An MCMC Algorithm for Haplotype Assembly from Whole-genome Sequence Data: Supplementary Methods

Vikas Bansal, Aaron L. Halpern, Nelson Axelrod, Vineet Bafna

1 Mixing time of Markov chains and Analysis of $\mathcal{M}(\Gamma_1)$

In the general Markov chain Monte Carlo framework for the haplotype assembly problem (described in the Methods section of the paper), for different values of Γ, one can have many different Markov chains $\mathcal{M}(X, q, \Gamma)$ each of which samples the haplotype space. We first show that the Markov chain $\mathcal{M}(X, q, \Gamma)$ has stationary distribution $Pr(H|X, q)$ when Γ includes a minimal collection of subsets $\Gamma_1 = \{\{1\}, \{2\}, \ldots \{n\}\}$.

Definition 1: Let $\mathcal{M}$ be a Markov chain on a finite state space Ω with transition matrix $P = (p_{ij})_{i,j\in\Omega}$. The chain is said to be ergodic if it is irreducible, i.e. $P^n_{ij} > 0$ for some natural number $n > 0$ and all $i, j \in \Omega$ and aperiodic. A Markov chain is said to be *time-reversible* if there exists a probability distribution $\pi = (\pi_i)_{i\in\Omega}$ that satisfies the *detailed balance* condition

$$p_{ij}\pi_i = p_{ji}\pi_j = Q(i, j), \text{ for all } i, j \in \Omega$$

It is well-known that this π satisfying the detailed-balance condition is the unique stationery distribution, as

$$(\pi P)[j] = \sum_i \pi_i P_{ij} = \sum_i \pi_j P_{ji} = \pi_j$$

Therefore, we can show the following:

Theorem 1: Let $\Gamma_1 = \{\{1\}, \{2\}, \ldots \{n\}\}$. For every fragment matrix X, error probabilities $q_i[j] > 0$ $(\forall i, j)$, and any $\Gamma \supseteq \Gamma_1$, $\mathcal{M}(X, q, \Gamma)$ is ergodic and has $Pr(H|X, q)$ as its stationary distribution.

Proof: If $\Gamma_1 \subseteq \Gamma$, then starting from any haplotype configuration H_1 we can reach any other haplotype configuration H_2 by flipping the locations (in an arbitrary order) that they differ in. Hence, for any pair of haplotypes H_1 and H_2, $P^t(H_1, H_2) > 0$ for some $t < n$ where n is the number of columns. Aperiodicity is ensured by the fact that with probability at least $1/2$, the Markov chain

remains in the state that it is currently in. Finally, it can be verified that the detailed balance condition is satisfied for $\pi_H = Pr(X|H,q)$, as for any neighboring H, H',

$$\pi_H Pr(H,H') = \pi_{H'} Pr(H',H) = \min\{Pr(X|H,q), Pr(X|H',q)\}$$

Therefore, the stationary distribution of the chain is $Pr(H|X,q)$.

♣

We give some definitions regarding the mixing time of Markov chains and their relation to *conductance* of the chain.

Definition 2: The distance of the Markov chain from the stationary distribution π at time t is defined as

$$||P^t, \pi|| = \max_{i\in\Omega} \frac{1}{2} \sum_{j\in\Omega} |P_{ij}^t - \pi_j|$$

For any $\epsilon > 0$, the mixing time of the Markov chain is defined as

$$\tau(\epsilon) = \min\{t : ||P^{t'}, \pi|| \leq \epsilon, \forall t' \geq t\}$$

A Markov chain is said to be rapidly mixing if the mixing time of chain can be bounded from above by a polynomial in n and $\log \epsilon^{-1}$ where n is the size of each state of the chain.

Definition 3: For a finite spaced Markov chain $\mathcal{M}$, consider the undirected weighted graph with vertex set Ω and edge set $E = \{(x,y) \in \Omega^2 : Q(x,y) > 0\}$. The conductance[Sinclair and Jerrum, 1989] of the Markov chain $\mathcal{M}$ is defined as

$$\Phi(\mathcal{M}) = \min_{\mathcal{S}\subset\Omega, 0<\pi(\mathcal{S})<1/2} \left(\frac{Q(\mathcal{S}, \overline{\mathcal{S}})}{\pi(\mathcal{S})} \right)$$

where $Q(\mathcal{S}, \overline{\mathcal{S}}) = \sum Q(x,y)$ for all pairs (x,y) such that $x \in \mathcal{S}$ and $y \in \overline{\mathcal{S}}$.

Intuitively, conductance measures the ability of the Markov chain to escape from any subset of the state space. A low conductance implies that the conditional probability of the Markov chain escaping from some subset $\mathcal{S} \in \Omega$ in a single step is small. The following result tightly relates the mixing time of a reversible Markov chain to the conductance.

