EFFICIENT ALGORITHMS FOR HAPLOTYPE PHASING WITH RFLPS

W. CASEY, B. MISHRA
Courant Institute Math Science, 251 Mercer St, NYC, NY-10012.
E-mail: {*wcasey,mishra*}*@courant.nyu.edu*

The determination of feature maps, such as STSs, SNPs or RFLPs maps, for each chromosome copy or *haplotype* in an individual has important potential applications to association studies. We present a method to recover RFLP feature maps for each haplotype starting from *genotype* data which is an ambiguous superposition of all haplotypes' data. Our method is an inference method which is able to interpret data in two key ways: 1) We determine when a feature expresses *polymorphic* diversity across the haplotypes, 2) We co-associate the alternatives of each pair of polymorphic feature thereby partitioning the genotype feature map into haplotype feature maps. We design an expectation maximization (EM) algorithm to detect the polymorphic markers. Secondly, we design an efficient algorithm to rapidly determine the co-associations of alternatives for each pair of polymorphic features: This process is called the *phasing* of polymorphims. The problem of optimizing existing methods of SNP (single nucleotide polymorphism) phasing have been investigated in [3] and found to be NP-hard. In contrast, using the RFLP (restriction fragment length polymorphism) markers, we show that our algorithm can produce marker phasing and hence haplotypes, when the genome-wide ordered restriction site data are produced by an available technology such as optical mapping [1]. A prior model of the data, comprising a set of *restriction fragment lengths*, allows us to analyze the proposed algorithm and provide a probabilistic guarantee for its correctness. Our algorithm can be suitably modified for a wide class of haplotyping problems, relying on unrelated markers and technologies. Independently, as a significant fraction of RFLP markers are directly caused by SNP's, the RFLP phasing may be an important tool for reducing the complexity of the SNP-phasing problem.

1 Introduction and Related Literature

A diploid organism contains two very similar copies of each chromosome, with the exception of sex-chromosomes. We call the pair of copies *haplotypes*, and refer to them individually as "Haplotype I" and "Haplotype II". In this paper, we discuss a restriction-enzyme-based experiment and algorithms capable of uncovering the diversity across the haplotypes.

A restriction enzyme (e.g., *Bam*H I) cuts double-stranded DNA at specific recognition sites (e.g. 'ggatcc') with high specificity. Thus, slight positional variations of these restriction sites on the mostly similar copies of a chromosome can potentially separate and identify the haplotypes. For instance, a segmental insertion or deletion between two consecutive restriction sites on one copy of the chromosome will be observed as a difference in the two restriction fragment lengths. This event represents one of two distinct causes that result in the underlying restriction site marker patterns on two haplotypes to differ—the other being an "in-del" or substitution of a nucleotide within the actual restriction site. Thus the second kinds of RFLPs represent a significant subset of SNP's (single nucleotide polymorphisms). While RFLPs have not received as wide an attention as SNPs, they also hold the same kinds of promise as SNPs for association studies and classification of genetic diseases—most likely for a significantly lower cost.

The base-pair length between two consecutive *restriction sites* are called *restriction fragment lengths* and are modeled as random variables, studied in [1]. When the homologous regions on the two haplotypes contain different lengths between consecutive restriction sites, they are said to result in a 'restriction fragment length polymorphism' (abbreviated, *RFLP*). These RFL's are subject to measurement errors (e.g., sizing error, partial restriction digestion and false positive site errors) and locally confound length-based polymorphism detection. Making haplotype maps directly is much more difficult than making genotype maps. Nonetheless, if a group of ordered restriction fragments can be sampled from either haplotype, a large number of such samples allows first to identify RFLPs, and then determine how these RFLPs co-associate locally. Furthermore, with the increase in the number of such samples and the increase in expected number of markers in each sample, it is possible to improve the accuracy and resolution of the haplotype maps in spite of the statistical errors alluded to. Note that, at the end of this process, the result can be interpreted as two haplotype ordered restriction maps, further annotated with the location of the RFLPs.

Similar haplotyping or 'the phasing problem' has been investigated in [2,?], but with a different data model. Of particular interest are the results of NP-hardness in [3] of a SNP phasing problem. Detailed statistical treatments of Optical mapping are found in [1]. The paper is organized as follow: In section 2, we reviews *Optical Mapping*. Section 3, discusses the recognition of polymorphic sites from data using an *EM-Algorithm* for parameter estimation for a mixture distribution. In section 4, we discuss the phasing problem mathematically. We provide simulations in section 6.

