Boosting EM for Radiation Hybrid and Genetic Mapping

T. Schiex[1], P. Chabrier[1], M. Bouchez[1] and D. Milan[2]

[1] Biometry and AI Lab.
[2] Cellular Genetics Lab.
INRA, Toulouse, France.
{Thomas.Schiex,Denis.Milan}@toulouse.inra.fr

Abstract. Radiation hybrid (RH) mapping is a somatic cell technique that is used for ordering markers along a chromosome and estimating physical distances between them. It nicely complements the genetic mapping technique, allowing for finer resolution. Like genetic mapping, RH mapping consists in finding a marker ordering that maximizes a given criteria. Several software packages have been recently proposed to solve RH mapping problems. Each package offers specific criteria and specific ordering techniques. The most general packages look for maximum likelihood maps and may cope with errors, unknowns and polyploid hybrids at the cost of limited computational efficiency. More efficient packages look for minimum breaks or two-points approximated maximum likelihood maps but ignore errors, unknowns and polyploid hybrids.
In this paper, we present a simple improvement of the EM algorithm [5] that makes maximum likelihood estimation much more efficient (in practice and to some extent in theory too). The boosted EM algorithm can deal with unknowns in both error-free haploid data and error-free backcross data. Unknowns are usually quite limited in RH mapping but cannot be ignored when one deals with genetic data or multiple populations/panels consensus mapping (markers being not necessarily typed in all panels/populations). These improved EM algorithms have been implemented in the CarHTaGène software. We conclude with a comparison with similar packages (RHMAP and MapMaker) using simulated data sets and present preliminary results on mixed simultaneous RH/genetic mapping on pig data.

1 Introduction

Radiation hybrid mapping [4] is a somatic cell technique that is used for ordering markers along a chromosome and estimating the physical distances between them. It nicely complements alternative mapping techniques especially by providing intermediate resolutions. This technique has been mainly applied to human cells but also used on animals, eg. [6].

The biological experiment in RH mapping 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.

The algorithmic analysis which follows the biological experiment, based on the retention patterns of the markers observed, aims at finding the most likely linear ordering of the markers on the chromosome along with distances between markers. The underlying intuition is that the further apart two markers are, the most likely it is that the radiation will create one or more breaks between them, placing the two markers on separate chromosomal fragments. Therefore, close markers will tend to be more often co-retained than distant ones. Given an order of the markers, the retention pattern can therefore be used to estimate the pairwise physical distances between them.

Two fundamental types of approaches have been used to evaluate the quality of a possible marker's permutation. The first, crudest approach, is a non parametric approach, called "obligate chromosomal breaks" (OCB), that aims at finding a permutation that minimizes the number of breaks needed to explain the retention pattern. This approach is not considered in this paper. The second one is a statistical parametric method of maximum likelihood estimation (MLE) using a probabilistic model of the RH biological mechanism. Several probabilistic models have been proposed for RH mapping [9] dealing with polyploidy, errors and unknowns. In this paper, we are only interested in a subset of the models that are compatible with the use of the EM algorithm [5] for estimating distances between markers. According to our experience, the simplest "equal retention model" is the most frequently used in practice and also the most widely available because it is a good compromise between efficiency and realism. Such models are used in the RHMAP and RHMAPPER packages. More recently, more efficient approximated MLE versions based on two points estimation have also been used [2] but they don't deal with unknowns and won't be considered in the sequel.

The older but still widely used genetic mapping technique [10] exploits the occurrence of cross-overs during meiosis. As for RH mapping, the underlying intuition is that the further apart two markers are, the most likely it is that a cross-over will occur in between. Because cross-overs cannot be directly observed, the indirect observation of allelic patterns in parents and children is used to estimate the genetic distance between them. There is a long tradition of using EM in genetic mapping [7]. This paper will focus on RH mapping but we must mention that the improvements presented in this paper have also been applied to genetic mapping with backcross pedigree. Actually, genetic and RH data nicely complement each other for *ordering* markers. Genetic data leads to myopic ordering: set of close markers cannot be reliably ordered because usually no recombination can be observed between them. On the contrary RH data leads to hypermetropic ordering: set of closely related markers can be reliably ordered but distant groups are sometimes difficult to order because too many breaks occurred between them. Dealing with unknown is unavoidable in genetic mapping since markers may be uninformative.