Theorem 2: [Sinclair et al.[Sinclair and Jerrum, 1989, Sinclair, 1992, Randall, 2006]] Let $\mathcal{M}$ be a finite, reversible, ergodic Markov chain with loop probabilities $P(x,x) \geq 1/2$ for all states x. Then

$$\frac{\log((2\epsilon)^{-1})}{4\Phi} \leq \tau(\epsilon) \leq \frac{2\log((\pi_y \epsilon)^{-1})}{\Phi^2}$$

for any choice of initial state y.

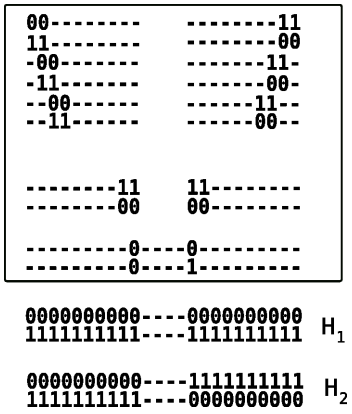

Figure 1: Example of a fragment matrix $(d = 2)$ for which two haplotypes H_1 and H_2 have equal likelihood.

We use these results to show that the mixing time for the Markov chain described in Figure 1 grows exponentially with d, the depth of fragment coverage.

From Figure 1, the two most likely haploytpes are given by

$$H^1 = (00000 \ldots 00000, 11111 \ldots 11111)$$

and

$$H^2 = (00000 \ldots 11111, 11111 \ldots 00000)$$

Theorem 3 Consider a fragment matrix $X(n, d)$ with n columns and $d*(n-2)+2$ rows with structure as described in Figure 1. Let q be the uniform error probability for all positions and define $\rho = \left(\frac{2pq}{p^2+q^2}\right)^d$. For any $\epsilon > 0$, the mixing time of the Markov chain $\mathcal{M}(\Gamma_1, X(n, d), q)$ is $\Omega\left(\log\left(\frac{1}{\epsilon}\right) n\rho^{-1}\right)$.

Proof: Choose $S = \{H^1\}$. There are exactly n edges between S and $\bar{S}$. Of these $n - 4$ have load exactly $\rho^2\pi(H^1)$ and 4 edges $(1, n/2, n/2 + 1, n)$ have weight $\rho\pi(H^1)$. Therefore,

$$Q(S, \bar{S}) = \frac{n-4}{n}\rho^2\pi(H^1) + \frac{4}{n}\rho\pi(H^1) \leq \left(\rho^2 + \frac{4\rho}{n}\right)\pi(H^1)$$

and

$$\pi(S) = \pi(H^1)$$

Therefore $\phi \leq \rho^2 + \frac{4\rho}{n} \leq 5\frac{\rho}{n}$ and the bound on the mixing time follows. ♣

2 Graph Partitioning Algorithm and Weighted Min-Cuts

Here we give a formal description of the recursive graph-partitioning algoritm for constructing Γ from the fragment matrix. We construct a graph $G(X)$ with vertex set V as the set of all columns of X and edge set E as all pairs of columns (i, j) of X for which there is at least one row in X covering both

columns, i.e. $X_k[i] \neq -$ and $X_k[j] \neq -$ for some row k. The weight of the edge (i, j) is the number of such rows covering both columns. A *cut* $(S, \overline{S})$ in $G(X)$ is a partition of the vertices of $G(X)$ into two disjoint set of vertices. The weight of a cut $(S, \overline{S}$ is equal to the sum of the weights of edges going across the cut from S to the $\overline{S}$.

GraphPartitioning(X,Γ)

1. If the number of columns in X is less than 2, return Γ
2. Compute a min-cut $(S, \overline{S})$ in the graph $G(X)$
3. $\Gamma = \Gamma \cup \{S, \overline{S}\}$
4. GraphPartitioning$(X(S), \Gamma)$
5. GraphPartitioning$(X(\overline{S}), \Gamma)$
6. return Γ

Instead of using the number of fragments for assigning weights to edges, we can use information about the actual variant calls and the haplotype pair H for assigning weights to the edges. Consider a pair of columns i, j in X. W.l.o.g, assume that the current haplotype H is described by $H = [00\ldots0, 11\ldots1]$. For a fragment f, we describe a function *match* that scores f on being consistent (or not) with the haplotype at positions i, j. Formally

$$\text{match}_H(f, i, j) = \begin{cases} 1 & \text{if } (f[i], f[j]) \in \{00, 11\} \\ -1 & \text{otherwise} \end{cases}$$

We use the match function for assigning weights to edges in the graph $G(X)$. Define

$$w_H(i, j) = \sum_f \text{match}_H(f, i, j)$$

If $w_H(i, j)$ is highly positive, it implies that the current haplotype pair H is consistent with the phasing suggested by the fragments for the pair (i, j). On the other hand, a negative value for $w_H(i, j)$ indicates that the phasing for the pair (i, j) in the haplotype pair H is likely to be incorrect. We denote the graph with edge weights $w_H(i, j)$ by $G(X, H)$, and would like to compute Min-Cuts. As negative weights on $w_H(i, j)$ make the problem equivalent to the Max-Cut problem, which is known to be computationally hard[Garey and Johnson, 1979], we use the heuristic of removing all edges (i, j) for which $w_H(i, j) < 0$ from $G(X, H)$. We denote the the graph partitioning algorithm based on $G(X, H)$ for computing Γ as *WeightedGraphPartitioning*(X, H).