In the full paper we give an analysis for the approximate algorithm, and detailed mathematical proofs for the propositions whose proofs are omitted in this presentation. The inital results on reasonable date sets are encouraging.

2 Optical Mapping

Optical mapping is a physical mapping approach that provides an ordered enumeration of the restriction sites along with the estimated lengths of the restriction fragments between consecutive restriction sites. A *restriction site* is the location of a short specific nucleotide sequence (4-8 bp long) where a particular restriction enzyme cleaves the DNA by breaking a phosphodiester bond. The fragment of DNA generated by cleaving at two consecutive restriction sites is a restriction fragment.

The physico-chemical approach underlying optical mapping is based on immobilizing long single DNA molecules on an open glass surface, digesting the molecules on the surface and visualizing the gaps created with fluorescence microscopy. Thus the resulting image, in the absence of any error, would produce an ordered sequence of restriction fragments, whose masses can be measured via relative fluorescence intensity and interpreted as fragment lengths in bps. The corrupting effects of many independent sources of errors affect the accuracy of an optical map created from one single DNA molecule, but can be tamed statistically by combining the optical maps of many single molecules covering completely or partially the same genomic region and by exploiting accurate statistical models of the error sources. To a rough approximation the resolution and accuracy of an optical map can be arbitrarily improved by simply increasing the number of enzymes and number of molecules involved.

2.1 Parameters of the experiment

We consider a set of M fragments of average length L which cover the genome of length G with coverage $c = \frac{ML}{G}$. On this genome we have a set of N restriction sites. Each molecule is a contiguous region from one of two haplotypes, and contained on the molecule are some restriction sites. Each molecule provides a local view of the ordered restriction sites taken from one haplotype. For each of the restriction sites found on a molecule we have data for position in the interval $[1, G]$. For the sake of simplicity, in this paper, we assume that non-digestion rates are negligible, and that distance data may be scaled to a consensus map so that positional data may be understood.

The consensus map may be represented by an ordered set of lengths. After the construction of a consensus map all of the observed data may be represented by an $M \times N$ matrix D. Each row of D represents a molecule. Each column of D represents a restriction site found in the consensus map. The entry found at D_{ij} is the position of restriction site j (corresponding to consensus restriction site j) found on molecule i, in the event that there is not a site j on molecule i then the entry D_{ij} is set to zero. The position may be specified by a metric amounting to a scaled distance in base-pairs from the 3' end of the chromosome. D is a large banded matrix whose band width is equal to the coverage c in expectation, hence the expected sparsity of matrix D is $\frac{c(N-1)}{MN} = \frac{c}{M}$ or $\frac{L}{G}$. We denote the smallest number which bounds the band width of this matrix by c'.

3 EM-Algorithm for Detection of RFLP's

This section details the problem of detecting the *RFLP* event. An expectation maximization or EM-algorithm is given for the estimation of a mixed distribution model, and is followed by a criteria for deciding if data supports an RFLP event.

Consider the jth column of D and take the non-zero entries as a column vector denoted by a and consider it as an instance of random vector A, an $n \times 1$ vector with $\langle A_i \rangle_{i=1:n}$ i.i.d. random variables representing the position of restriction sites taken from a distribution with p.d.f. function:

$$f(A_i = x) \quad = \quad q_1 \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu_1)^2}{2\sigma^2} + q_2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu_2)^2}{2\sigma^2}. \quad (1)$$

We do not know the parameters yet, but without any loss of generality, we may assume that $\mu_1 \leq \mu_2$. Also, $q_i(i = \{1,2\})$ may be interpreted as a probability that point x is derived from the Gaussian with mean μ_i. Further, we have $\sum_i q_i = 1$.

For each random column A of D there is an estimation problem: Namely, determine the values of $\Theta := \langle \mu_1, \mu_2, \sigma, q_1 \rangle$. Once this step is complete, we may detect RFLPs as events involving the distance between μ_1 and μ_2, and also compute probabilities of pairwise events.

