Techniques for constructing genome maps

Ashok Rajaraman

August 15, 2011

Abstract

The study of genome maps is pivotal to understanding the sequence of a genome. There are many algorithmic techniques that are used to find this sequence using data obtained from experimental results. This report covers some of these techniques, and their application to radiation hybrid maps.

1 Introduction

The ultimate goal when studying the genome of an organism is to obtain the complete genomic sequence. One of the steps towards obtaining this is genome mapping. This is the process of finding segments of the whole genome called *markers*, and reconstructing the order that these markers occur in on the genome. Markers are generally defined as orthologous segments of the genome when compared with the known sequence of another organism. While it is rare to find an exact one-to-one mapping between these genomic segments in the two organisms, two closely related organisms may have well defined markers. This will be assumed to be the case for the rest of the report. So, each organism we consider will be assumed to have exactly one copy of each marker in its genome.

The aim of genome mapping, as stated, is to reconstruct the order that these markers appear in. To do this, experiments are performed to obtain raw data that might provide evidence as to what the true order is.

The data obtained from these experiments is used as the input for many powerful algorithmic tools which try to reconstruct the true order. This report studies some of these tools and the principles behind them. Throughout the report, the genome we are reconstructing shall be assumed to be unichromosomal and linear, unless otherwise mentioned.

Section 2 discusses how the marker order on a reference genome can be used to reconstruct the marker order in the new genome. Section 3 focuses on the analysis of data obtained from a wet-experiment known as radiation hybridization. Section 4 expands on the methods discussed in section 2 to construct a set of 'good' marker orders. We also discuss some of the results that have been published using these methods in section 5.

2 Comparative genome approach to marker ordering

The following framework was provided by Faraut et al. [5]. Their model uses a reference genome find the order of orthologous genes (markers) in a related organism.

Let us identify the markers by the labels $\{1, 2, \ldots, n\}$. An ordering of these markers is a permutation π of these labels. In general, the ordering of the markers in a sequenced genome of an organism is taken as reference, and is assigned the identity permutation π_{ref}. This identifies each marker with a unique label.

To find the ordering of these markers in another genome, experiments are first used to obtain evidence for the true ordering. We shall call this evidence X. So, using Bayesian inference, our problem is to find the permutation, i.e. an ordering of the markers, $\bar{\pi}$ which maximizes the following probability.

$$Pr(\pi|X) = \frac{Pr(X|\pi) \cdot Pr(\pi)}{\sum_{\sigma \in \mathcal{S}_n} Pr(X|\sigma) \cdot Pr(\sigma)}. \tag{1}$$

Since the denominator is a constant for every permutation, we may instead maximize the right hand side of the following relation

$$Pr(\pi|X) \propto Pr(X|\pi) \cdot Pr(\pi). \tag{2}$$

The term $Pr(X|\pi)$ depends upon the outcome of the experiment, as well as parameters intrinsic to it. Estimating this quantity based on radiation-hybrid experiments will be part of the next section of this report.

The other term, $Pr(\pi)$ can be naively take to be $\frac{1}{n!}$, which says that the prior on each permutation is equal, or equivalently that each order is equally likely. Thus, the optimal order only depends on the posterior.

Faraut et al. modelled the probability of an order as a function of its evolutionary distance from the reference genome. The metric used was the number of adjacent markers in π which were not adjacent in π_{ref}. Thus, the metric is similar to the breakpoint distance. However, the direction of the markers is not important, and so the permutations are unsigned. In the report, we shall use the term *breakpoint distance* to denote this metric instead of the more classical definition of the same. The occurrence of a breakpoint between two markers was modelled using a Poisson process. The parameter used for the process is the expected breakpoint distance between the reference genome and the proposed gene order.