3 MEC score for haplotype assembly and posterior error probabilities

Given a haplotype assembly H, we would like to evaluate how "consistent" it is with the fragment matrix (sequenced reads) X. Each fragment represents a chunk of DNA from one of the two chromosomes and in the absence of sequencing errors, the alleles at variant sites covered by the fragment should perfectly match one of the two haplotypes. We define $\varepsilon_i[j,h] = 1$ if X_i and h disagree at position j, and let $\text{MEC}(X_i,h) = \sum_j \varepsilon_i[j,h]$ denote the number of alleles mis-matched between X_i and h.

For a haplotype pair $H = (h, \bar{h})$, let $0 \leq Z_i(H)) \leq 1$ denote the probability that the fragment X_i is derived from the haplotype h. Define

$$\text{MEC}(X_i, H) = Z_i(H) \cdot \text{MEC}(X_i, h) + (1 - Z_i(H)) \cdot \text{MEC}(X_i, \bar{h})$$

The total MEC score is defined as the fraction of mismatched variant calls[Bafna et al., 2005, Rizzi et al., 2002]

$$\text{MEC}(X, H) = \frac{1}{n} \sum_i \text{MEC}(X_i, H)$$

The MEC score gives an estimate of the quality of the reconstructed haplotypes, i.e. lower this number, better the haplotypes. For a single haplotype pair H, we can set $Z_i(H) = 1$ if $\text{MEC}(X_i, h) < \text{MEC}(X_i, \bar{h})$ and 0 otherwise. On the other hand, if we are given a probability distribution π on the haplotypes and the matrix q of error probabilities, we can compute

$$Z_i(H) = \frac{Pr(X_i|q,h)}{Pr(X_i|q,h) + Pr(X_i|q,\bar{h})}$$

and the expected MEC score as

$$E_\pi(\text{MEC}(X)) = \sum_H \pi_H \text{MEC}(X, H)$$

Additionally, π can also be used to used to compute posterior error probabilities on the base call $X_i[j]$ as

$$\Pr(\varepsilon_i[j] = 1|\pi) = \sum_H \pi_H \{Z_i(H) \cdot \epsilon_i[j,h] + (1 - Z_i(H)) \cdot \epsilon_i[j,\bar{h}])\}$$

An MCMC algorithm that samples from the posterior distribution of haplotypes can be used to compute the posterior error probabilities where $\pi_H = Pr(X|H,q)$. These error probabilities represent probabilistic estimates of the reliability of each variant call, i.e. an error probability of 0.9 implies a 90% chance of the call being incorrect. To illustrate, consider the example of a fragment matrix with two columns and two fragments: 00 and 01. Let $q_i[j] = \hat{q} = 0.05$ be indentical for all variant calls. The two haplotypes for this matrix are $H_1 = (00, 11)$ and $H_2 = (01, 10)$ with $\pi_{H_1} = \pi_{H_2} = 0.5$. The posterior error probability for each variant call can be easily computed to be ≈ 0.5. Therefore, none of the variant calls is reliable and there is no information about the phase between the two variants present in the data.

References

[Bafna et al., 2005] Bafna, V., Istrail, S., Lancia, G., and Rizzi, R., 2005. Polynomial and APX-hard cases of Individual Haplotyping Problems. *Theoretical Computer Science*, **335**(1):109–125.

[Garey and Johnson, 1979] Garey, M. R. and Johnson, D. S., 1979. *Computers and Intractability: A Guide to the Theory of NP-completeness.* W.H. Freeman and Company.

[Randall, 2006] Randall, D., 2006. Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science and Engineering*, **8**(2):30–41.

[Rizzi et al., 2002] Rizzi, R., Bafna, V., Istrail, S., and Lancia, G., 2002. Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI)*, pages 29–43.

[Sinclair, 1992] Sinclair, A., 1992. Improved Bounds for Mixing Rates of Markov Chains and Multicommodity Flow. *Combinatorics, Probability & Computing*, **1**:351–370.

[Sinclair and Jerrum, 1989] Sinclair, A. and Jerrum, M., 1989. Approximate counting, uniform generation and rapidly mixing markov chains. *Inf. Comput.*, **82**(1):93–133.