Detection of Recombination in Variable Number Tandem Repeat Sequences

Ezekiel Adebiyi, Eric Rivals

Detection of Recombination in Variable Number Tandem Repeat Sequences

Ezekiel Adebiyi, Eric Rivals

Department of Computer and Information Sciences, Covenant University, PMB 1023, Ota, Nigeria
L.I.R.M.M., Université de Montpellier II, CNRS UMR 5506, 161 rue Ada, F-34392 Montpellier Cedex 5, France

ABSTRACT

Tandem repeats are repeated sequences whose copies are adjacent along the chromosomes. They account for large portion of eukaryotic genomes and are found in all types of living organisms. Among tandem repeats, those with repeat unit of middle size are called minisatellites. These loci depart from classical loci because of the propensity to vary in size due to the addition or the removal of one or more repeat units. Due to this polymorphism, they prove useful in genetic mapping, in population genetics, and forensic medicine. Moreover, some specific tandem repeat loci are involved in diseases, like the insulin minisatellite, which is implicated in type I diabetes and obesity. Those loci also undergo complex recombination events. Presently, some programs to compare tandem repeats alleles exist and yield good results when recombination is absent, but none correctly handles recombinant alleles. Our goal is to develop an adequate tool for the detection of recombinant among a set of minisatellite sequences. By combining a multiple alignment tool and a method based on phylogenetic profiling, we design a first solution, called *MS_PhylPro*, for this task. The method has been implemented, tested on real data sets from the insulin minisatellite, and proven to detect recombinant alleles.

KEYWORDS: VNTR, tandem repeats, genetics, minisatellite, insulin, INS, recombinant, cross-over, phylogenetic profile, multiple alignment

1 INTRODUCTION

The genome length in base pairs (bps) displays huge variations among species: from about 10^5 bps for an archebacteria to 3.10^9 for humans, or even to more than 10^{11} bps for the protozoa *Amoeba dubia*. These differences are partly explained by the presence of regions called repeats that occur many times in genomes. Some molecular mechanisms allow the cell to duplicate a genome region. Among different classes of repeats, those whose copies are located one next to the other on the chromosome are termed *tandem repeats*. Particularly in tandem repeats, duplication and its dual event, contraction, may occur at very high frequencies, letting these loci acquire or lose one or more repeat units. These mechanisms make them the most variable regions (polymorphic loci) of the human genome. Among variable tandem repeat sequences (VNTR), one finds *minisatellites*, *i.e.*, tandem repeats whose repeat unit ranges between 7 and 100 bps. The length variability of minisatellites made them markers of choice in genetic mapping, in forensic medicine for individual identification or paternity testing, and to study genome variation inside or across populations (population genetics) [1]. Indeed, the repeat sequences (*alleles*) observed at a given locus in two individuals may be different. Over time, the repeat copies also undergo point mutations (substitutions, insertions, and deletions), which let them differ from each other. The sequence of the variants of the repeated unit of a minisatellite can be charted by a technique named *Minisatellite Variant Repeat PCR* [2], the result of which is a sequence called *map*, over the alphabet of variants (not the DNA alphabet). From the medical view-point, a lot of interest has been devoted to minisatellite since the discovery that some loci are involved in disease development [3]. For instance, the insulin minisatellite (INS) has been proven to be an important genetic factor in polycystic ovary syndrome, obesity, and type I diabetes [4, 5]. Like in hypervariable minisatellites and in many human minisatellites, the evolution of the INS locus involves recombination, which leads to exchanges of groups of variants between alleles [6], as illustrated by Figure 1.

Recently, progress has been made towards computational analysis of minisatellite sequence data. The main need is to compare the maps that represents the alleles of two individuals. The comparison should measure the differences between the two maps by accounting for the number of mutations needed to transform one into the other. Solutions to the minisatellite map alignment problem have been proposed for the case where mutations include point mutations, duplications, and contractions [7, 8, 9]. Basically these works extend dynamic programming approaches for the alignment of classical sequences (which do not undergo duplications nor contractions) to account for long range dependencies in the maps. The pairwise alignment program of [9] has been used to construct a multiple alignment program called *MS_Alimul* [10]. *MS_Alimul* starts with pairwise alignments, and grows them by adding new maps that are the closest to the