Assume that some order π has breakpoint distance k from the reference genome. In that case, the probability of the order will be given as follows:

$$\begin{aligned} Pr(\pi) &= Pr(\pi|\pi_{ref}) \\ &= Pr(\pi|k)\, Pr(k). \end{aligned} \tag{3}$$

Then, equation 2 becomes

$$Pr(\pi|X) \propto Pr(X|\pi) \cdot Pr(\pi|k)\, Pr(k). \tag{4}$$

Since the occurrence of breakpoints is modelled as a Poisson process, the term $Pr(k)$, which is the probability of observing k breakpoints, is given by $\frac{\lambda^k e^{-\lambda}}{k!}$, where λ is the

parameter controlling the Poisson process. The other term is simply the probability of observing π given that there are exactly k breakpoints. So,

$$Pr\left(\pi|k\right) = \frac{1}{O_n\left(k\right)}, \tag{5}$$

where $O_n\left(k\right)$ is the total number of permutations which are at breakpoint distance k from the reference genome.

To maximize the right hand side of 4, it is clear that an order with a low number of breakpoints with the reference order will be preferred, and this balances the posterior probability due to X. This brings us to the combinatorial question of finding the number of permutations at breakpoint distance k from the reference genome.

2.1 Number of permutations with k breakpoints

The problem of finding the number of permutations at a breakpoint distance of k from the reference genome is solved by setting up a system of recurrences. These recurrences capture the possible scenarios when a genome of size $n-1$ is expanded to a genome of size n.

A *segment* of a permutation is a maximal set of markers in the permutation such that all the markers in the segment are also adjacent in the reference genome. A segment with a single marker is called an *isolated* marker. Consider a permutation of length $n-1$, with $k-1$ breakpoints with the identity. Now we proceed by induction.

(i) If the marker $n-1$ is at the end of the segment, and we add the marker n at the end, we will not create any breakpoints. So, the new permutation on n markers will have $k-1$ break points.

(ii) If the marker n is inserted next to the marker $n-1$, but inside a segment instead of at an end, then the new permutation will have k breakpoints.

(iii) If we insert the marker n at the position of an existing breakpoint, we create one extra break, and the new permutation has k breaks.

(iv) If we insert the marker n at the end of the permutation, such that the end is not the marker $n-1$, the new permutation will have k breaks.

(v) If the marker n is inserted between markers $i, i+1$ in a segment, such that $i+1 \neq n-1$, then the new permutation will have $k+1$ breakpoints.

The position of the marker $n-1$, at the end of a segment or as an isolated marker, is the only information required to construct the next set of permutations. Now, we define the following quantities.

- $I_n^b\left(k\right)$ =The number of permutations of size n with k breakpoints, such that the marker n is isolated at the border of the permutation.
- $I_n^c\left(k\right)$ =The number of permutations of size n with k breakpoints, such that the marker n is isolated somewhere in the middle of the permutation (not at the border).

- $S_n^b(k)$ =The number of permutations of size n with k breakpoints, such that the marker n is part of a segment, and is at the border of the permutation.
- $S_n^c(k)$ =The number of permutations of size n with k breakpoints, such that the marker n is part of a segment, somewhere in the middle of the permutation.
- $O_n^b(k) = I_n^b(k) + S_n^b(k)$ =The number of permutations of size n with k breakpoints, such that the marker n is at the border of the permutation.
- $O_n^c(k) = I_n^c(k) + S_n^c(k)$ =The number of permutations of size n with k breakpoints, such that the marker n is in the middle of the permutation.

Clearly, $O_n^c(k) + O_n^b(k) = O_n(k)$. Using the initial values $I_2^b(0) = I_2^c(0) = S_2^c(0) = 0, S_2^b(0) = 1$, we can define the following recurrence relations by using inclusion-exclusion.

$$I_n^b(k) = O_{n-1}^b(k-1) + 2O_{n-1}^c(k-1) \tag{6a}$$
$$I_n^c(k) = (k-1)O_{n-1}(k-1) + (n-k)O_{n-1}(k-2) - S_n^c(k-1) \tag{6b}$$
$$S_n^b(k) = O_{n-1}^b(k) \tag{6c}$$
$$S_n^c(k) = I_{n-1}^b(k) + 2I_{n-1}^c(k) + S_{n-1}^c(k) + S_{n-1}^c(k-1) + S_{n-1}^b(k-1) \tag{6d}$$

Using these recurrences, it is easy to compute all $O_n(k)$ for some $n \leq N$ and $k < N$ is time quadratic in N.

