C C C A T G T T A G A C T C A G G T T A	0.058333	0.0583	0.056754
C A C G T C C T A T A T C G G A C T C A A G C T A	0.041667	0.0407	0.039054
C G T A T G C C A T G T C G G A C T C T A A C A A	0.025000	0.0250	0.018000
C A C G C C C T A T G T T A G A C T T A G G T T A	0.016667	0.0157	0.014444
C G T A T G C T A T A T C G G A C T C A A G C T A	0.008333	0.0083	0.007566
C A C G C C C T A T G T T A G A C C C A G G T T A	0.008333	0.0083	0.006454
C A C G T C C T A T A T C G G A C T T A A G C T A	-	0.0010	-

is reported[6], where the pairwise measures D' (Fig.3b - left diagonal elements) and r^2 (Fig.3b - right diagonal elements) summaries the Linkage Disequilibrium in the region; high D' or r^2 values indicate in position m, n a block relation between the site m and n. The table, built using exact reconstructed haplotypes with a-priori knowledge, confirms our results.

The second data set used consists of 11 SNPs taken from interlukin-1 cluster on

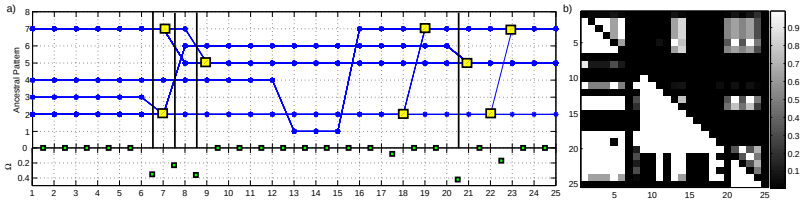

Fig. 3. a) Viterbi paths (top) and correspondent Ω plot (bottom). Splits/Joins are indicated with yellow rectangles; block boundaries are shown with a bar; b) pairwise LD table: D' (left diagonal elements) and r^2 (left diagonal elements) values confirm block boundaries found with our method.

human chromosome 2q12-2q14 presented in [9]. In figure 4 are depicted all the paths over the ancestral patterns inferred after the model training. No splits or joins are present, thus only a haplotype block is present here, as confirmed by a-priori knowledge on the data. Haplotype reconstruction results are optimal, but not reported here due to the lack of space.

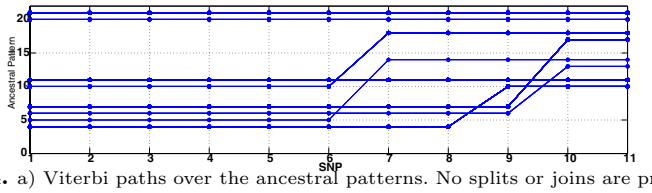

Fig. 4. a) Viterbi paths over the ancestral patterns. No splits or joins are present.

References

1. Gabriel, S.B., et al.: The structure of haplotype blocks in the human genome. Science **296** (2002) 2225–2229
2. Zhang, K., et al.: Haplotype block structure and its application to association studies: Power and study designs. Am. J. Hum. Genet. **71** (2002) 1386–1394
3. Stephens, M., Donnelly, P.: A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. **73** (2003) 1162–1169
4. Chen, Y., Lin, C.H., Sabatti, C.: Volume measures for linkage disequilibrium. BMC Genetics (**7**)
5. Jojic, N., et al.: Joint discovery of haplotype blocks and complex trait associations from snp sequences. (In: Proceedings of the UAI-04)
6. Rabiner, L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. of IEEE **77** (1989) 257–286
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B **39** (1977) 1–38
8. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001)
9. Gohlke, H., Illig, T., et al.: Association of the interleukin-1 receptor antagonist gene with asthma. Am J Respir Crit Care Med **169** (2004) 1217–1223

[6] Pairwise LD table is a widely used method for block-discovery, that needs *exact* haplotypes to accurately estimate blocks, *scarcely* robust to reconstruction errors.