Email: Ezekiel Adebiyi e.adebiyi@dkfz.de, Eric Rivals rivals@lirmm.fr

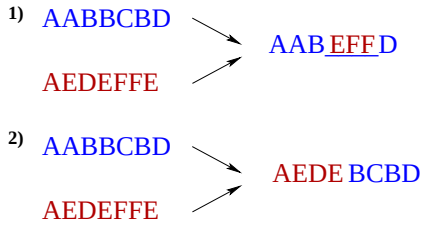

Figure 1: Two examples of recombination between two minisatellite alleles, which give rise to a recombinant allele. All alleles are given as maps, i.e., *sequences of symbols taken from the alphabet of the variants of the repeat unit (here* $\{A,B,C,D,E,F\}$*). On the left the two progenitor alleles (top one in blue, bottom one in red), on the right the recombinant created. In the recombinant allele of example* 1*, the variant symbols from positions* 4 *to* 6 *comes from the red progenitor, while the segments before and after come from the blue one. In this recombinant, there are two cross-over points; the first lies between positions* 3 *and* 4*, while the second is between* 6 *and* 7*. In other cases, the cross-over point may lie outside the repeat and the recombinant can be split in two segments (instead of three), one originating from each progenitor. This case is depicted in Example* 2*, the prefix comes from the red progenitor and the suffix from the blue one.*

maps already in the alignment. The choice of the maps is performed following a *guide tree* obtained by hierarchical clustering of the maps according to their pairwise alignment distances (all pairwise alignments are first precomputed). Alignments are progressively "merged" using the procedure for aligning alignments from [11]. This step is performed iteratively until one obtains a complete multiple alignment. This strategy is known as *progressive alignment* in the literature [12].

Both pairwise and multiple alignments make sense from the biological point of view if the maps under scrutiny are homologous, that is derive from a common ancestor. The assumption is that mutations represent a small portion of the maps and it is possible to find a putative series of mutations to transform one map in the other. Recombination is able to create a new allele, the *recombinant*, from pieces of two different alleles. These parent alleles may not be homologous or their common ancestor may be so ancient that their maps are now dissimilar. A recombinant allele can thus be aligned by pieces, but not completely, with each of its parents or with any map similar to one of its parents. Aligning pairwise a recombinant with such a map yields an alignment of high similarity in the regions of homology and of low similarity in the other regions. In the case of multiple alignment, it pertubates the construction of the progressive alignment by inserting many gaps in the non-homologous regions.

A long biological literature is devoted to the detection of recombinant sequences in non-repetitive genetic sequences, and the available methods have been reviewed and compared in [13] (see references therein). Proposed approaches either (1) rely on the knowledge of a phylogenetic tree, (2) use the pattern of nucleotidic substitution site-by-site, or (3) compute a distance based criterion over a window of a multiple alignment. In this last class, one finds the *Phylogenetic profile* method of [14] (also developed in RAT [15]).

Up to now, for VNTR in general and for minisatellite data in particular, none of the available alignment methods deals with recombinant alleles. As the availability of a phylogenetic tree of the VNTR alleles is rare, and as for such data, variant duplications are much more frequent than nucleotidic substitutions (and make it more difficult to obtain surely homologous sites in a column of a multiple alignment), approaches of type (1) and (2) are unadapted to the detection of recombination. Thus, we turn ourselves towards a distance based method. Here, we present a first algorithm to detect if an allele is a recombinant. We adapt the phylogenetic profile method [14] to VNTR data and combine it with the multiple alignment procedure to detect putative recombination positions in a map. We have implemented our method in a program dubbed *MS_PhylPro* and tested it on different recombinant and non-recombinant alleles of the INS minisatellite provided in [4].

The sequel of the paper is organised as follows: in Section 2 we summarise the phylogenetic profile approach. In Section 3, we explain *MS_PhylPro* and its procedure for the selection of progenitors, while we present the material and the validation tests on real data in Section 4.2. The strengths and limitations of the approach, as well as some perspectives are discussed in Section 5.

2 DETECTION OF RECOMBINATION USING PHYLOGENETIC PROFILES

Phylogenetic profile is a distance based method for recombination detection [14]. It relies solely on the sequence data and does not require the knowledge of an evolutionary tree (which is rarely available for VNTR sequences). Let us give a brief overview of this method.