In either RH or genetic mapping, the most obvious computational barrier is the shear number of possible orders. For n markers, there are $\frac{n!}{2}$ possible orders (as an order and its reverse are equivalent), which is too large to search exhaustively, even for moderate values of n. In the simplest case of error-free unknown-free data, it has been observed by several authors that the MLE ordering problem is equivalent to the famous traveling salesman problem [13, 1], an NP-hard problem. The ordering techniques used in existing packages go from branch and bound [2], to local search techniques and more or less greedy heuristics. In all cases, finding a good order requires a very large number of MLE calls. In practice, the cost of EM execution is still too heavy to make branch and bound or local search methods computationally usable on large data sets and the greedy heuristic approach remains among the most widely used in practice.

In this paper, we show how the EM algorithm for RH/genetic mapping can be sped up when it is applied to data sets where each information is either completely informative or completely uninformative. This is the case eg. for error-free haploid RH data with unknown or error-free backcross data with unknowns. In practice, for RH mapping, it has been observed in [9] that "analyzing polyploid radiation hybrids as if they were haploid does not compromise the ability to order markers" which makes this restriction to haploid data quite reasonable. For genetic mapping, most phase-known data can be reduced (although with some possible loss of information) to backcross data. In practice, we have applied it to diploid RH data and complex animal (pig) pedigree with very limited discrepancies (see section 5) and a speed-up factor of two orders of magnitude.

Interestingly, this boosted EM algorithm is especially adapted to the unknown/known patterns that appear in multiple population/panels mapping when a single consensus map is built: when two or more data sets are available for the same organism (and with a similar resolution for RH), a possible way to build a consensus map is to merge the two data sets in one. markers that are not typed in one of the 2 data sets are marked as unknown in this case. In this case, we show that each iteration of the EM algorithm may be in $O(n)$ instead of $O(n.k)$, where n is the number of markers and k the number of individuals typed.

These improved EM algorithms along with several TSP-inspired ordering strategies for both framework and comprehensive mapping have been implemented in the free software package Car$^{\mathrm{T}}_{\mathrm{H}}$aGène [13] which allows for multiple population/panels mapping using either shared or separate distance estimation for each pair of data sets. This allows, among other, for mixed genetic/RH mapping (with separate estimations of genetic and RH distances) which nicely exploits the complementarity of genetic and RH data.

2 The EM algorithm for RH data

In this section, we will explain how EM can be optimized to deal with haploid RH data sets with unknowns. This optimization also applies to backcross genetic data sets with missing data. It has been implemented in the genetic/RH mapping software CarHTaGène [13] but has never been described in the literature before.

Suppose that genetic markers M_1,... M_n are typed on k radiation hybrids. The observations for a given hybrid, given the marker order $(M_1, \ldots, M_n)$ can be written as a vector $x = (x_1, \ldots, x_n)$ where $x_i = 1$ if the marker M_i is typed and present, $X_i = 0$ if the marker is typed and absent and $x_i = -$ if the marker could not be reliably typed. Such unknowns are relatively rare in RH mapping but are much more frequent in genetic mapping or in multiple population/panel consensus mapping.

The probabilistic HMM model for generating each sequence x in the case of error-free haploid "equal retention" data is defined by one retention probability denoted r (probability for a fragment to be retained) and $n-1$ breakage probability denoted $b_1, \ldots, b_{n-1}$ (b_i is the probability of breakage between marker M_i and M_{i+1}). Breakage and retention are considered independent processes.