2.2 Extension to multichromosomal genomes

While the recurrence relations presented above hold for linear, unichromosomal genomes, it is easy to modify the same for multichromosomal genomes. To do so, we first glue together the chromosomes of the reference genome in some prespecified order. This gives us a unichromosomal genome, which can be labelled as before, but the gluing points are treated as breakpoints instead of adjacencies. Now, assuming that the reference genome had r chromosomes, and the order under consideration has n_i markers from chromosome i for $1 \leq i \leq r$, we notice that while adding markers, each time we finish adding markers from chromosome i and start adding markers from chromosome $i+1$, we will have to introduce a breakpoint. So, when we add the $\left(1 + \sum_{i=1}^{j} n_i\right)^{th}$ marker to the new order (note that this is not the identification number of the marker, but the step in which it is added to the new genome), where $1 \leq j \leq r$, then we have to make the following modifications to our recurrence relations.

$$I_n^b(k) = 2O_{n-1}(k-1) \tag{7a}$$
$$I_n^c(k) = (k-1)O_{n-1}(k-1) + (n-k)O_{n-1}(k-2) \tag{7b}$$
$$S_n^b(k) = S_n^c(k) = 0 \tag{7c}$$

The final case says that if we are adding a marker from a new chromosome, it is not possible to create an adjacency.

2.3 Final steps to setting up the problem

Having calculated the probability of choosing the order π given the reference, we can look at the following log-likelihood maximization problem.

$$\ln Pr\left(\pi|X\right) = \ln Pr\left(X|\pi\right) + \ln\left[Pr\left(\pi|k\right) Pr_{\lambda}\left(k\right)\right] + C, \tag{8}$$

where C arises from the normalization constant. This is clearly equivalent to the problem as stated in equation 4. The next step is to compute $Pr\left(\pi|k\right)$ for $0 \leq k \leq n-1$, and to find a least squares fit $a + bk$ to $\ln\left[Pr\left(\pi|k\right) Pr_{\lambda}\left(k\right)\right]$. After this reduction is made, the problem is generally reduced to an instance of the Travelling Salesman Problem, as we shall see in the case of Radiation Hybrid maps in section 3.

2.4 Using more than one reference

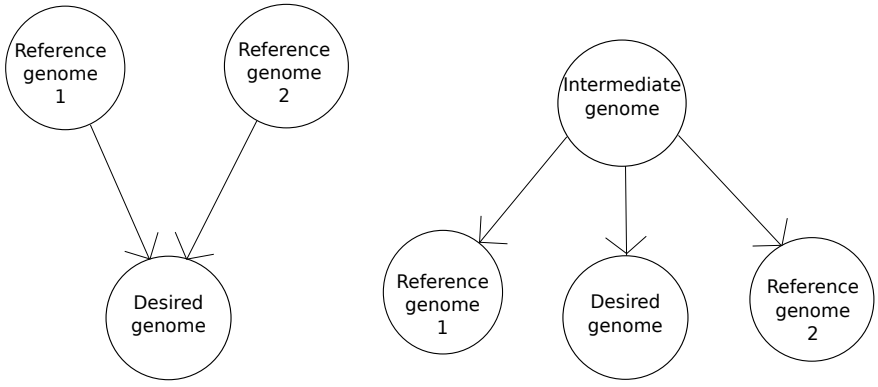
If there is more than one genome which supports the case for an ordering, then we simply need to treat the two reference orders are independent of each other, as shown in figure 1a. Since the terms $\ln\left[Pr\left(\pi|k\right) Pr_{\lambda}\left(k\right)\right]$ in equation 8 are independent of the position of the markers, and dependent solely on the number of breakpoints with the reference genome, we can easily generalize this. Assuming that the two reference orders given to us are π_{ref1} and π_{ref2} (we can take one to be the identity without loss of generality),

$$Pr\left(\pi|\pi_{ref1}, \pi_{ref2}\right) = Pr\left(\pi|k_1\right) Pr\left(\pi|k_2\right) Pr_{\lambda_1}\left(k_1\right) Pr_{\lambda_2}\left(k_2\right),$$

where k_1 and k_2 are the number of breakpoints that π has with π_{ref1} and π_{ref2} respectively, and λ_1 and λ_2 are the controlling parameters of the respective Poisson processes of introducing breakpoints. Using the recurrence relations, these are easily computable, and the least squares fit for both can be calculated.

$$\ln Pr\left(\pi|X\right) = \ln Pr\left(X|\pi\right) + a_1 + b_1 k_1 + a_2 + b_2 k_2 + C'. \tag{9}$$

The process can also be extended when using more than two references.