Assume one investigates if a candidate sequence t is a recombinant of some sequences from a set S, and assume that the two progenitor sequences are in S. Just after the recombination event, a recombinant sequence having one cross-over point is made of two pieces: the region from the beginning up to the cross-over point comes from one progenitor sequence, the remaining region after this point comes from the second progenitor sequence. Thus, if one measures the sequence similarity on these two regions between t and all sequences in S, the maximum of similarity for the left region is attained with the first progenitor, while for the right region is reached with the second progenitor. If one compares the vectors of pairwise sequence similarities of the two regions, their correlation is minimal at the cross-over point. Thus, the algorithm considers all positions in turn in sequence t, looks at the sequence similarities on a left- and right-windows separated only by the current position, and selects a possible cross-over point where the correlation between the left and right similarity vectors is minimal.

This would work perfectly if time has stopped just after the recombination and if the two progenitors are in the set S. However, recombination occurred in the past and since then the sequences have evolved and accumulated point-mutations (substitutions, insertions, deletions), which blurred the recombination signal. Only descendants of the original progenitors are in set S. Moreover, some sequences in S may themselves be recombinants.

One encounters another problem: whatever the sequence, there is (at least) one position at which the correlation is minial. One must decide with the correlation value if this position is a true cross-over point. Is the correlation value low enough? To solve this problem, an estimation of the significance of the correlation has been added to the method named *PhylPro* [13]. In an empirical procedure, the algorithm generates a user-defined number of randomised alignments by shuffling the columns of the original multiple alignment, and recomputes the minimum correlation value among all positions. The *P*-value for the null hypothesis of recombination absence equals the number of times the simulated correlation value was smaller than the one observed with the original data.

Important parameters of the phylogenetic profile method are the window size, the distance used to measure the sequence similarity, and the correlation coefficient. For nucleotidic sequences, the Hamming distance in combination with the linear correlation coefficient yields good results. In practice, *PhylPro* takes as input a multiple alignment of a set of sequences and tests in turn for each sequence if it is a recombinant sequence. The window size is set as a function of the total number of variable sites in the multiple alignment (*i.e.*, the columns of the alignment that contains at least two different symbols). As output, the method plots the correlation coefficient along the sequence position, but does not automatically predict the progenitor sequences. A visual examination of the plot is needed to predict whether a sequence is a recombinant and predict its putative progenitors [14]. An option of *PhylPro* offers the possibility to remove the alignment sites that are too polymorphic, since these suggest the presence of homoplasy.

3 A NEW ALGORITHM FOR DETECTING RECOMBINATIONS IN TANDEM REPEATS

We propose and develop a program named *MS_PhylPro* to detect recombination events in VNTR data. It follows the principle of *PhylPro* exposed above and combines it with a multiple alignment procedure called *MS_Alimul*, which is specifically designed for the alignment of VNTR maps. Compared to *PhylPro*, we implement an additional feature to select a subset of alleles among the set of putative progenitors present in the input multiple alignment. For this we propose an efficient procedure and describe it in Section 3.2. This improvement is a first reason why we choose to adapt the phylogenetic profile method to VNTR data, instead of fitting the data into existing programs. The second reason is because we find it more practical to have a program that takes as input an alignment of the sole putative progenitors, as such alignments can be obtained from the literature or with *MS_Alimul*.

3.1 Overview of *MS_PhylPro*

In *MS_PhylPro*, we choose to separate the input into (i) a single test allele for which we want to assess if it is a recombinant or not, and (ii) an already computed multiple alignment M_S of a set S of alleles, which may be putative progenitors of the test allele.

As *MS_PhylPro* builds on *MS_Alimul*, we first give a short description of the latter. It is a heuristic progressive multiple alignment procedure that was designed for tandem repeat maps. It follows the algorithmic scheme used in ClustalW to align classical (*i.e.*, non-repetitive) sequences [16]. It first computes an optimal pairwise global alignment for each possible pair of maps using *MS_Align* [9], and then uses the resulting distance matrix between alleles to infer a guide tree for the multiple alignment with a Neighbour-Joining method [17]. This tree resembles an evolutionary tree in that the alleles are at the leaves and internal nodes group alleles according to similarity. To each internal node corresponds the set of alleles at the leaves of the subtree rooted by this node. An example of a guide tree for five alleles is displayed in Figure 2. In a third step, the multiple alignment is built progressively for larger and larger subsets of alleles corresponding to the internal nodes of the guide tree, until reaching the root, whose associated set comprises all alleles. Finally, the complete multiple alignment is optimised with local modifications.