The typical approach to such problems is through maximum likelihood estimators (MLE), whereby one maximizes the probability that a particular parameter vector Θ may produce the observed data. $L(\Theta) \quad = \quad P(A = a : \Theta) \quad = \quad \Pi_i f_\Theta(A_i = a_i)$

In the full paper we show that our EM-Algorithm attains the same results as a related Maximum likelihood optimization problem.

We choose to treat one of the parameters q_{1j} as the probability of a hidden random variable for each derivate a_i. Let Y_{1i} be a Bernoulli random variable whose p-value is equal to q_{1i} and represents the probability that the data point a_i is derived from the Gaussian with the leftmost mean.

We note that we have a distribution:

$$P(A_i = x, Y_{1i} = \nu|\Theta) \;=\; \mathbb{I}_{\nu=1} f_1(a_i) + \mathbb{I}_{\nu=2} f_2(a_i)$$

$$f_1(a_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(a_i - \mu_1)^2}{2\sigma^2} \quad \text{and} \quad f_2(a_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(a_i - \mu_2)^2}{2\sigma^2}$$

whose marginals are:

$$P(A_i = x|\Theta) \;=\; \mathbb{E}_\nu \left[\mathbb{I}_{\nu=1} f_1(a_i) + \mathbb{I}_{\nu=2} f_2(a_i)\right] \;=\; q_{i1} f_1(a_i) + q_{i2} f_2(a_i)$$

We now formulate the EM-algorithm by use of Jensen's inequality, omitting details:

$$\begin{aligned} L'(\Theta) &= \log(L(\Theta)) = \log \Pi_{i=1:N} P(A_i = x|\Theta) \\ &\geq F(Q, \Theta) \end{aligned}$$

Where:

$$F(Q, \Theta) \;:=\; \sum_{i=1:N} \sum_{\nu=1:2} Q(Y_i = \nu) \log P(A_i = x, Y_i = \nu|\Theta) + H(Q_i)$$

Here Q is an arbitrary measure and $H(Q) = \sum_i Q_i \log(\frac{1}{Q_i})$ is the Entropy Function on probability vectors:

The EM-algorithm is a process of increasing the F function value [5]. We note that a gradient ascent may be performed on the "likelihood surface" by alternately maximizing Q followed by Θ.

E-Step: $Q_{k+1} \leftarrow \{Q^* : \max_Q F(Q, \Theta_k) = F(Q^*, \Theta_k)\}$.

Lemma 1 *E-Step.*

Let Q be a vector $\langle q_{i\nu} \rangle_{i=1:N, \nu=1:2}$ where N is the number of non-zero entries in the column of data. The Arg-Max can be solved for explicitly with :

$$Q_{k+1} = \langle q_{i1}^{k+1} \rangle_{i=1:N} \quad \text{and} \quad q_{i1}^{(k+1)} \leftarrow \left(\frac{1}{\exp\left(\frac{(a_i - \mu_1^{(k)})^2}{2\sigma^{(k)}} - \frac{(a_i - \mu_2^{(k)})^2}{2\sigma^{(k)}} \right) + 1} \right)$$

The proof is omitted.

M-Step: $\Theta_{k+1} \leftarrow \{\Theta^* : \max_\Theta F(Q_{k+1}, \Theta) = F(Q_{k+1}, \Theta^*)\}$.

Lemma 2 *M-Step.*

The Arg-Max can be solved for explicitly with :

$$\mu_1^{(k+1)} \leftarrow \frac{\sum_{i=1:N} q_{i1}^{(k+1)} a_i}{\sum_{i=1:N} q_{i1}^{(k+1)}}, \qquad \mu_2^{(k+1)} \leftarrow \frac{\sum_{i=1:N} q_{i2}^{(k+1)} a_i}{\sum_{i=1:N} q_{i2}^{(k+1)}}$$

$$\sigma^{(k+1)} \leftarrow \sqrt{\frac{1}{2N} \sum_{\nu=1:2} \sum_{i=1:N} q_{\nu i}^{(k+1)} (a_i - \mu_\nu^{(k)})^2}, \qquad \Theta_{k+1} = \langle \mu_1^{(k+1)}, \mu_2^{(k+1)}, \sigma^{(k+1)} \rangle$$

The proof is omitted.

Lemma 3 *In the limit, the EM algorithm converges to a local maximum of the likelihood function in the parameter space.*

The proof is omitted.