The structure of the HMM model for 4 markers ordered as `M1`,...,`M4` is sketched as a weighted digraph $G = (V, E)$ in figure 1. Vertices correspond to the possible state of being respectively retained, missing or broken. An edge $(a, b) \in E$ that connects two vertices a and b is simply weighted by the conditional probability of reaching the state b from the state a, noted $p(a, b)$. For example, if we assume that M_1 is on a retained fragment, there is a probability b_1 that a new fragment will start between M_1 and M_2 and a probability $(1-b_1)$ that the fragment remains unbroken. In the first case, the new fragment may either be retained (r) or not $(1-r)$. In the second case, we know that M_2 is on the same retained fragment.

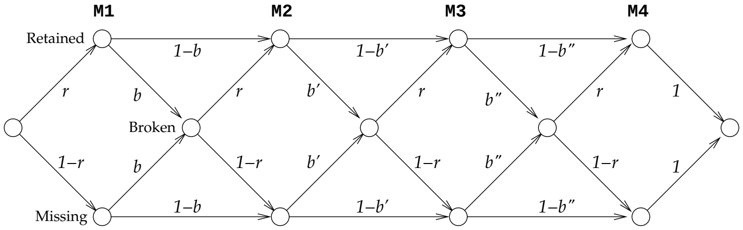

Fig. 1. A graph representation of the "equal retention model" for RH mapping

The EM algorithm [5] is the usual choice for parameter estimation in hidden Markov models [11] where it is also known as the Forward/Backward or Baum-Welsh algorithm. This algorithm can be used to estimate the parameters $r, b_1, b_2 \ldots$ and to evaluate the likelihood of a map given the available evidence (a vector x of observation for each hybrid in the panel).

If we consider one observation $x = (0, -, 0, 1)$ on a given hybrid, the graph can be restricted to the partial graph of figure 2 by removing vertices and edges which are incompatible with the observation (dotted in the figure). Every path in this graph corresponds to a possible reality. The path in bold corresponds to the fact that a fragment has been a breakage between each pair of markers, each fragment being successively missing, retained, missing and retained. If we define the probability of such a source-sink path as the product of all the edges's probabilities, then the sum of the probabilities of all the paths that are compatible with the observation is precisely its likelihood.

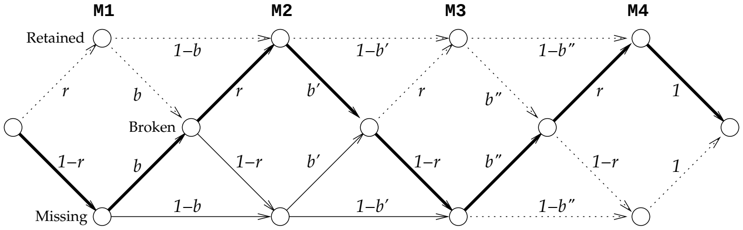

Fig. 2. The graph representation for $x = (0, -, 0, 1)$

Although there is a worst-case exponential number of such paths, dynamic programming, embodied in the so-called "Forward" algorithm [11] may be used to compute the likelihood of a single hybrid in $\theta(n)$ time and space. For any vertex v, if we note $P_l(v)$ the sum of the probabilities of all the paths that exist in the graph from the source to the vertex v, we have the following recurrence equation:

$$P_l(v) = \sum_{u \text{ s.t. } (u,v)\in E} P_l(u).p(u,v)$$

This simply says that in order to reach v from the source, we must first reach a vertex u that is directly connected to v (with probability $P_l(u)$) then go to v (with probability $p(u,v)$). We can sum up all these probabilities that correspond to an exhaustive list of distinct cases. This recurrence can simply be used by initializing the probability P_l of the source vertex to 1.0 and applying the equation "Forward" (from left to right, using a topological ordering of the vertices). Obviously, P_l for the sink vertex is nothing but the likelihood of the hybrid. One should note that the same idea can be exploited to compute for each vertex $P_r(v)$, the sum of the probabilities of all paths that connect v to the sink (simply reverse all edges and apply the "forward" version).