In *MS_PhylPro*, we first align the test allele t with the multiple alignment M_S. This is done with the procedure of *MS_Alimul* that aligns one allele with a multiple alignment of a set of sequences. It is an extension of a pairwise alignment method *MS_Align* that adapts the algorithm of [11]. This yields a new multiple alignment, call it M, for the set $S \cup \{t\}$. Then we apply *PhylPro* on M for the test allele t. The window size W is set automatically as in *PhylPro*. To compute the distance between t and any other allele in S, we use the Hamming distance over the two windows: it is simply the number of columns in which the characters of the two maps differ. However, one can account for the number of nucleotidic differences between each pair of aligned variants provided their nucleotidic sequences are known. This has been implemented in the last version of *MS_Align* [9], on which *MS_Alimul* is built, but not yet in *MS_PhylPro*. Then *PhylPro* selects the possible recombination junction and outputs a *P*-value as detailed above. The algorithm of *MS_PhylPro* is sketched in Figure 3. Note that we conserved the option that enables the removal of polymorphic sites. However, we show experimentally in Section 4.2 that this option impairs the recombination detection and does not seem useful at least in the case of INS data.

3.2 Progenitor selection

As mentioned in Section 2, *PhylPro* solely predicts recombination junction, but not the progenitor alleles. Once a potential recombinant has been detected, a naive solution for the prediction of progenitors is to test the recombination in any triple of alleles including the recombinant. This requires a time at least proportional to the square of the number of sequences, which is not practical.

Here, we propose to take advantage of the guide tree used to construct the multiple alignment M. This tree classifies the alleles in clusters according to their similarity: to each internal node corresponds the set of alleles at the leaves of the subtree rooted by this node. As alleles from the same cluster (*i.e.*, in the same subtree) are similar, it is often not meaningful to test all possible pairs of alleles from the same cluster. In our solution, the user gives as

parameter a depth in the tree. The depth of a node is the number of nodes that are its ancestors on the path from the root to this node. One can consider an internal node at this depth in the tree and draw a horizontal line that cuts the tree. Each edge crossed by the line leads to a subtree, that is to a cluster of related alleles. Thus, to a depth in the tree correspond as many clusters as the number of edges crossed by the horizontal line (see Figure 2). We propose to randomly sample one allele per cluster at the given depth and to test for recombination between t and this subset of alleles. In this way, we ensure that the selected alleles are not the most similar and restrict the number of possibilities.

We implemented this idea in a procedure called, *Selection of Progenitors by Depth* or *SPD* for short. It takes as input a tree T and a integer *depth* representing the given depth (in the range 1 to $\log_2(\#S)$). It outputs a subset S_p of S. The recombinant status of t can be tested against alleles in S_p, if the correlation value is lower than for the whole set and it suggests that alleles resembling the original progenitors are in S_p.

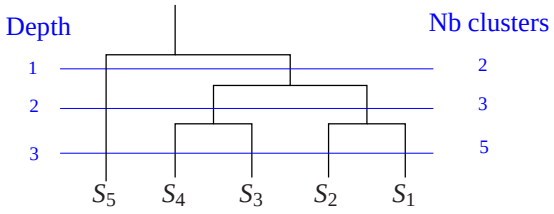

Figure 2: A guide tree for a set of 5 *sequences* S_1, S_2, S_3, S_4, *and* S_5. *Illustration of the notion a depth used in the procedure called* Selection of Progenitors by Depth *(SPD). A depth is represented by a horizontal line; at its right the number of clusters associated to that depth is shown.*

1. *MS_PhylPro*(M_S: multiple alignment, t: test allele, T: guide tree, *depth*: depth parameter)
2. $T \leftarrow$ *MS_Alimul*(M_S, t)
3. $S_P \leftarrow$ *SPD*(T, *depth*)
4. P-value $\leftarrow$ *PhylPro*(S_P, t)
5. Return(Is-Significant(P-value))

Figure 3: The algorithm *MS_PhylPro*.

3.3 Algorithm Complexity

As we account for point mutations and duplications/contractions, the multiple alignment problem considered here generalises the classical multiple alignment problem with a sum-of-pairs score, which is NP-complete [18]. Apart from the multiple alignment step, *PhylPro* takes $O(lnw)$, where l, n, and w denote respectively the number of sequences in the multiple alignment, its number of columns, and the window length. The procedure for progenitor selection takes $O(l)$, if implemented with a Breadth First Search on the guide tree.