With the lemmas we assume we have procedures called ESTEP and MSTEP. The EM-Algorithm is now:

Algorithm 1

```
EM( A )
    QPREV ← .5*ONES( MAX(SIZE(A)), 2)
    M ← MEAN(A )
    S ← STD(A )
    TPREV ← ( M( 1- S), M( 1 + S), S )
    QNEW ← INF
    TNEW ← INF
    WHILE( MAX( NORM( QPREV - QNEW ) , NORM( TPREV - TNEW ) ) > ε )
        QNEW ← ESTEP( QPREV, TPREV )
        TNEW ← MSTEP( QNEW, TPREV )
    ENDWHILE
    return ( QNEW, TNEW )
```

Denote the return values QNEW with matrix $\left[\hat{q}_1, \hat{q}_2\right]_{M\times 2}$, and TNEW with vector $\langle \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma} \rangle$; the "over-script hats" denoting that these quantities are estimates.

3.1 Detection of RFLPs

We define a detected RFLP as an outcome to our EM algorithm, it is an event such that $|\mu_2 - \mu_1| > \delta$ for some positive $\delta(c)$ as a function of local coverage c.

$$\text{detected RFLP } = (|\mu_2 - \mu_1| > \delta(c)) \tag{2}$$

4 Mathematics of Phasing

The *Phasing problem* consists of inferring two haplotype data sources which combine to provide the observed genotype data.

Each polymorphism gives rise to two alternatives for a diploid organism. We may enumerate the alternatives of polymorphism 1 by the set $\{\uparrow_1, \downarrow_1\}$, and likewise for polymorphism 2 the alternatives are a set $\{\uparrow_2, \downarrow_2\}$. Haplotype I must have one of the alternatives from each set while haplotype II must have the others. The potential co-association for the haplotypes are shown below:

$$\boxed{\begin{array}{c} \text{Haplotype I } \uparrow_1 \uparrow_2 \\ \text{Haplotype II } \downarrow_1 \downarrow_2 \end{array}} \quad (3)$$

$$\boxed{\begin{array}{c} \text{Haplotype I } \uparrow_1 \downarrow_2 \\ \text{Haplotype II } \downarrow_1 \uparrow_2 \end{array}} \quad (4)$$

$$\boxed{\begin{array}{c} \text{Haplotype I } \downarrow_1 \uparrow_2 \\ \text{Haplotype II } \uparrow_1 \downarrow_2 \end{array}} \quad (5)$$

$$\boxed{\begin{array}{c} \text{Haplotype I } \downarrow_1 \downarrow_2 \\ \text{Haplotype II } \uparrow_1 \uparrow_2 \end{array}} \quad (6)$$

These four events (3 – 6) give all possible outcomes to pairwise events.

We may identify events (3) and (6) as the event that alternative $\uparrow_1$, and $\uparrow_2$ are found on the same haplotype, and denote this *covariant* event by the symbol $\uparrow\uparrow$. Similarly we may identify events (4) and (5) as the event that $\uparrow_1$, and $\downarrow_2$ are found on the same haplotype and denote this *contravariant* event by the symbol $\uparrow\downarrow$. One can compute the probabilities of these pairwise events $\uparrow\uparrow$, and $\uparrow\downarrow$ with the use of a continuous multiplicative group, to be introduced below.

4.1 Data maps to Group Elements, MLE homomorphism

The results of EM on each column A_j of D is value $\langle Q(j), \Theta(j)\rangle = \langle \hat{q}_j(\cdot), 1 - \hat{q}_j(\cdot), \mu_{j1}, \mu_{j2}, \sigma_j\rangle$. Consider a data point in the jth column $d_{i'j}$, as such it is derived from a distribution of the form equation(1) whose parameters are given by the results of EM. where $\hat{q}_j$ is an interpolated functional estimate of $q(x)$ a random function giving p-values of the point x being derived from the distribution with μ_{j1}. Let $q_{ji'}$ be the the probability that $d_{i'j}$ derives from the left distribution.