(a) Independent reference genomes

(b) Through an intermediate genome

Figure 1: Using more than one reference

If, on the other hand, we wish to be more rigorous, and not assume the independence of the two reference genomes, we might wish to consider constructing an intermediate ordering of the markers which is the most probable ordering based on the two reference genomes, depicted in figure 1b. Using Bayesian inference, the two reference orders will be conditionally independent of each other, given the parent. Thus, we want to find an order $\bar{\pi}_{inter}$ which maximizes the following probability.

$$Pr\left(\pi_{inter}|\pi_{ref1},\pi_{ref2}\right) = \frac{Pr\left(\pi_{ref1}|\pi_{inter}\right)Pr\left(\pi_{ref2}|\pi_{inter}\right)}{\left(\sum_{\sigma\in\mathcal{S}_n} Pr\left(\pi_{ref1}|\pi_{\sigma}\right)\right)\left(\sum_{\sigma\in\mathcal{S}_n} Pr\left(\pi_{ref2}|\pi_{\sigma}\right)\right)}$$

Note that there is no posterior evidence for $\bar{\pi}_{inter}$ apart from the reference orders. We can then use this intermediate genome as the new reference to find the optimal order $\bar{\pi}$.

3 Radiation Hybrid maps

The construction of a reliable order of markers on a chromosome, or a *mapping of the markers* is an important step towards sequencing chromosomal DNA. Radiation hybrid (RH) mapping is a technique that is used to construct such maps, and estimate the distance between the markers on the chromosome.

The experimental stage of RH mapping consists of irradiating the cells of the organism on whose chromosome we need to order the markers, and fusing these cells with healthy cells of another organism. The irradiation breaks the healthy chromosome of the original organism at random intervals, into many *fragments*, which are subintervals of the original chromosome, and contains the markers in this subinterval. Fusing the cells results in the *rescue* of a subset of these fragments, by recombination with healthy cells. This creates a *hybrid clone*, whose chromosome can be tested for the presence or absence of each marker.

The experiment is repeated several times, and provides us with data to analyze, and with parameters that can be incorporated into the analytical model that we adopt. The algorithmic part of the RH process aims to deduce the most likely order of the markers on the original chromosome, given the retention pattern of the markers, i.e. the absence or presence of the markers on each hybrid.

The question of 'most likely order' is generally solved by reducing the problem to a maximum likelihood estimation (MLE) setting, or by minimizing the number of *obligate chromosomal breaks* (OCBs), as we shall explain later. Both these instances can be further reduced to the travelling salesman problem [1, 3]. Methods to approach the problem also include trying to construct a minimal weighted Hamiltonian path [2], but we shall restrict ourselves to a high level discussion of the reduction of the problem to a TSP, rather than focusing on the heuristics used to solve the TSP.

A variant of the maximum likelihood approach aims to order the markers with respect to some reference order. Furthermore, this approach can be extended to provide a *map distribution*, i.e. other possible maps that can have led to the radiation hybrid data, assuming uncertainty in the reconstructed order.

The following section introduces the object constructed from radiation hybrid data, on which all these techniques can be applied. This shall be followed by brief dis-

cussions of the minimization of obligate chromosomal breaks, and of the maximum likelihood methods.

3.1 The Object

Assume that there were m independent hybridization experiments performed on the same chromosome. This produces m hybrids.

If we know that the original chromosome had n markers, we can check each hybrid for the presence or absence of these markers. Assuming an arbitrary order of the markers, the presence or absence of the markers in these hybrids can be given in a matrix $X = (x_{ij})_{m \times n}$, where x_{ij} is 0 if marker j is not present in hybrid i, and 1 otherwise.