4 EXPERIMENTAL EXPERIENCE

4.1 Adaptation to VNTR sequences and the INS minisatellite data set.

Here, we emphasise the differences of recombination detections with VNTR sequence data compared to classical nucleotidic sequences. In the case of minisatellite maps, the symbols of maps represent variants of the repeat unit. The alphabet of symbols is fixed experimentally in the MVR-PCR assay [2]. In most cases the nucleotidic sequence of each variant is known and it is possible to compute a number of differences, a distance between any pair of variants. Thus, by computing difference between maps windows, it is possible to account for the nucleotidic differences between each pair of aligned variants.

For our experiments, we use maps data from the INS minisatellite, which lies upstream from the Insulin gene on Chromosome 11 in the Human genome [4]. The maps were obtained through Minisatellite Variant Repeat-PCR. In Caucasian populations, somatic INS alleles cluster in two main classes, I and III, of respectively small and large alleles [4]. In contrast to somatic alleles, mutant alleles arising in sperm often display complex rearrangement involving intra- and inter-allelic recombination. The INS data is well suited to test recombination detection in VNTR sequences. All INS minisatellite maps were retrieved from `http://www.le.ac.uk/genetics/ajj/insulin`.

First, the data include mutant alleles derived from the sperm of four men. In each case, the progenitor alleles of these men are known. Among these mutants, the authors report in Figure 5 of [4] eleven cases of mutant resulting from complex inter-allelic recombination. This gives a data set of immediate recombinant alleles with the knowledge of their true progenitors alleles. These are positive examples for our tests. However, in real cases one does not know the progenitor alleles, but only other somatic alleles sampled in the population (see below). Thus, we composed our eleven positive test cases with one sperm mutant allele and a subset of somatic alleles from the same class as its progenitors.

Second, the data contain maps of somatic alleles from different individuals living in UK; these alleles were classified into 4 subclasses IC, ID, IIIA, and IIIB. Most alleles in a class are similar to each other and can be correctly aligned with *MS_Alimul* [9]. Investigations of the mutation processes at work in somatic cells at the INS minisatellite show that most variations arise through simple intra-allelic duplication of variants. Combined with the fact that alleles within each class or subclass can be multiply aligned without major gaps in the alignment, this is strong evidence in favor of the absence of recombination within these subclasses. Therefore, we composed from the whole data set ten groups of alleles that either do not contain a recombinant allele or contain one but no alleles related to its progenitor (see Table 5 in the appendix). These last datasets are termed "non-recombinant" data sets and are used to test the specificity of *MS_PhylPro*, *i.e.*, its ability to predict no recombination when there is none.

4.2 Implementation and Experimentations

The above algorithm *MS_PhylPro* has been implemented in C and tested on using a PC with Pentium IV CPU and 512MB RAM. A visual presentation of the algorithmic flow of *MS_PhylPro* with an example output are shown in Figure 4.

Figure 4: Workflow and example output of MS_PhylPro with recombinant allele $S1-46.1$ *and progenitors* $P_1 := IIIA143.1$ *and* $P_2 := ID40.2$.

We first perform experiments in which the input consists a subset of alleles from the four subclasses and the test allele is a recombinant. Table 1 summarises the results with and without the option of removal of highly polymorphic sites. Let us consider that recombination is detected whenever the *P*-value is below 0.01. The results show that without removal, recombination is always detected in these positive test cases. The automatically set window size is given next to the *P*-value. It is interesting to note that the result is positive although both the number of alleles in the set and the size of the segment exchanged by recombination vary respectively from 36 to 113 for the former (see Column 3 of Table 4), and from 4 to 29 repeat units for the latter. The second outcome is that removing polymorphic sites of the alignment reduces the latter drastically (see the window size) and prevents detection. Obviously, this option is not appropriate for such highly polymorphic markers as the INS minisatellite, and finds its application only in less polymorphic DNA sequences. All subsequent experiments are performed without removal.

Second, we run the same type of experiments with the ten non-recombinant data sets. In those cases, *MS_PhylPro* outputs *P*-values ranging from 0.14 to 1.00 with an average of 0.67 (see Table 2). These examples suggest that *MS_PhylPro* specifically detects recombination and not simple allele dissimilarity generated through tandem duplications and point mutations.

In our last experiments, we tested *MS_PhylPro* with and without the procedure for selection of progenitors. Results are listed in Table 3. These illustrate the effect of the procedure $S_P \leftarrow SPD(T,\ d)$,which automatically selects a subset of relatively dissimilar putative progenitors. The smaller the subset the faster is the overall computation. One observes that it divides the running time by at least 2 without sacrificing the sensitivity of *MS_PhylPro* for detecting recombination.