Let $\hat{p}_{ji'} = 1 - \hat{q}_{ji'}$ and identify the data point $d_{i'j}$ with the 2×2 matrix $\begin{bmatrix} \hat{q}_{ji'} & \hat{p}_{ji'} \\ \hat{p}_{ji'} & \hat{q}_{ji'} \end{bmatrix}$. We similarly define a map for each element in the jth column of data:

$$\Phi_j : d_{i'j} \rightarrow \begin{bmatrix} \hat{q}_{ji'} & \hat{p}_{ji'} \\ \hat{p}_{ji'} & \hat{q}_{ji'} \end{bmatrix}$$

The map is an injection into the set $\mathcal{G}$ the 2×2 symmetric bi-stochastic matrices. $\mathcal{G}$ is a set with a natural Abelian group structure under the usual matrix

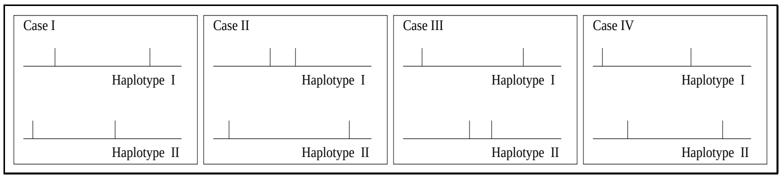


Figure 1. Case I, II, III, IV

multiplication when the degenerate, idempotent (or dead state) element $\begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}$ is excluded from the set.

4.2 Computing Pairwise Events

Let us focus on two restriction sites: Site j and site k that are believed to be RFLPs. These sites have non-constant $\hat{q}_j(x)$,$\hat{q}_k(x)$ functions.

Consider molecules that span both RFLPs, these molecules contain data points x and y which were used to estimate both the functions $\hat{q}_j(x)$ and $\hat{q}_k(x)$. Let us look at the possible haplotypes producing points x and y:

See figure (1) We denoted case 1 by: $(j_{12}, k_{12}) \cap (j_{21}, k_{21})$, case 2 by $(j_{12}, k_{11}) \cap (j_{21}, k_{22})$, case 3 by $(j_{11}, k_{12}) \cap (j_{22}, k_{21})$, and case 4 by $(j_{11}, k_{11}) \cap (j_{22}, k_{22})$. As mentioned before, we care to determine which pairs are found together on a haplotype, the events of interest are:

$$\begin{aligned} E_1 &= ((j_{11}, k_{11}) \cap (j_{22}, k_{22})) \cup ((j_{12}, k_{12}) \cap (j_{21}, k_{21})) \,, \\ E_2 &= ((j_{11}, k_{12}) \cap (j_{22}, k_{21})) \cup ((j_{12}, k_{11}) \cap (j_{21}, k_{22})) \,. \end{aligned}$$

Now we compute the probability that molecule ζ with point x_ζ and point y_ζ support the event E_1:

$$\begin{aligned} P(E_1|\zeta) &= P\left((j_{11}, k_{11}, \zeta \in H_1) \cap (j_{22}, k_{22}, \zeta \in H_2)\right) \cup \left((j_{12}, k_{12}, \zeta \in H_1) \cap (j_{21}, k_{21}, \zeta \in H_2)\right) \\ &= q_j(x_\zeta) q_k(x_\zeta) + p_j(y_\zeta) p_k(y_\zeta) \\ &\approx \hat{q}_{j\zeta} \hat{q}_{k\zeta} + \hat{p}_{j\zeta} \hat{p}_{k\zeta} \,. \end{aligned}$$

Similarly $P(E_2) \approx \hat{q}_{j\zeta} \hat{p}_{k\zeta} + \hat{p}_{j\zeta} \hat{q}_{k\zeta}$.

The connection with the group structure is given by the formula:

$$\begin{bmatrix} P(E_1|\zeta) & P(E_2|\zeta) \\ P(E_2|\zeta) & P(E_1|\zeta) \end{bmatrix} = \begin{bmatrix} \hat{q}_{j\zeta} & \hat{p}_{j\zeta} \\ \hat{p}_{j\zeta} & \hat{q}_{j\zeta} \end{bmatrix} * \begin{bmatrix} \hat{q}_{k\zeta} & \hat{p}_{k\zeta} \\ \hat{p}_{k\zeta} & \hat{q}_{k\zeta} \end{bmatrix}$$

note $P(E_1)$ is the entry on the diagonal while $P(E_2)$ is the entry on the off diagonal of the product: Since sites on different molecules are independent, various probabilities of events (E_1 and E_2) are computed as follows:

$$\begin{bmatrix} P(E_1|\cup_{v=1:m}\zeta_v) & P(E_2|\cup_{v=1:m}\zeta_v) \\ P(E_2|\cup_{v=1:m}\zeta_v) & P(E_1|\cup_{v=1:m}\zeta_v) \end{bmatrix} = \sum_{v=1:m} w_v \begin{bmatrix} \hat{q}_{j\zeta_v} & \hat{p}_{j\zeta_v} \\ \hat{p}_{j\zeta_v} & \hat{q}_{j\zeta_v} \end{bmatrix} *_{\mathcal{G}} \begin{bmatrix} \hat{q}_{k\zeta_v} & \hat{p}_{k\zeta_v} \\ \hat{p}_{k\zeta_v} & \hat{q}_{k\zeta_v} \end{bmatrix}$$

Here $\sum_{v=1:m} w_v = 1$ and is the general form of a prior, for example when all molecules are equally "informative" we have: $w_v = \frac{1}{m}$

Given two restriction sites α and β, we define the *support* of the pair as: $\text{Supp }(\alpha,\beta) = \{\zeta : d_{\zeta\alpha} \neq 0 \wedge d_{\zeta\beta} \neq 0\}$ or equivalently as the number of molecules indexed by ζ that span both sites. The *phase* between two sites: RFLP α and RFLP β may be defined as:

$$\phi_{\alpha,\beta} = \frac{1}{|\text{ Supp }(\alpha,\beta)|} \sum_{\zeta \in \text{ Supp }(\alpha,\beta)} \Phi_\alpha(x_\zeta) *_{\mathcal{G}} \Phi_\beta(y_\zeta)$$

We can also define the distance between two fragments as $d_{\alpha,\beta} = \frac{1}{|\text{ Supp }(\alpha,\beta)|}$.

Computing all pairwise spins can be done with a few sparse matrix multiplications:

Algorithm 2

```
PWS ( P )
    DIST ← ( P != 0 )'*( P != 0 )
    Q ← ONES( SIZE( P )) - P
    θ ← ( (P'*P) + (Q'*Q ))./DIST
     return ( θ )
```

For use in large data sets, we may round values to the singular matrix, for reasons discussed in the section on Chernoff bounds. We define a *dead state* as a spin which is rounded to the singular matrix.

5 Algorithms

We define the phasing problem as follows: *Given a sequence of polymorphisms (whose parameters and distributions have been estimated from a mixture model), use pairwise data to assign the polymorphisms to the haplotypes (i.e., a consistent phasing structure) such that the local assignments are consistent with the data, in the sense of maximum likelihood.*

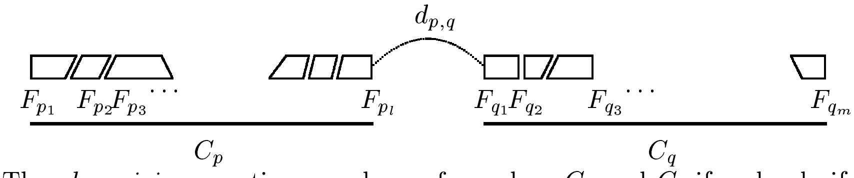
5.1 Weighted k−Neighbor Phase-Contig Algorithm

We can define the phased contigs recursively as follows: The base case is a singleton contig: $C_i = \{F_i\}$ shall be phased as follows:

$$\Phi(C_i) = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \text{if } F_i \text{ is a detected RFLP, (nontrivial contigs) ;} \\ \\ \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix}, & \text{otherwise (trivial contigs).} \end{cases}$$

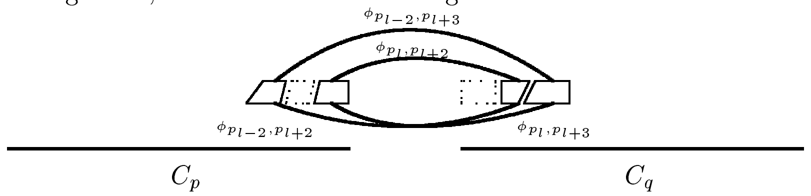
Induction cases: If $C_p = \{F_{p_1}, F_{p_2}, \ldots, F_{p_l}\}$ and $C_q = \{F_{q_1}, F_{q_2}, \ldots, F_{q_m}\}$ are phased contigs with well defined *phasing*, then the union $C_p \cup C_q$ may be phased by a *phase-join* operation.