For example, consider 3 hybrids with 4 markers in the original chromosome. A possible matrix would be:

$$\begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

The matrix construction described above is valid for haploid, error free data. This matrix can be modified for the following cases:

1. For diploid data, we get a matrix with entries $0, 1, 2$, where x_{ij} is 2 if both copies of the marker are rescued and found on the hybrid.

2. There may be typing (reading) errors, in which case we may be uncertain whether a certain marker is present or absent. Using the notation of Ben-Dor et al., these entries will be depicted as $?$.

It is easy to see that there are $n!/2$ possible orderings of the columns. Typically, n is large, of the order of hundreds, or even thousands.

The assumption that markers that are close together will lie on the same fragment, i.e. there will be no break between them, means that we desire an ordering of the columns of the matrix that tries to group all the 1's together. This leads to the first optimization criterion.

3.2 Obligate Chromosomal breaks(OCB)

An obligate chromosomal break in a hybrid for a certain marker order is scored whenever a 1 is immediately followed by a 0, or a 0 is immediately followed by a 1 while traversing the corresponding row of the matrix. The OCB score for the entire matrix for a given marker order is the sum over the OCB scores of each row/hybrid. In the matrix given before, there are 2 OCBs in the first row, and 7 OCBs in total. Non-informative entries in the matrix, i.e. those marked by '?', do not count towards breaks, and changes from 0 to $?$ or 1 to 0 etc., is ignored while counting the number of OCBs.

Finding a column order with the minimum number of obligate chromosomal breaks translates to finding a marker order in which the least number of radiation induced breaks have occurred. If there is an order with no obligate chromosomal breaks, then

the matrix, by definition, has the consecutive ones property. Otherwise, the problem of optimizing the number of such breaks is performed by reducing the problem to an instance of the TSP.

3.2.1 Reduction to TSP

Assume that we are given a matrix $X = (x_{ij})_{m \times n}$ of m hybrids and n markers, with possible typing errors, but no diploid data. The matrix is used to construct a complete graph, whose vertex set is the set of markers and a separate source vertex s. Then, the following rules are used to weigh the edges.

1. The edges from s to each vertex is given a weight of 0.
2. The edge from a vertex u to v is given the following weight:

$$w_{uv} = \frac{\text{\# of entries } i \text{ in which } x_{iu} \text{ and } x_{iv} \text{ differ by } 1}{\text{\# of entries } i \text{ in which both } x_{iu} \text{ and } x_{iv} \text{ are not '?'}}$$

   The edge has a higher weight if there are more hybrids in which only one of marker u and v is retained.

Assuming we have full information, i.e. no non-informative entries in the matrix, the minimum number of obligate chromosomal breaks will be m times the optimal Travelling salesman tour in this graph. Thus, the TSP solution to this graph will minimize the number of OCBs, and yield an optimal ordering of the markers.

3.2.2 Ambiguous entries

In the case when there are ambiguous entries, the TSP solution need not be optimal. consider the following matrix:

$$\begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1 & ? & 1 \\ 1 & ? & 1 \\ 1 & ? & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}.$$

In this case, the optimal TSP tour is given by the permutation $(1, 3, 2)$ of the markers, but induces 3 breaks. The identity order, though, induces only 2 breaks.

This last case is generally claimed to be close to the actual solution, since most data has low fraction of ambiguous entries.

3.3 Maximum Likelihood Estimation (MLE)

The second optimization criterion is to estimate an order and the distance between the markers such that the probability of the resulting RH data, given that order, is maximized. This is called the likelihood of that map/order. We introduce some notation here that will be used in this section, as well as the others to follow it.

The probability of rescue or retention p is the probability that a fragment created by irradiating the cells is recombined to form a hybrid. The variable q will be used to denote $1-p$, i.e., the probability that a fragment is not rescued. p is estimated by the ratio of the number of 1's in the hybrid data to the number of 1's and 0's. If the data is fully informative, i.e. no ambiguous entries, then this is the ratio of 1's to the size of the matrix. The breakage probability θ_{ij} between two markers i and j is the probability that a radiation induced break occurs between marker i and marker j.