Data set	Remove Polymorphic sites		Don't remove Polymorphic sites	
1	0.944(N)	$W=6$	0.002(Y)	$W=34$
2	0.133(N)	$W=10$	0.000(Y)	$W=114$
3	1.000(N)	$W=2$	0.000(Y)	$W=100$
4	1.000(N)	$W=2$	0.000(Y)	$W=100$
5	1.000(N)	$W=2$	0.000(Y)	$W=100$
6	1.000(N)	$W=2$	0.000(Y)	$W=100$
7	1.000(N)	$W=2$	0.000(Y)	$W=100$
8	1.000(N)	$W=2$	0.000(Y)	$W=100$
9	1.000(N)	$W=2$	0.000(Y)	$W=100$
10	1.000(N)	$W=2$	0.000(Y)	$W=100$
11	1.000(N)	$W=2$	0.000(Y)	$W=100$

Table 1: Tests of MS_PhylPro with eleven positive data sets in which the test allele is a recombinant. Tests are performed with (column 2) and without (column 3) the option for removal of polymorphic sites. In both cases, we give the P-value, the outcome (Y/N), and the window length indicated with W. We consider that recombination was significantly detected when P-value<0.01.

Set #	1	2	3	4	5	6	7	8	9	10
P-value	0.52	0.97	0.25	0.14	0.69	0.83	0.33	0.98	1	1
Outcome	N	N	N	N	N	N	N	N	N	N

Table 2: Results of MS_PhylPro with ten non-recombinant data sets: the data set number on the first line, output P-value and binary outcome on the second and third lines, respectively. Data sets are described in Table 5 of the appendix.

Data	Without *SPD*(*T*,depth)		With *SPD*(*T*,depth)	
set	*P*-value	Run time (s)	*P*-value	Run time (s)
1	0.023(Y)	10.92	0.017(Y)	4.23
2	0.000(Y)	448.72	0.000(Y)	190.11
3	0.000(Y)	76.44	0.000(Y)	34.25
4	0.000(Y)	74.90	0.000(Y)	33.72
5	0.000(Y)	79.33	0.000(Y)	33.79
6	0.000(Y)	81.89	0.000(Y)	34.20
7	0.000(Y)	88.15	0.000(Y)	34.32
8	0.000(Y)	76.85	0.000(Y)	34.22
9	0.000(Y)	75.24	0.000(Y)	33.05
10	0.000(Y)	75.82	0.000(Y)	33.04
11	0.000(Y)	84.80	0.000(Y)	34.28

Table 3: Effect of using or not using the progenitor selection SPD (T,depth) with a fixed window size. Here, polymorphic sites were not removed. Results are given as in previous tables. The main effect is an increase in the computational speed.

5 CONCLUSION AND FUTURE WORK

We provide the first program, *MS_PhylPro*, to detect recombinant sequences among a set of minisatellite data. Our approach combines the multiple alignment program *MS_Alimul* [10] and adapts the phylogenetic profile method of [14] for recombinant detection. Moreover, to predict putative progenitors of a recombinant allele, we propose a method to select a smaller set against which the recombinant status of the test allele is evaluated. It avoids an exhaustive exploration of all allele pairs, accelerates the computations and does not impair correct predictions. Our program was tested on real, positive and negative datasets from the INS minisatellite, and in all cases yielded correct results. This human minisatellite was chosen because cases of recombinant alleles were detected experimentally and are well documented in the article where the relationships between the progenitors and the recombinant are shown [4]. Test cases of this type are rare in the literature. Positive test cases are sperm alleles for which one knows the direct progenitor and not a descendant of this progenitor. This definitely renders the detection of recombination easier.

MS_PhylPro uses a multiple alignment of maps, which we usually compute with the prototype program *MS_Alimul*, which to our knowledge, is the sole computational solution available to date for this task. As mentioned above, one must keep in mind that multiple alignment of sequences is a hard problem (both in the general case as for VNTR alleles), for which only heuristic algorithms are computationally practical. The multiple alignment reported by *MS_Alimul* requires a posterior manual editing to improve the legibility of the alignment, especially when applied to a set of divergent maps.