The EM algorithm is an iterative algorithm that starts from given initial values of the parameters $r_0, b_0, b'_0 \dots$ and that goes repeatedly through two phases:

1. Expectation: for each hybrid $h \in H$, given the current value of the parameters, the probability $P_h(u,v)$ that a path compatible with the observation x for hybrid h uses edge (u,v) can simply be computed by:

$$P_h(u,v) = P_l(u).p(u,v).P_r(v)$$

If for a given parameter p, and given hybrid h, we note $S_h^+(p)$ the set of all edges weighted by p and $S_h^-(p)$ the set of all edges weighted by $1-p$, an expected number of occurrence of the corresponding event can be computed by:

$$E(p) = \sum_{h \in H} \frac{\sum_{(u,v) \in S_h^+(p)} P_h(u,v)}{\sum_{(u,v) \in S_h^+(p) \cup S_h^-(p)} P_h(u,v)}$$

2. Maximization: the value of each parameter p is updated by maximum likelihood under the assumption of complete data by:

$$p_{i+1} = \frac{E(p)}{k}$$

It is known that EM will produce estimates of increasing likelihood till a local maximum is reached. The usual choice is to stop iteration when the increase of log-likelihood is lower than a given tolerance. Several iterations are usually needed to reach eg. tolerance 10^{-4}, especially as the number of unknowns increases.

Each of the forward, backward and $E(p)$ computation of the E phase successively treat each pair of adjacent markers in constant time (there are at most 6 edges between each pair of markers). This will be called “steps” in the sequel with the aim of getting a better idea of complexity than Landau’s notation can offer (and which can anyway be derived from the number of steps needed since each step is constant time). From the previous simplified presentation of EM, we can observe that each EM iteration needs $3(n+1)k$ steps since $n+1$ steps are required for the Forward phase, the Backward phase and the $E(p)$ computation phase. The M phase is in $\theta(n)$ only.

2.1 Speeding up the E phase

To make the E phase more efficient, the idea is to try to sum up the data for all hybrids in a more concise way in order to try to avoid, as far as possible the k factor in the complexity of the E phase. The crucial property that is exploited is that when a loci status is known (`0` or `1`), then the probabilities P_l and P_r for the corresponding “Retained” vertex are both equal to 0.0 and 1.0 respectively (1.0 and 0.0 respectively for the “Missing” vertex) and this independently of the markers around.

Any given hybrid can therefore be decomposed in segments of successive loci of 3 different types as illustrated in figure 3:

- **dangling** segments are either segments that start at the first loci and are all of unknown status except the rightmost one (dangling left) or segments that end at the last loci and are all of unknown status except the leftmost one (dangling right).
- **known pairs** are segments composed from a pair of adjacent loci which are both of known status.

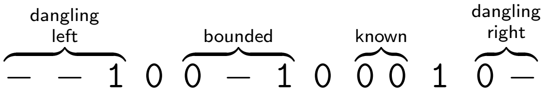

Fig. 3. Decomposing hybrid into segments

– **bounded** segments are segments that start and stop at loci of known status but which are separated by loci of unknown status.

Given the hybrid data set H, it is possible to precompute for all pairs of markers the number of known pairs of each type. This is done only once, when the data set is loaded. For a given loci ordering, we can precompute in a $O(n.k)$ phase the number of dangling and bounded segments of each type that occurs in the data set. Then the EM algorithm may iterate and perform the E phase by computing expectations for each of the cases and multiplying the results by the number of occurrences. Maximizing remains unchanged.

For known pairs, the expectation computation can be done in one step and there are at most $4(n-1)$ different types of pairs which means $4(n-1)$ steps are needed. For all other segments, dangling or bounded, the expectation computation needs a number of steps equal to the length of the fragment. So, if we note u the total number of unknowns in the data set, the total length of these fragments is less than $3u$ and the expectation computation can be done in at most $9u$ steps. We get an overall number of steps of $4(n-1)+9u$ which is usually very small compared to the $3(n+1)k$ needed before. From an asymptotic point of view, this is still in $O(nk)$ because u is in $O(nk)$ but it does improve things a lot in practice. Also, decomposing hybrids into fragments guarantees that repeated patterns occurring in two or more hybrids are only processed once.

