BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

A comparative genome approach to marker ordering

T. Faraut [a, *, †], S. de Givry [b, †], P. Chabrier [b], T. Derrien [c], F. Galibert [c], C. Hitte [c] and T. Schiex [b]

[a]Laboratoire de génétique cellulaire and [b]Laboratoire de mathématiques et informatique appliquées, INRA, BP 52627, 31326 Castanet Tolosan, France
[c]UMR 6061 Génétique et Développement, CNRS-Université de Rennes 1, Faculté de Médecine, 2 Av du Pr Leon Bernard, CS 34317, 35043 Rennes, France

ABSTRACT

Motivation: Genome maps are fundamental to the study of an organism and essential in the process of genome sequencing which in turn provides the ultimate map of the genome. The increased number of genomes being sequenced offers new opportunities for the mapping of closely related organisms. We propose here an algorithmic formalization of a genome comparison approach to marker ordering.

Results: In order to integrate a comparative mapping approach in the algorithmic process of map construction and selection, we propose to extend the usual statistical model describing the experimental data, here radiation hybrids (RH) data, in a statistical framework that models additionally the evolutionary relationships between a proposed map and a reference map: an existing map of the corresponding orthologous genes or markers in a closely related organism. This has concretely the effect of exploiting, in the process of map selection, the information of marker adjacencies in the related genome when the information provided by the experimental data is not conclusive for the purpose of ordering. In order to compute efficiently the map, we proceed to a reduction of the maximum likelihood estimation to the Traveling Salesman Problem. Experiments on simulated RH data sets as well as on a real RH data set from the canine RH project show that maps produced using the likelihood defined by the new model are significantly better than maps built using the traditional RH model.

Availability: The comparative mapping approach is available in the last version of (de Givry *et al.*, 2004), a free[1] mapping software in C++, including LKH (Helsgaun, 2000) for maximum likelihood computation.

1 INTRODUCTION

Since the discovery of the molecular basis of genes, the time devoted to mapping has dramatically increased, reaching its apogee with the advent of whole genome sequence projects. Although the complete sequence provides the ultimate map of a genome, the problem of constructing a map from experimental data remains an active area of research (Bø *et al.*, 2002; Crane and Crane, 2004; Mester *et al.*, 2003; Wu *et al.*, 2003). Maps are key to the study of organisms that are not planned to be sequenced in the near future. In addition, the availability of detailed maps offers great advantage in the process of whole genome sequencing (Havlak *et al.*, 2004). The production of whole genome sequences therefore doesn't dismiss the need for gene mapping. It suggests however alternative mapping strategies. Having in hand the exhaustive gene catalog of a completely sequenced genome, makes it possible to take advantage of the conservation of chromosome segments with a related genome of interest. This approach, also called comparative mapping, has been extensively used for many years as a guideline for the construction of maps in animals as well as in plants (Bowers *et al.*, 2005; O'Brien *et al.*, 1990). The comparative mapping strategy is also of great value in the context of whole genome sequence assembly (Havlak *et al.*, 2004; Pop *et al.*, 2004).

We propose here a novel approach to gene mapping, in the context of radiation hybrid (RH) mapping, provided that a closely related completely sequenced genome is available. Unlike the traditional approach, the map of the reference organism is used at the very first step of marker ordering for the construction and evaluation of the candidate maps. Although devised in the context of RH mapping, we believe that the proposed method applies equally to other mapping strategies such as genetic mapping. Sections 2 and 3 describe a new statistical model that takes into account both the experimental RH data and the order in a related organism. Section 4 deals with the algorithmic aspects of searching the space of all possible maps, trying to find the best one according to the predefined criterion, without evaluating the $\frac{n!}{2}$ possible marker orders. Finally, the interest of this approach is evaluated on both simulated and real data, showing a significant improvement in map quality over the traditional approach.

2 THE STATISTICAL MODEL

Our presentation is restricted to the case of radiation hybrid mapping which can be described by a simple statistical model (Boehnke *et al.*, 1991). In order to focus our presentation on the new comparative approach for marker ordering, the RH mapping technique and the associated statistical model are described in details in the appendix of this paper. We implicitly develop our comparative approach principle in the particular case of haploid error-free data due to the approximation using 2-point likelihood (see below and appendix). The comparative principle is however not closely interlinked to the 2-point likelihood approach and could be extended to other approaches of RH mapping (see discussion).