The mutation rate and turnover processes vary across minisatellite loci and from species to species. Our method yields satisfactory result on a variable human minisatellite, INS, but may reveal unadapted to hypervariable VNTR that undergo complex mutational processes (for instance CEB1 [6] or at least for some haplotypes, MSY1 [19]). However in such cases, simple pairwise alignment is often inappropriate. Fortunately, this type of minisatellite loci are exceptional and do not seem to exist in other mammalian species like mouse [20, 21]. This suggests that *MS_PhylPro* may be useful for a majority of VNTR loci.

Another present limitation is due to the single variant duplications/contractions we consider now. Indeed, in VNTR loci, duplications may copy a complete block of adjacent variants at once. Block duplications are rare compared to single variant events, but are not accounted for in our alignment model [9]. On a large dataset, they do not hinder a correct analysis of the alleles evolutionary relationships as testified in [22]. However, such an event may be erroneously detected as the product of a recombination.

A future line of research is to carry on testing *MS_PhylPro* on a larger number of either simulated or real data sets. We wish to test more thoroughly both the sensitivity and the specificity *MS_PhylPro* on different minisatellite data, like some polymorphic Mouse minisatellites [3] or some hypervariable GC-rich human minisatellites [23]. Because of length variation among alleles, it is sometimes difficult to align globally the recombinant with putative progenitors. A solution may be to use a local alignment algorithm to align the candidate recombinant with the multiple alignment of the putative progenitors. However, such an algorithm must first be designed. Further work includes the detection of multiple recombination events in a single sequence, a precise prediction of the progenitors, and the incorporation of *MS_PhylPro* in a graphical user-friendly interface.

ACKNOWLEDGMENTS

We thank Weiller, Posada and Etherington for useful discussions. We are grateful to Posada for sending us his C implementation of *PhylPro*. Part of this work was done while the first author was at LIRMM, Montpellier, France on a CNRS-NEPAD Bioinformatics Initiative special grant from the CNRS. Eric Rivals is supported by the ACI IMP-Bio REPEVOL (http://www.lirmm.fr/~rivals/REPEVOL), the Languedoc Roussillon Genopole, and a grant from BIOSTIC LR.

REFERENCES

[1] A. J. Jeffreys, V. Wilson and S. L. Thein. "Individual-specific 'fingerprints' of human DNA". *Nature*, vol. 316, no. 6023, pp. 76–79, 1985.

[2] A. J. Jeffreys, A. MacLeod, K. Tamaki, D. L. Neil and D. G. Monckton. "Minisatellite repeat coding as a digital approach to DNA typing". *Nature*, vol. 354, no. 6350, pp. 204–209, 1991.

[3] P. Bois and A. J. Jeffreys. "Minisatellite instability and germline mutation". *Cellular and Molecular Life Sciences*, vol. 55, no. 12, pp. 1636–1648, 1999.

[4] J. D. H. Stead and A. J. Jeffreys. "Allele diversity and germline mutation at the insulin minisatelliete". *Human Molecular Genetics*, vol. 9, no. 5, pp. 713–723, 2000.

[5] J. D. Stead, J. Buard, J. A. Todd and A. J. Jeffreys. "Influence of allele lineage on the role of the insulin minisatellite in susceptibility to type 1 diabetes". *Human Molecular Genetics*, vol. 9, no. 20, pp. 2929–2935, 2000.

[6] J. Buard and G. Vergnaud. "Complex recombination events at the hypermutable minisatellite CEB1 (D2S90)". *EMBO J.*, vol. 13, no. 13, pp. 3203–3210, 1994.

[7] S. Bérard and E. Rivals. "Comparison of minisatellites". *J. of Computational Biology*, vol. 10, no. 3-4, pp. 357–372, 2003.

[8] M. Sammeth, T. Weniger, D. Harmsen and J. Stoye. "Alignment of Tandem Repeats with Excision, Duplication, Substitution and Indels (EDSI)". In R. Casadio and G. Myers (editors), *Proceedings of the 5th Workshop on Algorithms Bioinformatics (WABI-05)*, vol. 3692 of *Lecture Notes in Computer Science*, pp. 426–437. Springer-Verlag, Heidelberg, 2005.

[9] S. Bérard, F. Nicolas, J. Buard, O. Gascuel and E. Rivals. "A Fast and Specific Alignment Method for Minisatellite Maps". *Evolutionary Bioinformatics*, vol. 2, pp. 327–344, 2006.

[10] E. Rivals. "Multiple alignment of minisatellite maps". Unpublished.