There is a specific important case where asymptotic complexity may improve. When multiple population/panels consensus mapping is performed, a marker which is not typed in a data set will be unknown for all the hybrids/individuals of the data-set. This may induce dangling and bounded segments that are shared by all the hybrids/individuals of the data-set but that will be processed only once at each EM iteration. In this case, known pairs and all dangling/bounded segments induced by data-set merging will be handle in $O(n)$ time instead of $O(nk)$.

3 The CarHTaGène package

This improved EM algorithms have been implemented in the CarHTaGène package [13]. Beyond RH and backcross data, CarHTaGène can also handle genetic intercross, recombinant inbred lines and phase known outbred data. Although perfectly able to handle single population mapping, CarHTaGène is oriented towards multiple population mapping: data sets can be hierarchically "merged" and the user decides which data sets share (or not) distance estimations. Markers ordering are always evaluated using a true maximum likelihood multipoint criteria. Beyond multiple population genetic mapping or multiple panels RH mapping, Car$\overset{\mathrm{T}}{\mathrm{H}}$aGène allows to perform mixed genetic/RH mapping by merging genetic and RH data and estimating genetic/RH distances separately using a common loci order.

Considering loci ordering, Car$\overset{\mathrm{T}}{\mathrm{H}}$aGène offers a large spectra of tools for building and validating both comprehensive and framework maps. Most of these tools have been derived from the famous Traveling Salesman Problem (TSP) solving technology. From its beginning [13], Car$\overset{\mathrm{T}}{\mathrm{H}}$aGène has exploited the analogy between backcross genetic mapping and the TSP. As it has been observed in [1], this analogy also exists for RH mapping. Car$\overset{\mathrm{T}}{\mathrm{H}}$aGène has been implemented in C++ and has a Tcl/Tk based graphical interface. It is available on the web at www-bia.inra.fr/T/CarthaGene. It runs on most Unix platforms and Windows.

4 Empirical evaluation of the boosted EM algorithm

To better evaluate the gains obtained by the improved EM algorithm, we have compared it to the EM implementations available in the RHMAP RH mapping package [3] and in the MapMaker genetic mapping package [7]. Since we only wanted to compare EM efficiency and not loci ordering efficiency, all tests have been done by estimating the log-likelihood of a fixed loci ordering using the same convergence tolerance, the same starting point, on a Pentium II 400Mhz machine running Linux, using GNU compilers (respectively g77, gcc and g++ for RHMAP, MapMaker and Car$\overset{\mathrm{T}}{\mathrm{H}}$aGène), on 1000 EM calls.

Simulated data sets have been generated using the underlying probabilistic model both for backcross and RH data:

- the first tests have been done single panel/population data. The RH data uses 50 markers evenly distributed on a 10 ray long chromosome, 100 hybrids and 5% of unknowns. The genetic data, uses 50 markers evenly distributed on a 1 Morgan long chromosome., 200 individuals and 25% of unknowns. This corresponds to typical framework mapping situations.
- the second tests have been done by merging two panels/population. The data sets have been generated using the same parameters except that each data set shares half of the markers with the other data set. When a marker is not typed in a panel/population, the corresponding hybrid/individual is marked as unknown[1]

[1] Note that for RH data, this situation corresponds to the merging of two panels that have been irradiated using a similar level of radiation. If this is not the case, one should rather perform separate distance estimations per panel. More complex models, using proportional distances, are available in RHMAP but are not compatible with the use of the EM algorithm.