We note A the reference organism and B the organism of interest. For B an RH data set X for n markers is available. We make the assumption that there is a one-to-one correspondence between the markers in B and their orthologs in A. The complete genome sequence of A provides a map π_A of these markers in A. Our aim is to build a map, identified by a marker permutation π, for the n

[*] to whom correspondence should be addressed

[†] Both authors contributed equally to this work

[1] The LKH part is free for academic use only.

markers of organism B. Let $P(X|\pi,\theta)$ denote the likelihood of the data for a given order π and a set of parameters (nuisance parameters such as the retention fraction and breakage frequencies for radiation hybrids). In the traditional maximum likelihood approach, the likelihood associated with each order is the maximum over all possible values of θ:

$$L(\pi|X) = \max_{\theta} P(X|\pi,\theta) \tag{1}$$

and the candidate map is the order π that maximizes this likelihood. Although the situation is generally complicated by the fact that the estimation of θ depends on the particular choice of π, we will consider an approximation of this likelihood, using the product of 2-point maximum likelihoods strictly equivalent to the likelihood only for haploid error-free data, which breaks this dependencies between θ and π (see appendix and Agarwala *et al.*, 2000 for a detailed description of 2-point likelihoods and a discussion of the relevance of such an approximation).

Using this approximation, we can consider the likelihood of the data as depending solely on π:

$$L(\pi|X) = P_\theta(X|\pi) \tag{2}$$

and proceed to the Bayesian inversion

$$P_\theta(\pi|X) = \frac{P_\theta(X|\pi)P(\pi)}{\sum_\pi P_\theta(X|\pi)P(\pi)} \propto P_\theta(X|\pi)P(\pi) \tag{3}$$

In this framework, the information provided by the existing map π_A for the corresponding orthologous genes in A can be incorporated by defining a non-uniform prior distribution on the possible orders for the map in B. We suppose that the probability of an order is a function of its evolutionary distance to the reference map, measured with the number of *breakpoints* between the proposed order π and the reference order π_A. This distance, denoted as k, is the number of adjacent markers in π which are not adjacent in π_A.

As the choice of a particular order π implies a unique breakpoint distance with the reference map, the previous equation can be written as

$$P_\theta(\pi|X) \propto P_\theta(X|\pi)P(\pi|k)P(k) \tag{4}$$

where k is the number of breakpoints. Assuming a Poisson prior for the law of breakpoints $P(k) = P_\lambda(k)$, the only expression which is not yet determined is $P(\pi|k)$, the likelihood of a given order for a fixed number of breakpoints. For a given breakpoint distance, we assume that all the orders are equally probable and hence follow a uniform distribution. The likelihood takes the following form:

$$P(\pi|k) = \frac{1}{O_n(k)} \tag{5}$$

where $O_n(k)$ denotes the number of different orders having exactly k breakpoints with the identity permutation of size n. We show in the next section how to compute this number. For $n = 100$ markers for example, we have $O_n(k) = 1, 293, 79349, 19071365, \cdots$ for $k = 0, 1, 2, 3, \cdots$. Intuitively, this new objective function states that the risk of making an additional breakpoint to the reference order is taken if the gain in likelihood of the data balances the risk of jumping from a search space of size $O_n(k)$ to a search space of size $O_n(k+1)$ (and from k to $k+1$ in the Poisson law). In the sequel, we note L_c the likelihood including the comparative information defined by (4) and L the usual likelihood (2). Finding the map maximizing L will be termed simple 2-point RH approach while searching for the one maximizing L_c will be termed comparative 2-point approach.

3 NUMBER OF ORDERS AT A GIVEN BREAKPOINT DISTANCE

We describe first the case of single chromosome genomes and then extend our results to the case of multiple chromosomes. Since complete map reversals define the same order, a permutation and its complete reversal will be considered equivalent in the sequel.

3.1 Single chromosome genomes

We assume that the reference order π_A is the identity permutation. Consider an arbitrary permutation π. We define a *segment* in this permutation as a maximal set of markers in the permutation that contains no breakpoint with π_A. The single order exempt of breakpoints with π_A is π_A itself. With a fixed breakpoint, the two resulting segments can be ordered in 3 different ways:

$$1\cdots j \mid n\cdots j{+}1 \quad j\cdots 1 \mid n\cdots j{+}1 \quad n\cdots j{+}1 \mid 1\cdots j$$

In the general case we proceed by induction on n, the size of the permutations and k the number of breakpoints. When a segment is reduced to a single marker, the marker is said to be isolated. When adding the new marker n in an existing configuration, 3 possible outcomes must be considered

(0) 0 breakpoint is created when n is inserted before or after marker $n-1$, at the border of a segment;

(1) 1 breakpoint is created when n is inserted (i) inside a segment next to marker $n-1$, (ii) at the position of an existing breakpoint or (iii) at one of the two ends (borders) of the permutation except next to $n-1$;

(2) 2 breakpoints are created when n is inserted anywhere inside in a segment, except next to $n-1$.