Apart from these, we can also have error rates for false positive and false negative observations, as well as factors that come into play when we consider the ploidity of the chromosome. For our purposes we shall examine only the methods used to analyze haploid, error free data.

3.3.1 Estimating breakage probabilities

The first step in maximum likelihood estimation is to find the breakage probability θ_{ij} between each pair of markers. For a single hybrid h_k, we can denote the possible cases for the presence or absence of the markers i and j, and the probabilities associated with each, as follows.

i / j	0	1
0	$q(1-\theta_{ij}p)$	$\theta_{ij}qp$
1	$\theta_{ij}qp$	$p(1-\theta_{ij}q)$

For example, both markers i and j will be absent on the hybrid if one of two cases is satisfied:

1. Marker i is lost (probability q), and marker j was on the same fragment (no break probability $1-\theta_{ij}$).
2. A break occurred (probability θ_{ij}) and both i and j are lost.

This gives us the first entry in the table.

Let n_{00} denote the number of hybrids in which both marker i and j are not present, n_{10} the number of hybrids in which only marker i is present, n_{01} the number of hybrids in which only marker j is present, and n_{11} the number of hybrids in which both are present. Then the probability of observing the columns (x_{ki}) and (x_{kj}), where $1 \leq k \leq m$, together, given θ_{ij} is equal to

$$Pr\left((x_{ki}),(x_{kj})\,|\theta_{ij}\right) = \left(q\left(1-\theta_{ij}p\right)\right)^{n_{00}}\left(\theta_{ij}qp\right)^{n_{01}+n_{10}}\left(p\left(1-\theta_{ij}q\right)\right)^{n_{11}}.$$

This probability is maximized when the derivative of this term with respect to θ_{ij} is zero. A quadratic polynomial is obtained, which can be solver for θ_{ij}, and the solution chosen is one which lies in $[0,1]$ and maximizes the probability.

Using this method, the breakage probability for every pair of markers i,j is estimated. The probability thus maximized is the likelihood of seeing the columns (x_{ki})

and (x_{kj}) together. The likelihood of observing a single column (x_{ki}), $Pr(\pi(1))$, is given by $q^{n_0}p^{n_1}$, where n_1 is the number of 1's in the column, and n_0 is the number of 0's. If we assume that the probability of observing a column is dependent only on its immediate predecessor (i.e. the columns before it do not affect it), then we can define the two-point likelihood of an order π of the markers as follows:

$$Pr(X|\pi) = Pr((\pi(1)))\, Pr((\pi(2)) \,|\, (\pi(1))) \ldots Pr((\pi(n)) \,|\, (\pi(n-1))), \quad (10)$$

where each $(\pi(i))$ is the column of the marker in position i as observed in X. Since we need to find the marker order which gives us the greatest probability of observing the data X, we need to find an order which maximizes this quantity.

3.3.2 Reduction to TSP

The case of reducing the MLE problem to TSP is slightly more involved than the case of OCBs. The graph is the same complete graph on vertices labelled by the markers and a source vertex, which is a dummy marker. Before the weights are calculated, we define transition probabilities between markers.

- The transition probability t_i from s to the marker i, is given by

  $$t_i = p^{n_1/2} q^{n_0/2},$$

  where n_1 is the number of hybrids in which the marker i is present, and n_0 is the number of hybrids in which it is absent.

- The transition probability t_{ij} from a marker i to j or vice-versa, is given by

  $$t_{ij} = (q(1-\theta_{ij}p))^{n_{00}} (\theta_{ij}qp)^{(n_{01}+n_{10})/2} (p(1-\theta_{ij}q))^{n_{11}}.$$

  where the definitions of $n_{00}, n_{01}, n_{10}, n_{11}$ are the same as given before.

This definition of transition probabilities can be exploited using Karp et al.'s result that, for a given permutation π of the markers, the product of the transition probabilities is equal to the likelihood of that order of markers [6]. This was left as an exercise in the original paper, and was proved in the paper by Aggarwal et al. for completeness [1].

To maximize the likelihood, we can minimize the negative log-likelihood. Thus, we weight each edge ij by $-\ln t_{ij}$ and si by $-\ln t_i$. Now, the solution to the TSP will minimize the weight of the total tour, which maximizes the likelihood, and the order of traversal again gives us the most probable order.