[11] J. Kececioglu and W. Zhang. "Aligning alignments". In M. Farach-Colton (editor), *Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching*, no. 1448 in Lecture Notes in Computer Science, pp. 189–208. Springer-Verlag, Berlin, Piscataway, NJ, 1998.

[12] S. Abdeddaim and L. Duret. *Multiple alignments for structural, functional or phylogenetic analyses of homologous sequences*, chap. 3, pp. 51–76. Oxford Univ. Press, 2000.

[13] D. Posada. "Evaluation of methods for detecting recombination from DNA sequences: Empirical data". *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 708–717, 2002.

[14] G. F. Weiller. "Phylogenetic Profiles: A graphical method for detecting genetic recombinations in homologous sequences". *Molecular Biology and Evolution*, vol. 15, no. 3, pp. 326–335, 1998.

[15] G. Etherington, J. Dicks and I. Roberts. "Recombination analysis tool (RAT): a program for the high-throughput detection of recombination". *Bioinformatics*, vol. 21, no. 3, pp. 278–281, 2005.

[16] J. Thompson, D. G. Higgins and T. J. Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[17] R. Desper and O. Gascuel. "Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle". *J. of Computational Biology*, vol. 9, no. 5, pp. 687–705, 2002.

[18] D. Gusfield. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, 1997.

[19] N. Bouzekri, P. G. Taylor, M. F. Hammer and M. A. Jobling. "Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization". *Human Molecular Genetics*, vol. 7, no. 4, pp. 655–9, 1998.

[20] P. Bois, G. Grant and A. Jeffreys. "Minisatellites Show Rare and Simple Intra-allelic Instability in the Mouse Germ Line". *Genomics*, vol. 80, no. 1, pp. 2–4, July 2002.

[21] P. R. J. Bois. "Hypermutable minisatellites, a human affair?" *Genomics*, vol. 81, no. 4, pp. 349–355, April 2003.

[22] F. Bonhomme, E. Rivals, A. Orth, G. Grant, A. Jeffreys and P. Bois. "Species-wide distribution of highly polymorphic minisatellite markers suggests past and present genetic exchanges among House Mouse subspecies". *Genome Biology*, vol. 8, p. R80, 2007. URL `http://genomebiology.com/2007/8/5/R80`.

[23] M. A. Jobling, N. Bouzekri and P. G. Taylor. "Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1)". *Human Molecular Genetics*, vol. 7, no. 4, pp. 643–53, 1998.

A APPENDIX A

Data set	allele t length	Set # alleles	M length
1	46	36	1524
2	46	113	12336
3	66	50	3433
4	58	50	3433
5	56	50	3433
6	56	50	3433
7	54	50	3433
8	51	50	3433
9	48	50	3433
10	47	50	3433
11	46	50	3433

Table 4: Positive data sets contain one recombinant allele (parameter t of *MS_PhylPro*), whose length is given in column 2, and a set of putative progenitor alleles (parameter M_S), the number of which and their cumulated lengths are stated in columns 3 and 4. The recombinant alleles are those mutant alleles detected in sperm and displayed in Figure 5 of [4], while the subsets of progenitor are taken from the subclasses of somatic alleles (see Figure 1 of that reference).

Set #	List of alleles ID	Nb	Size
1	ID39.1,38.2,38.1,s1-51.1	4	166
2	ID43.1,44.1,43.7,43.8,43.6,43.5,41.4, 42.4,43.9,41.2,42.5,42.3,39.3,41.3	14	590
3	ID39.1,38.2,38.1,39.4,40.4,41.5,39.2, 40.2,39.6,39.5,40.3,s2-66.1	12	498
4	ID38.4,40.1,41.1,43.4,44.2,42.1,43.2, 42.2,43.3,38.3,37,1,s2-66.1	12	517
5	IIIB143.3,145.3,145.1,142.1,144.3, 143.4,141.1,143.1,143.2,144.1,145.2,144.2	12	1722
6	IIIA138.1,139.1,138.2,146.3,145.3, 149.14	6	855
7	IIIA144.3,148.2,144.1,143.1	4	579
8	IIIA150.5,150.4,148.10,149.9,149.8, 149.7,147.1	7	1042
9	IIIA156.1,157.1,158.4,157.2,158.2, 159.2,158.3,158.1,159.1	9	1420
10	IIIA150.7,150.1,149.1	3	449

Table 5: Allele composition of the ten non-recombinant data sets. The allele identifiers as in [4] are listed for each set, as well as the number of alleles and their cumulated lengths.