Note that the knowledge of the position of $n-1$, isolated or not, in a central position or at one of the two extremities, is the only relevant information needed prior to the introduction of n. Consider the set of all permutations with k breakpoints with the reference order. In order to compute the cardinality of this set, we define a partition into four components according to the position of marker n (see figure 1):

- $I_n^b(k)$: permutations with n isolated at one of the two extremities of the permutation
- $I_n^c(k)$: permutations with n isolated but in a central position (anywhere except at the extremities)
- $S_n^b(k)$: permutations with n at one of the extremities of the permutation and at the border of a segment
- $S_n^c(k)$: permutations with n on the border of a segment but in a central position

Using the same notation for a set and its cardinality, let $O_n^b(k) = I_n^b(k) + S_n^b(k)$ and $O_n^c(k) = I_n^c(k) + S_n^c(k)$. We have $O_n(k) = O_n^b(k) + O_n^c(k)$. The following induction relations enable to compute the number of permutations sharing a fixed number of

$$\begin{array}{ll}
1\,2\cdots j \;\mid\; n{-}3\cdots j{+}1 \;\mid\; n{-}1\;n{-}2 & \in S^c_{n-1}(2) \\
\\
n \;\mid\; 1\,2\cdots j \;\mid\; n{-}3\cdots j{+}1 \;\mid\; n{-}1\;n{-}2 & \in I^b_n(3) \\
1\,2\cdots j \;\mid\; n \;\mid\; n{-}3\cdots j{+}1 \;\mid\; n{-}1\;n{-}2 & \in I^c_n(3) \\
1 \;\mid\; n \;\mid\; 2\cdots j \;\mid\; n{-}3\cdots j{+}1 \;\mid\; n{-}1\;n{-}2 & \in I^c_n(4) \\
1\,2\cdots j \;\mid\; n{-}3\cdots j{+}1 \;\mid\; n{-}1\;n \;\mid\; n{-}2 & \in S^c_n(3) \\
1\,2\cdots j \;\mid\; n{-}3\cdots j{+}1 \;\mid\; n\;n{-}1\;n{-}2 & \in S^c_n(2)
\end{array}$$

Fig. 1. An example of initial permutation with $n-1$ elements followed by 5 different possibilities of inserting marker n illustrating the sets $I^b_n(k)$, $I^c_n(k)$ and $S^c_n(k)$. The only set not shown, $S^b_n(k)$, can be illustrated by simply reverting the rightmost segment of the last permutation. Breakpoints are represented as vertical bars.

breakpoints with the identity permutation:

$$\left\{\begin{array}{lcl}
I^b_n(k) & = & O^b_{n-1}(k-1) + 2O^c_{n-1}(k-1) \\
I^c_n(k) & = & (k-1)O_{n-1}(k-1) + (n-k)O_{n-1}(k-2) \\
 & & -S^c_n(k-1) \\
S^b_n(k) & = & O^b_{n-1}(k) \\
S^c_n(k) & = & I^b_{n-1}(k) + 2I^c_{n-1}(k) + S^c_{n-1}(k) \\
 & & +S^b_{n-1}(k-1) + S^c_{n-1}(k-1)
\end{array}\right.$$

A configuration with n isolated at one border can only be obtained through the operation described in (1)(iii) leading to the induction relation for $I^b_n(k)$. The other relations can be derived by a similar analysis. Setting all quantities to 0 for $k < 0$ and using initial values of $I^b_2(0) = I^c_2(0) = S^c_2(0) = 0$, $S^b_2(0) = 1$, a simple dynamic programming procedure can compute all $O_n(k)$ values for $n \leq N$ and $k \leq N-1$ in quadratic time.

3.2 Multiple chromosome genomes

Generalization to multiple chromosomes implies to distinguish obligate breakpoints created by the concatenation of markers from different chromosomes from other breakpoints. If the chromosome maps of the reference organism are arbitrarily concatenated before the numbering process, some adjacencies in this new reference map must be considered as breakpoints. Let $n_1, \ldots, n_r$ denote the number of markers on the chromosomes $1, \ldots, r$ of the reference organism A involved in a single linkage group of the genome of interest B. In the induction process, when incorporating the first marker from a new chromosome in the permutation, i.e of the type $\sum_{i=1}^{j} n_i + 1$ for $j = 1, \ldots, r-1$, one has to ensure that an additional breakpoint is always created. The number of permutations at a given breakpoint distance k when n spans the $n_1 + \cdots + n_r$ markers uses the same induction relations as defined in 3.1 with the following modifications for the particular cases where $n = \sum_{i=1}^{j} n_i + 1$ ($j = 1, \ldots, r-1$):

$$\left\{\begin{array}{lcl}
I^b_n(k) & = & 2O_{n-1}(k-1) \\
I^c_n(k) & = & (k-1)O_{n-1}(k-1) + (n-k)O_{n-1}(k-2) \\
S^b_n(k) & = & S^c_n(k) = 0
\end{array}\right.$$