We define the *distance* between two phased contigs as the minimum distance between two fragments within contigs.

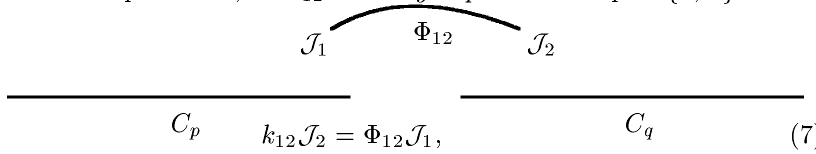


The *phase-join* operation may be performed on C_p and C_q if and only if there is a molecule ζ which contains a data point x_ζ from a restriction site F_{p_l} found in contig C_p, and a data point y_ζ from a restriction site F_{q_1} found in contig C_q, as otherwise the distance is undefined.

For every pair $F_\alpha \in C_p$ and $F_\beta \in C_q$ there are pairwise "phasing" variables to consider in the phase-join. These pairwise phasings tell us how to orient the phased-contig C_q relative to the phased-contig C_p: we will consider a weighted combination of this information, where weights depend on distance between fragments, confidence in RFLP assignment etc.



To attempt a join of C_p to C_q we compute a *mean group action* which is a 'least squares' rotation to be applied similarly to all variables in the right contig to make a "fit" for all pairwise spins in the union of $C_p \cup C_q$. To compute the *group action* for a pair of RFLPs , one in each of the phased-contigs, having spin assignment $\mathcal{J}_1$ and $\mathcal{J}_2$, and pairwise spin Φ_{12}, we derive the chain of computations, let k_{12} be the *group action* for pair $\{1, 2\}$.



$$k_{12}\mathcal{J}_2 = \Phi_{12}\mathcal{J}_1, \tag{7}$$

$$k_{12} = \mathcal{J}_2^{-1}\Phi_{12}\mathcal{J}_1. \tag{8}$$

Solving for k_{12} we find the best rotation for this pair, as after we update the phasing $\mathcal{J}_2 \leftarrow k_{12}\mathcal{J}_2$ the variables would be in a state which satisfy the pairwise spin estimates. Thus in our algorithm the pair $1, 2$ casts a "weighted vote" for $k_{12} = \mathcal{J}_2^{-1}\mathcal{J}_1\Phi_{12}$ as the *mean group element* needed to phase contig C_q correctly. In summary, the contigs are phased by the *mean group action* Φ_{MGA}:

$$\Phi_{MGA} = \sum_{F_\alpha \in C_p, F_\beta \in C_q \beta - \alpha < k+1} w_{\alpha\beta} k_{\alpha\beta}, \text{ where } \sum w_{\alpha\beta} = 1.$$

Now if the resulting *mean group action* $\Phi_{MGA} = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$ is not near the dead state (degenerate matrix), the phase-join operation is successfully executed, and a parent contig $C_p \wedge C_q$ is assigned the value $\Phi(C_p \wedge C_q) \leftarrow \Phi_{MGA}$. We omit the discussion of the special attention that needs to be given to the situation when either of the constituent contigs is trivial.

We omit all the details of an efficient implementation, Assuming that the maximum molecular coverage at any region is c_{max}, the worst case complexity of the phasing algorithm is bounded by $O(c_{max}^2 N\gamma(c_{max}^2 N))$. In practice, γ is a very slow growing function, and the parameters c_{max} and k are likely to be small constants, the algorithm performs almost linearly in the number of polymorphic markers N.

6 Simulations and Examples

We demonstrate our algorithm on two simulated data sets. The views below are broken up into bands, the simulated haplotypes are in the bottom–most band of the layout. Above that is the haplotype molecule map for a diploid organism, these molecule maps are available to the algorithm as mixed data, the segmentation shown is unknown to the algorithm. The third band indicates estimate values and here we can see what features the EM algorithm for mixed Gaussian chooses as RFLPs. Mistakes occur with the lack of a deep library. The forth band in the layout indicates the history of contig–operations and from this tree one can view: 1) the developing k–neighborhoods used to compute mean group action, and 2) the distinct phased contigs. The top band in the layout gives the algorithmic output to this problem, complete with phasing in subsets that span the distance indicated by the bars. Areas where phase structure overlaps but cannot extend indicate regions that are of interest to target with more specific sequences to extend the phasing.