3.3.3 Utilising the comparative genome approach

In section 2, we saw a method for introducing a prior on the ordering of the markers using a reference order. Faraut et al. used their own approach to attack RH-data sets. The term $\ln Pr(X|\pi)$ in equation 8 is the taken to be the log-likelihood that we see the RH-data X for a given marker ordering π. Using Karp et al.'s reduction, and the linear regression fit discussed in section 2.3, the equation reduces to

$$\ln Pr(\pi|X) = \ln t_{\pi(1)} t_{\pi(2)\pi(1)} \ldots t_{\pi(n-1)\pi(n)} t_{\pi(n)} + a + bk + C. \quad (11)$$

Now, we change the weights to each edge ij as follows:

(i) Each edge si shall be weighed $-\ln t_i - a/2$.

(ii) An edge ij shall be weighed $-\ln t_{ij} - b \times \gamma_{ij}$, where γ_{ij} is 0 if i and j are adjacent in the reference order, and 1 otherwise.

This modification in the weights of the graphs weighs the edges such that a TSP solution will maximize the probability of the order conditional on both the RH-data as well as the reference order.

4 Statistical confidence measures

In section 2, we discussed the model used by Faraut et al. to find a prior distribution for marker orderings. Servin et al. [7] used this information to evaluate the uncertainty in a constructed genome map. The essence of their work is that instead of calculating a specified order of the markers, they compute the probabilities for certain 'good' orders which agree with some reference genome, and with the evidence. Based on these orders, they construct what they call a *robust map*, which captures the possible ordering of the markers.

4.1 Determining the confidence measure

Servin et al. proposed a Markov Chain Monte Carlo approach to estimate the confidence measure in a constructed map. Recall equations 3 and 4, which together give:

$$Pr\left(\pi|X\right) \propto Pr\left(X|\pi\right) \cdot Pr\left(\pi|\pi_{ref}\right).$$

To compute the likelihood of an order π means navigating through all $n!/2$ possible orders of n markers. Instead, we would like to look at permutations that are somehow close to the permutation π. The MCMC algorithm does this in two steps. The first step is a Metropolis Hastings sampling step, which is performed once for every iteration. The second step is a Gibbs sampling step which is performed $2n$ times, where n is the number of markers.

Metropolis Hastings sampling of inversions: The first step in the MCMC looks at the map $I\left(\pi, i, j\right)$, $1 \leq i \leq n-1, i+1 \leq j \leq n$, which is the permutation $\pi^{'}$ obtained by inversion of the segment between markers i and j (i and j included) in π. For each $\pi^{'} = I\left(\pi, i, j\right)$, the quantity $Q\left(\pi, i, j\right) = Pr\left(X|\pi^{'}\right) Pr\left(\pi^{'}|\pi_{ref}\right)$ is computed. Using this, we get the following probability distribution:

$$Pr\left(\pi'|\pi\right) = \frac{Q\left(\pi, i', j'\right)}{\sum_{(i,j)} Q\left(\pi, i, j\right)}. \tag{12}$$

We sample an inversion from this probability distribution, and accept it with the following probability:

$$\min\left(1, \frac{Pr\left(\pi'|X\right) Pr\left(\pi|\pi'\right)}{Pr\left(\pi|X\right) Pr\left(\pi'|\pi\right)}\right).$$

So, there is a greater chance of accepting the inversion if it increases the likelihood of the order π'.

Gibbs sampling of marker replacement: Once the Metropolis Hastings step is run, the Gibbs sampling step is run $2n$ times when there are n markers. $T(\pi, i, i')$ is defined as the map in which the marker at position i in π exchanges places with the marker at position i'. This defines a probability distribution:

$$Pr(i'|\pi_{-i}, X) = \frac{Pr(X|T(\pi,i,i'))\, Pr(T(\pi,i,i')|\pi_{ref})}{\sum_k Pr(X|T(\pi,i,k))\, Pr(T(\pi,i,k)|\pi_{ref})}, \tag{13}$$

where π_{-i} is the order π with the marker at position i removed. An iteration of this step is completed by choosing an order from this distribution.