4 MAXIMUM LIKELIHOOD COMPUTATION REDUCED TO SOLVING A TSP

In order to compute efficiently the maximum likelihood estimation of π under the model defined by (4) we reduce the corresponding optimization problem to the Traveling Salesman Problem (TSP). The principle of this reduction is to write the likelihood of an order as a weighted path visiting all the markers in that order. Practically, this entails constructing a distance measure on the set of markers. We consider the log-likelihood

$$\log P_\theta(\pi|X) = \log P_\theta(X|\pi) + \log\left[P(\pi|k)P_\lambda(k)\right] + C$$

and follow the approach of Agarwala *et al.*, 2000 for the first term:

$$\log P_\theta(X|\pi) = \log[t_{x_1} \times t_{x_1,x_2} \times \cdots \times t_{x_{n-1},x_n} \times t_{x_n}]$$

where $t_{x_i,x_{i+1}}$ is the contribution of the radiation hybrid data associated with marker interval $[x_i, x_{i+1}]$ to the likelihood of the map defined by π (see appendix and Agarwala *et al.*, 2000). Due to the exponential nature of $O_n(k)$, the additive contribution of each interval for the breakpoint counterpart of the likelihood is obtained by a linear regression $y = a + bk$ on the data $y = log\left[P(\pi|k)P_\lambda(k)\right]$ ($k = 0, \ldots, n-1$) using the exact computation of $P(\pi|k)$ given by the recurrence formula[2] of section 3 and a predefined parameter λ for the Poisson law. Setting

$$w_{x,y} = \log t_{x,y} + b \times 1_{x|y} \qquad (6)$$

with

$$1_{x|y} = \left\{\begin{array}{ll} 0 & \text{if } x \text{ and } y \text{ are adjacent in the reference order} \\ 1 & \text{otherwise} \end{array}\right.$$

fully defines the TSP reduction

$$\log P(\pi|X) = \sum_{i=0}^{n} w_{x_i,x_{i+1}}$$

with $w_{x_0,x_1} = \log t_{x_1} + a$ and $w_{x_n,x_{n+1}} = \log t_{x_n}$.

Solving the resulting TSP instances can be done in several ways using either complete methods such as branch and cut or heuristic methods. We have tried both state-of-the-art complete and/or heuristic methods available in CONCORDE (Applegate *et al.*, 1998) and LKH (Helsgaun, 2000). The likelihood computation has been implemented above the CARTHAGÈNE (de Givry *et al.*, 2004) C++ and LKH (Helsgaun, 2000) C libraries.

For the purpose of comparing the performance of the comparative 2-point and the simple 2-point approaches, all the TSP instances in the sequel are resolved using the LKH heuristic of Helsgaun, 2000.

5 SIMULATED RADIATION HYBRID DATA SETS

The following protocol is used to generate RH data sets and reference orders. N markers are randomly distributed according to the uniform distribution on a chromosome of size S Ray giving rise to the target map or true order. The inter-marker expected breakage frequencies $\theta_{i,i+1} = 1 - e^{-\delta_{i,i+1}}$ corresponding to the inter-marker distance $\delta_{i,i+1}$ are subsequently used to generate random RH data sets for I individuals according to the haploid equal retention model (Boehnke *et al.*, 1991) with the retention fraction r, a false positive/negative error rate p_{error} and a proportion of missing data p_{miss}. Finally, a reference order is generated by applying

[2] This computation can be easily precomputed once for different number of markers and the results made available as a table.

a sequence of rearrangement events (reversal, transposition, inverted transposition) on the true order with an expected number of events, or *evolutionary distance*, set to E (Moret *et al.*, 2001). Note that an inversion creates 2 breakpoints while the two other rearrangements produce 3 breakpoints so that the expected number of breakpoints is $\frac{8}{3} \times E$ if the 3 rearrangements are equiprobable. In addition, a parameter H controls the proportion of known orthologous relationships which are randomly selected among the N possible ones. Whenever a marker x has no identified ortholog, $1_{x|y}$ is set to 1 in (6) mimicking therefore a breakpoint. In the experiments, we tried the following values for the generator parameters: $N = 100, S \in [4, 40], I = 40, r = 37\%, p_{miss} = 3\%, p_{error} = 3\%, E \in \{2, 4, 8\}, H \in [0, 100]$. Each reported experimental result is a mean over 100 randomly generated RH data sets and reference orders by following the previous protocol with a fixed value of the parameters.

In order to assess the effectiveness of our comparative mapping approach, two performance metrics were used to evaluate the accuracy of the proposed maps: (a) proportion of the correctly reconstructed maps, (b) the longest increasing subsequence (LIS). Since, in our simulations, the true order is represented by the identity permutation, the longest increasing subsequence of a candidate order indeed measures how accurate the candidate map is. Let $\pi = (\pi_1 \dots \pi_n)$, the longest increasing subsequence is the largest subset $(\pi_{i_1}, \dots \pi_{i_r})$ such that $\pi_{i_1} < \cdots < \pi_{i_r}$. We note $LIS(\pi) = r$ the size of this set. LIS computation is folklore in algorithmic and has already been used for the evaluation of mapping strategies (Bø *et al.*, 2002).