Software	Data	Panels/	CPU time	Improvement
package	type	populations	(1000 EM)	ratio
	RH	1		
RHMAP 3.0			110”64	
CarTHaGène			5”57	19.8
	RH	2		
RHMAP 3.0			2376”48	
CarTHaGène			125”95	18.9
	BC	1		
MapMaker			60”99	
CarTHaGène			8”46	7.2
	BC	2		
MapMaker			232”27	
CarTHaGène			70”94	3.3

For radiated hybrid data, the speed-up exceeds one order of magnitude. More modest improvements are reached on genetic data. These improvements may reduce day of computation time to hours and enable the use of more sophisticated ordering techniques, without any approximation being made. These numbers still leave room for improvements since CarTHaGène does not exploit the strategy of precomputing known pairs once and for all but recomputes them at each EM call.

5 Application to real genetic and radiation hybrid data

The improvements obtained apply only to haploid RH data and phase known backcross genetic data. This could be considered as quite restrictive. In this section we show how RH polyploid data and complex genetic data can be reduced to these cases and evaluate the impact of these reductions. This also illustrates how genetic and RH data can be mixed in order to exploit the different resolutions for marker ordering.

Considering genetic data, the USDA porcine reference pedigree consists in a two generations backcross population of 10 parents (2 males from a White composite line and 8 F1 females) and 94 progeny [12]. The height F1 females were obtained after mating White composite females with Duroc, Fengjing, Meishan or Minzhu boars. To reduce this to backcross data, phases have been set to the most probable phase, identified using Chrompic from the CRIMAP package. Then the data is encoded as two backcross (one for the paternal allele and one for the maternal allele). The original data set can be processed by CRIMAP (but not by MapMaker).

Considering RH data, the IMpRH panel consists in 118 radiated hybrid clones produced after irradiation of porcine cells at 7000 rads [14]. Using this panel a first generation whole genome radiation map has been established using 757 markers [6] including 699 markers already mapped on the genetic map from [12].

These two data-sets have been merged and a framework consensus order built for chromosome 1 using the `buildfw` command of CarthaGène. The resulting order contains 38 markers. This order has been validated using simulated annealing and other usual techniques (local permutations, swapping of markers…) and could not be improved, the second best order having a log-likelihood 4.48 below the best. The simulated annealing step alone took few hours on Pentium-II 400Mhz and could probably not have been done in reasonable time using standard EM algorithms.

Using CarthaGène instead of CRIMAP/RHMAP, we made the assumptions that using an haploid model on diploid data, fixing the phase and considering outbreds as double backcross does not change *differences* in likelihood too much. In order to check these assumptions, we compared CarthaGène to CRIMAP and RHMAP applied separately on the 4 best orders identified by CarthaGène. We used CarthaGène haploid model, RHMAP haploid model and RHMAP diploid model on the RH data alone. The following table indicates the log-likelihoods obtained in each case and the differences in log-likelihood with the best order. The results are consistent with our assumption.

	Order 1	Order 2	Order 3	Order 4
CarthaGène	-927.61	-928.37	-932.65	-931.21
RHMAP haplo.	-927.61	-928.37	-932.65	-931.21
RHMAP diplo.	-926.45	-927.26	-931.36	-930.07
Dif. haplo.	0.00	-0.76	-5.04	-3.60
Dif. diplo.	0.00	-0.81	-4.91	-3.62

The same comparison was done using CRIMAP outbred model and CarthaGène backcross model on the derived double backcross data. The results obtained are again consistent with our assumption. Note that the important change in log-likelihood is not surprising: fixing phases brings in information, while double backcross projection removes some information. The important thing is that differences in log-likelihood are not affected.

	Order 1	Order 2	Order 3	Order 4
CRIMAP	-366.54	-370.26	-366.54	-379.22
CarthaGène	-422.29	-426.01	-422.29	-434.97
Dif. CRIMAP	0.00	-3.72	0.00	-12.68
Dif. Cartha.	0.00	-3.72	0.00	-12.68

We completed this test by a larger comparison, using more orders, and it appears that the differences in log-likelihood are well conserved: a difference of difference greater than 1.0 was observed only for orders whose log-likelihood was very far from the best one (more than 10 LOD). CarthaGène can thus be used to build framework and comprehensive maps, integrating genetic and RH maps in reasonable time. For better final distances between markers, one can simply reestimate them with RHMAP diploid model and CRIMAP.