Parameters of the simulations are summarized in the table:

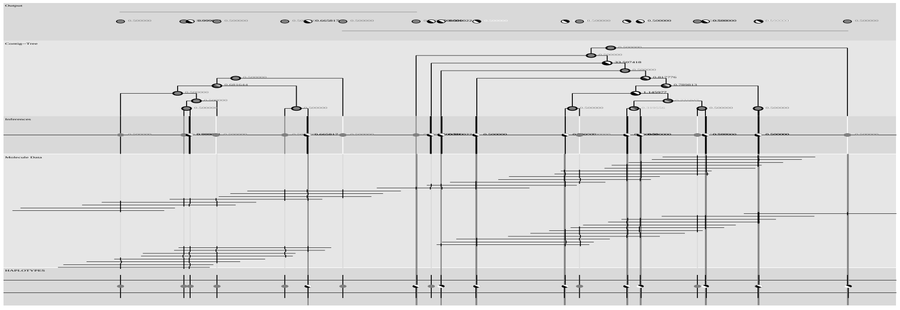


Figure 2. Data set I

Parameter	Symbol	Data Set 1	Data Set 2
number of molecules	M	80	150
number of fragments RFLP and non RFLP	F	20	100
size of the genome	G	12000	50000
expected molecule size	EMS	2000	2000
variance in molecule size	VMS	50	500
variance in fragment length size	VFS	1	20
P–value that any given Fragment is an RFLP	P–BIMODE	.5	.3
Expected separation of means for RFLP	ERFLPSEP	10	50
Variance in the separation of means for RFLP	VRFLPSEP	.01	6

Any parameter with both an expectation and variance is generated with a normal distribution. From these parameters one can compute some additional symbols that we use in the paper $L = \text{ EMS }$ and $c = \frac{LM}{G}$.

For the first simulation on data set I seen in figure 5.1 a relatively small set is chosen so that one can view the action of the algorithm, here the neighborhood size is set to $k = 5$ and there is no ϵ guard of the dead state, still things work pretty well, and one can see that any mistakes are due to the low coverage library.

In the second simulation on data set II seen in figure 5.1 we illustrate that similar results may be achieved on large data sets.

References

1. T.S. Anantharaman, B. Mishra and D.C. Schwartz. "Genomics via Optical Mapping II: Ordered Restriction Maps," **Journal of Computational Biology**, **4(2)**:91–118, 1997.
2. A. Clark. "Inference of Haplotypes from PCR-Amplified Samples of Diploid Populations," **Mol. Biol. Evol**, **7**:111–122,1990.
3. D. Gusfield. "Inference of Haplotypes from Samples of Diploid Populations: Complexity and Algorithms," **Journal of Computational Biology**, **8-3**:305–323,2001.

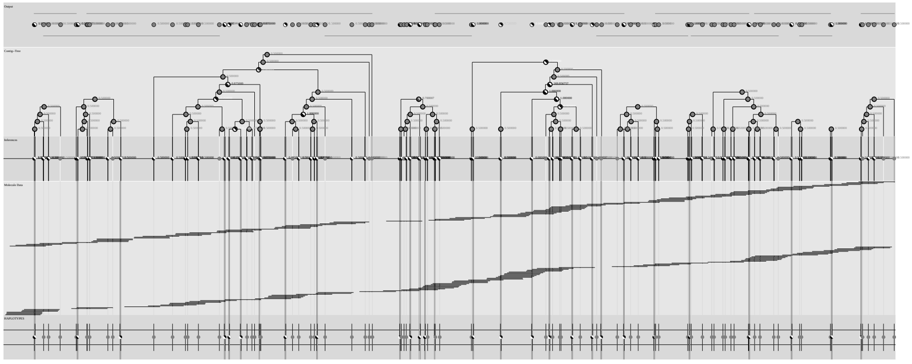


Figure 3. Data set II

4. L. Parida, and B. Mishra. "Partitioning Single-Molecule Maps into Multiple Populations: Algorithms And Probabilistic Analysis," *Discrete Applied Mathematics*, (The Computational Molecular Biology Series), **104**(l-3):203–227, August, 2000.
5. S. Roweis, and Z. Ghahramani. "A Unifying Review of Linear Gaussian Models," **Neural Computation**, **11(2)**:305–345, 1999