6 SIMULATION RESULTS

The robustness of our approach was studied with respect to 3 different factors: the influence of the evolutionary distance with the reference genome, the influence of chromosome size (or marker density), and the proportion of known orthology relationships within the dataset.

As expected, the availability of a complete map for a closely related organism significantly improves mapping efficiency, the improvement being dependent on the evolutionary distance between the two maps (figure 2). In these experiments, for S=15 Ray for example, the true order was never found by the simple 2-point RH mapping approach while the comparative 2-point RH mapping recovered the true order from 16 up to 65 times depending on the evolutionary distance. The proportion of correctly reconstructed order is however too crude for a metric: as the number of marker increases, the probability of recovering the true order decreases rapidly (see Ben-Dor and Chor, 1997 for a formal analysis of this behavior). The LIS criterion in contrast, by measuring the size of the largest subset correctly ordered in the proposed map, enables to quantify the distance to the true map. Comparison using this criteria, shown in figure 3, confirms the benefit of the comparative mapping approach. Less than 10% of the markers were wrongly positioned when the chromosome size belongs to the interval $[5, 15]$ Ray in the case of comparative 2-point RH mapping with a medium-size evolutionary distance $E = 4$. On the contrary, simple 2-point RH approach got 33% of incorrectly positioned markers at its best ($S = 10$ Ray).

As shown in both figures, there is a clear influence of marker density on the mapping accuracy. Indeed, the linkage between markers is respectively loose and tight for large and small chromosomes. In both extreme cases, the RH data set is not very informative for the purpose of ordering and the reference order provides therefore a valuable information. The robustness of the comparative approach to marker densities, due to the fact that the evolutionary breakpoints are independent from the number of markers, is of great value when the objective is to produce dense maps.

In our experiments, the expected number of breakpoints between the true order and the reference order, or λ, was set to 1 in the Poisson prior $P_\lambda(k)$. This value is generally unknown for the mapping process. However, no clear improvement in terms of both criteria was observed when using for each instance the exact number of breakpoints, available in the context of simulation (results not shown).

Finally, we studied the impact of diminishing the proportion of known orthologous relationships. Figure 4 shows the results for the LIS criterion on a 10 Ray chromosome with $H \in [0, 100]$.

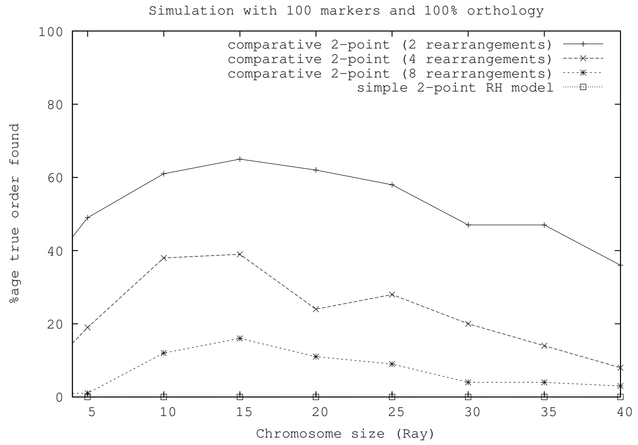

Fig. 2. Effect of marker density (chromosome size) and evolutionary distance on the percentage of true order found. Simulated radiation hybrid data sets with 100 markers randomly distributed on a single chromosome the size of which varies from 4 Ray to 40 Ray. For the comparative approach, the reference order of 100 orthologous markers is at an evolutionnary distance of respectively 2, 4 and 8 (see text).

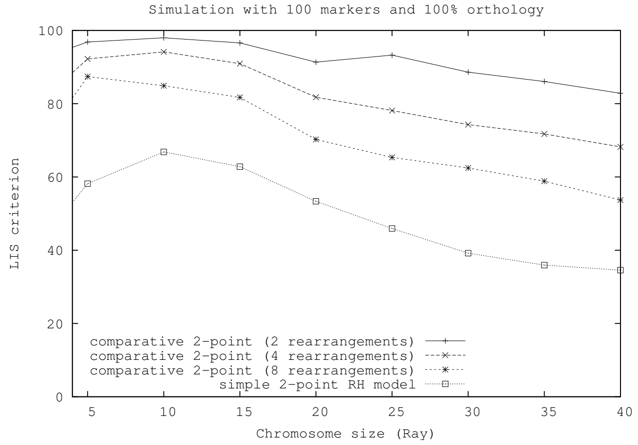

Fig. 3. Effect of marker density (chromosome size) in terms of the longest increasing subsequence (LIS) criterion.