6 Conclusion

We introduced a simple improvement for the EM algorithm in the framework of HMM based RH and genetic maximum likelihood estimation. One of the limitation of this improvement is that it can only deal with observations that either completely determine the hidden states or leave the hidden state undetermined (unknown). It can therefore not be extended to eg. handle intercross genetic data, diploid RH data or to models that explicitly represent typing errors [9]. However, as we experienced on real data, these restrictions can be easily be dealt with.

This is especially attractive considering that our application to haploid radiation hybrid and backcross genetic mapping shows that the boosted EM algorithms can lead to speed-ups of more than one order of magnitude compared to the traditional EM approach, without any loss in accuracy. Ordering techniques such as simulated annealing or taboo search requires an important number of calls to the evaluation function. The efficiency of the boosted EM algorithm makes the application of such approaches practical, even in the presence of unknowns. This is crucial for multiple population/panel mapping as it has been done eg. in [8].

Acknowledgements

We would like to thank Gary Rohrer (USDA) for letting us use the USDA porcine reference genetic data and Martine Yerle (INRA) for the RH porcine data.

References

[1] Amir Ben-Dor and Benny Chor. On constructing radiation hybrid maps. *J. Comp. Biol.*, 4:517–533, 1997.

[2] Amir Ben-DOr, Benny Chor, and Dan Pelleg. RHO – radiation hybrid ordering. *Genome Research*, 10:365–378, 2000.

[3] Michael Boehnke, Kathryn Lunetta, Elisabeth Hauser, Kenneth Lange, Justine Uro, and Jill VanderStoep. *RHMAP: Statistical Package for Multipoint Radiation Hybrid Mapping*, 3.0 edition, September 1996.

[4] D.R. Cox, M. Burmeister, E.R. Price, S. Kim, and R.M.Myers. Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, 250:245–250, 1990.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser.*, 39:1–38, 1977.

[6] R.J. Hawken, J. Murtaugh, G.H. Flickinger, M. Yerle, A. Robic, D. Milan, J. Gellin, C.W. Beattie, L.B. Schook, and L.J. Alexander. A first-generation porcine whole-genome radiation hybrid map. *Mamm. Genome*, 10:824–830, 1999.

[7] E.S. Lander, P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1:174–181, 1987.

[8] E. Laurent, V.and Wajnberg, B. Mangin, T. Schiex, C. Gaspin, and F. Vanlerberghe-Masutti. A composite genetic map of the parasitoid wasp *Trichogramma brassicae* based on RAPD markers. *Genetics*, 150(1):275–282, 1998.

[9] Kathryn L. Lunetta, Michael Boehnke, Kenneth Lange, and David R. Cox. Experimental design and error detection for polyploid radiation hybrid mapping. *Genome Research*, 5:151–163, 1995.

[10] Jurg Ott. *Analysis of human genetic linkage.* John Hopkins University Press, Baltimore, Maryland, 2nd edition, 1991.

[11] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

[12] G.A. Rohrer, L.J. Alexander, J.W. Keele, T.P. Smith, and C.W. Beattie. A microsatellite linkage map of the porcine genome. *Genetics*, 136:231–245, 1994.

[13] T. Schiex and C. Gaspin. Cartagene: Constructing and joining maximum likelihood genetic maps. In *Proceedings of the fifth international conference on Intelligent Systems for Molecular Biology*, Porto Carras, Halkidiki, Greece, 1997. Software available at www-bia.inra.fr/T/CartaGene.

[14] M. Yerle, P. Pinton, A. Robic, A. Alfonso, Y. Palvadeau, C. Delcros, R. Hawken, L. Alexander, C. Beattie, L. Schook, D. Milan, and J. Gellin. Construction of a whole-genome radiation hybrid panel for high-resolution gene mapping in pigs. *Cytogenet. Cell Genet.*, 82:182–188, 1998.