These steps are run after the comparative approach is used construct an initial order $\bar{\pi}$, and, after some 'burning' iterations, converge to some orders in the neighbourhood of $\bar{\pi}$.

Once the algorithm has run, we obtain a map distribution, and a posterior probability associated to each order in this distribution. These are the confidence measures for the maps obtained, i.e. the probability that the map agrees with the data and the reference.

4.2 Constructing robust maps

The maps constructed by the algorithm may number in the thousands, but since they are all somehow 'close', these maps share various structures. We are interested in three kinds of structures:

- *Sequences* of markers, i.e. markers that are organized consecutively in all maps.
- Sequences of sequences, or *metasequences*, which are consecutive sequences in all maps.
- *Common intervals*, sets of markers that occur together, but not necessarily consecutively in all maps.

Finding the sequences and meta-sequences in the orders is relatively easy. The algorithm of Bergeron et al. [4] can be used to find all the common intervals between the orders. These structures can then be arranged in the form of a PQ-tree (called a *metamap*), which is a *robust map*, in the sense that the probability of finding these structures in the true marker order is very high. The P nodes hold the probabilities of concurrent order of the markers associated with them, and the Q nodes hold the probabilities of each orientation of the associated markers.

5 Discussion of the results

The method of Faraut et al. [5] was used to find an ordering of the canine chromosome 2 with 426 markers typed to obtain an RH dataset, using the human genome as a reference map. Each marker occurs on average after 200kb. Using a Poisson constant of

1 for modelling the occurrence of breakpoints, the approach used was in closer agreement with the dog genome sequence than the order obtained by naive TSP calculation. The effect of the constant was judged to be minimal after experiments on simulated data.

The same dataset was used by Servin et al. [7], only they used 423 markers. The map distribution constructed using the MCMC algorithm consisted of over 10000 maps. The main observation was that areas of conflict in the reconstructed map by Faraut et al. and the true genome order were regions where other orderings were seen in the map distribution, which makes the case for a confidence measure. In their own words, "our method allows us to pinpoint regions where the assembly order disagrees with an RH map order that is strongly supported by RH data" [7].

6 Conclusion

The report discusses some advances in the field of marker ordering in genome maps, in particular RH maps. The methods of Faraut and Servin are general frameworks that can be applied to the reconstruction of maker order from experimental data. Their methods and results lean heavily on the framework provided by Karp, Aggarwal and Ben-Dor for reconstructing marker orderings from RH data, which utilizes an elegant reduction to the Travelling Salesman Problem. The question of using more than one reference order to reconstruct the orders remains an open problem, but the framework seems to be flexible enough to allow us to incorporate this.

References

[1] R. Agarwala, D.L. Applegate, D. Maglott, G.D. Schuler, and A.A. Schäffer, *A fast and scalable radiation hybrid map construction and integration strategy*, Genome Research **10** (2000), no. 3, 350.

[2] A. Ben-Dor and B. Chor, *On constructing radiation hybrid maps*, Journal of Computational Biology **4** (1997), no. 4, 517–533.

[3] A. Ben-Dor, B. Chor, and D. Pelleg, *RHO-radiation hybrid ordering*, Genome Research **10** (2000), no. 3, 365.

[4] A. Bergeron, C. Chauve, F. De Montgolfier, and M. Raffinot, *Computing common intervals of k permutations, with applications to modular decomposition of graphs*, Algorithms–ESA 2005 (2005), 779–790.

[5] T. Faraut, S. De Givry, P. Chabrier, T. Derrien, F. Galibert, C. Hitte, and T. Schiex, *A comparative genome approach to marker ordering*, Bioinformatics **23** (2007), no. 2, e50.

[6] R. Karp, W. Ruzzo, and M. Tompa, *Algorithms in molecular biology-lecture notes. 1996*, Department of Computer Science and Engineering, University of Washington, Seattle, WA.

[7] B. Servin, S. de Givry, and T. Faraut, *Statistical confidence measures for genome maps: application to the validation of genome assemblies*, Bioinformatics **26** (2010), no. 24, 3035.