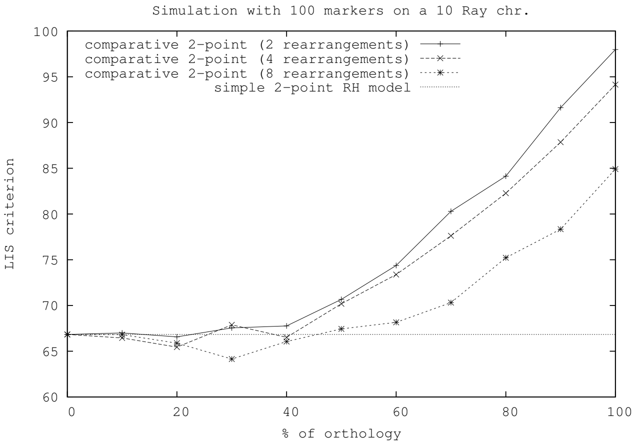

Fig. 4. Effect of the number of orthologous relationships in terms of the LIS criterion. The horizontal line correspond to the simple 2-point RH model.

When $H = 0$, the method reduces to a simple 2-point RH mapping approach. When H was greater than $40 - 50\%$, we observed a clear improvement in terms of map quality for the comparative 2-point mapping approach compared to the simple 2-point RH model. Below this threshold, the knowledge of a partial reference order can be counterproductive, especially if the evolutionary distance is high. An explanation for this negative result, in the case of $E = 8$ and $H = 30$, is the fact that the number of breakpoints was close to the number of orthologous relationships (in the experiments, $E = 8$ corresponds to 18.74 breakpoints for 100 markers and still 11.27 breakpoints for $H = 30$ orthologous markers) and the TSP reduction provided a coarse approximation because of the arbitrary weight $w_{x,y}$ assigned in the absence of orthologs (see section 5).

7 EXPERIMENTS WITH A DOG RADIATION HYBRID DATA SET

In order to test the efficiency of our method on a real example, we applied this comparative approach to the construction of a RH map of a whole canine chromosome (CFA2 - figure 5) using a set of 426 markers typed on the RHDF9000 dog radiation hybrid panel (Hitte *et al*., 2005). The human genome sequence was used as a reference map. As the RH markers consisted essentially in gene-based fragments, the corresponding orthologous position was determined for all 426 markers using a simple reciprocal best hit principle with the human gene catalog (Kirkness *et al*., 2003). The 426 markers cover the entire canine chromosome 2 (87 Mb) corresponding to a marker every 200kb on average. We constructed RH maps of CFA2 using both the simple 2-point RH method and the comparative 2-point approach. The comparative mapping approach showed a clear improvement over the simple 2-point method in that the proposed map was in better agreement with the dog genome sequence (Lindblad-Toh *et al*., 2005) than the map built using the simple 2-point RH mapping approach. An illustration of this improvement is given in figure 5.

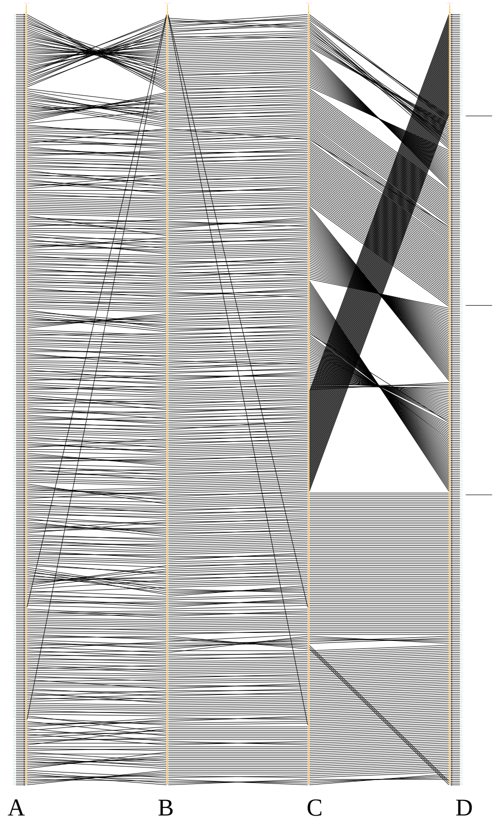

Fig. 5. Consensus maps of 426 markers for Dog Chromosome 02 found by (from left to right) simple 2-point RH mapping (A), sequence assembly (B), comparative 2-point RH mapping (C), and following the Human genome (4 segments, 100% orthology) order (D). LIS criteria are: $LIS(A) = 212$ and $LIS(C) = 317$. Computing maps A and C took less than 10 seconds each on a Pentium IV 2 GHz.

8 DISCUSSION

As frequently pointed out (Agarwala *et al.*, 2000; Ben-Dor and Chor, 1997; Ben-Dor *et al*., 2000), and illustrated in the previous sections, the major impediments to producing dense high-quality RH maps are the panel resolution power and experimental data quality and not computation. The traditional avenue to overcome this problem is the construction of framework maps: only a subset of markers is ordered with the counterpart that the proposed order is significantly better (usually in a ratio of 1000:1 of the likelihood) than all other orders with the same markers. Unfortunately this has a cost as the remaining unplaced markers (typically 50 to 80% of initial dataset) are then positioned into bins of confindence leading to a placement map which may encounter many discrepancies with the true order. We propose here a novel approach that defines a new objective function which takes into account the information provided by a closely related completely sequenced genome: a genome for which an exhaustive map is available. The efficiency of the method is clearly dependent on the evolutionary distance between the reference genome and the genome one wishes to map but also on the quality of orthologous relationships. The proposed objective function performs significantly better than the simple 2-point likelihood on both simulated and real data for the range of parameters typically observed for the mammalian species (relative low number of breakpoints and ability to detect orthology relationships). While the experiments are here restricted to the comparison with the simple 2-point approach of RH mapping, our purpose was to demonstrate the benefits of incorporating comparative mapping information in an existing statistical framework, principle which should be applicable to other RH mapping strategies.

We would like to emphasize that the comparative approach could also be of interest within a species, in the context of genetic mapping, when a map of the markers already exists and one wishes to incorporate this prior knowledge into the statistical model while studying additional experimental data set corresponding to another breed or cultivar for example.

Future directions should address the case of multiple closely related sequenced genomes.

APPENDIX

A radiation hybrid experiment can be rapidly sketched as follows: cells from the organism under study are irradiated. The radiation breaks the chromosomes at random locations into separate fragments. A random subset of the fragments is then rescued by fusing the irradiated cells with normal rodent cells, a process that produces a collection of hybrid cells. The resulting clone may contain none, one or many chromosome fragments. This clone is then tested for the presence or absence of each of the markers. This process is performed a large number of times producing a radiated hybrid panel, previously called RH data set in Section 2.

More formally, given N markers and I hybrid cells, a panel is a collection of I vectors of identical size N, containing boolean values 0 for the absence of a marker and 1 for its presence.

The radiation breakage frequencies between two markers, estimated from their co-occurrence pattern in a panel of radiated hybrid cells (possible configuration patterns are (11), (10), (01), or (00) in vectors), provides, in a similar manner to the recombination fraction in genetic mapping, a measure of the distance separating the markers. The distance unit is the *Ray*, corresponding to a segment length where one break is expected. Let r denote the retention fraction and θ the breakage probability between markers y and z. The conditional probabilities of the status Z of marker z, knowing the status Y of marker y, is given by the following formulas (Boehnke *et al.*, 1991):

$$\begin{cases} P(Z=1\,|\,Y=1) & = & p_{1|1} & = & (1-\theta)+\theta r \\ P(Z=1\,|\,Y=0) & = & p_{1|0} & = & \theta r \\ P(Z=0\,|\,Y=1) & = & p_{0|1} & = & \theta(1-r) \\ P(Z=0\,|\,Y=0) & = & p_{0|0} & = & (1-\theta)+\theta(1-r) \end{cases}$$

Let $p_1 = r$ and $p_0 = 1 - r$. The probability of observing a hybrid with marker y present and marker z absent is for example $p_1 p_{0|1}$ and, by a simple refactorization, the likelihood for the data Y and Z associated to a panel of hybrids takes the following form

$$L(Y,Z|\theta) = L(Y|\theta)L(Z|Y,\theta) \tag{1}$$

with $L(Y|\theta) = p_0^{n_{0\cdot}} p_1^{n_{1\cdot}}$ and the 2-point likelihood

$$L(Z|Y,\theta) = p_{1|1}^{n_{11}} p_{1|0}^{n_{10}} p_{0|1}^{n_{01}} p_{0|0}^{n_{00}}$$

where θ is the extended set of parameters (θ, r) and n_{ij} the cardinality of the different configurations outlined above with $n_{i\cdot}$ the marginal cardinality $n_{i0} + n_{i1}$.

The maximum likelihood estimate of r is simply the ratio of the total number of 1s to the total number of 1s and 0s (the number of 1 in the panel divided by $I \times N$). The maximum likelihood estimate of the breakage frequency θ can be derived analytically from (1) (see for example Agarwala *et al.*, 2000 for a detailed description).

The natural mathematical framework for radiation hybrid mapping depicts the succession of loci on a chromosome as successive steps of a Markov chain. The likelihood of a hybrid for a given order $\pi = (x_1 \cdots x_n)$ is the probability to observe the data X under the associated Markov model

$$L(X|\theta,\pi) = P(X_1|\theta_1) \prod P(X_i|X_{i-1},\theta_i) \tag{2}$$

Considering simultaneously all the hybrids, the likelihood can be rewritten in the following form

$$L(X|\theta,\pi) = L(X_1|\theta_1) \prod L(X_i|X_{i-1},\theta_i) \tag{3}$$

where θ_i is the set of parameters restricted to the interval between two consecutive markers. In particular, the maximization over the parameters θ on one side and the order parameter π can be conducted independently. We call $L_\theta(X_i|X_{i-1})$ the 2-point maximum likelihoods :

$$L_\theta(X_i|X_{i-1}) = \max_{\theta_i} L(X_i|X_{i-1},\theta_i)$$

This value can be computed using the maximum likelihood estimation procedure of r and θ described above. The likelihood of an order π can be computed directly from these maximum likelihoods:

$$L_\theta(X|\pi) = L_\theta(X_0) \prod L_\theta(X_i|X_{i-1}) \tag{4}$$

A reduction to a symmetric TSP implies a symmetric treatment of the different loci, dropping the reference to θ for simplicity, we note

$$t_x = \sqrt{L(X)} \text{ and } t_{x,y} = \frac{L(X,Y)}{t_x t_y}$$

In a straightforward manner

$$\begin{aligned} t_{x_1}\left(\prod t_{x_i,x_{i-1}}\right) t_{x_n} &= t_{x_1}\left(\prod \frac{L(X_i|X_{i-1})t_{x_{i-1}}}{t_{x_i}}\right) t_{x_n} \\ &= L(X_1) \prod L(X_i|X_{i-1}) \end{aligned}$$

therefore

$$L(X|\pi) = t_{x_1} \times t_{x_0,x_1} \times \cdots \times t_{x_{n-2},x_{n-1}} \times t_{x_n} \tag{5}$$

and the TSP reduction is completed (see Ben-Dor *et al.*, 2000; Agarwala *et al.*, 2000 for analytical formulas).

In general however, the correct Markov formalization implies some hidden properties (model including the diploid nature of the genome or typing errors) and equality (4) no longer holds. It has been argued that the product of 2-point maximum likelihoods provides however a good approximation of the likelihood (Ben-Dor *et al.*, 2000; Agarwala *et al.*, 2000).

ACKNOWLEDGEMENT

Thanks to Lauranne Duquenne who wrote the extension of CARTHAGÈNE taking into account a reference order data set and Brigitte Mangin for fruitful discussions concerning the statistical model.

REFERENCES

Agarwala,R. *et al.* (2000) A fast and scalable radiation hybrid map construction and integration strategy. *Genome Research*, **10**, 350–364.

Applegate,D. *et al.* (1998) On the solution of traveling salesman problems. In *Proc. of ICM III*, 645–656.

Ben-Dor,A. and Chor,B. (1997) On constructing radiation hybrid maps. *J. Comp. Biol.*, **4**, 517–533.

Ben-Dor,A. *et al.* (2000) RHO – radiation hybrid ordering. *Genome Research*, **10**, 365–378.

Bø,T. H. *et al.* (2002) A fast top-down method for constructing reliable radiation hybrid frameworks. *Bioinformatics*, **18**, 11–18.

Boehnke,M. *et al.* (1991) Statistical methods for multipoint radiation hybrid mapping. *Am. J. Hum. Genet.*, **49**, 1174–1188.

Bowers,J.E. et al. (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc. Natl. Acad. Sci. USA*, **102**, 13206–13211.

Crane,C.F. and Crane,Y. M. (2004) A nearest-neighboring-end algorithm for genetic mapping. *Bioinformatics*, **21**, 1579–1591.

de Givry,S. *et al.* (2004) CARTHAGENE: multipopulation integrated genetic and radiated hybrid mapping. *Bioinformatics*, **21**, 1703–1704. www.inra.fr/mia/T/CarthaGene.

Havlak, P. *et al.* (2004) The atlas genome assembly system. *Genome Res.*, **14**, 721–732.

Helsgaun,K. (2000) An effective implementation of the lin-kernighan traveling salesman heuristic. *European Journal of Operational Research*, **126**, 106–130. www.dat.ruc.dk/˜keld/research/LKH.

Hitte,C. *et al.* (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat. Rev. Genet.*, **6**, 643–648.

Kirkness,E.F. *et al.* (2003) The dog genome: Survey sequencing and comparative analysis. *Science*, **301**, 1898–1903.

Lindblad-Toh,K. *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.

Mester,D. *et al.* (2003) Constructing Large-Scale Genetic Maps Using an Evolutionary Strategy Algorithm. *Genetics*, **165**, 2269–2282.

Moret,B.M. *et al.* (2001) New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, **17**(Suppl. 1), S165–S173,.

O'Brien,S.K. and Graves,J.A. (1990) Report of the committee on comparative gene mapping. *Cytogenet. Cell Genet.*, **55**, 406–433.

Pop,M. *et al.* (2004) Comparative genome assembly. *Brief. in Bioinformatics*, **5**, 237–248.

Wu,J. *et al.* (2003) Monte carlo simulations on marker grouping and ordering. *Theor. Appl. Genet.*, **107**, 568–573.