International Journal of Foundations of Computer Science
Vol. 15, No. 2 (2004) 225–257

World Scientific
www.worldscientific.com

A SURVEY ON ALGORITHMIC ASPECTS OF TANDEM REPEATS EVOLUTION

ERIC RIVALS
L.I.R.M.M., CNRS U.M.R. 5506
161 rue Ada, F-34392 Montpellier Cedex 5, France
rivals@lirmm.fr

Received 29 May 2003
Accepted 15 October 2003
Communicated by Juhani Karhumaki

Local repetitions in genomes are called *tandem repeats.* A tandem repeat contains multiple, but slightly different copies of a repeated unit. It changes over time as the copies are altered by mutations, when additional copies are created by *amplification* of an existing copy, or when a copy is removed by *contraction.* Theses changes let tandem repeats evolve dynamically. From this statement follow two problems. Tandem Repeat History aims at recovering the history of amplifications and mutations that produced the tandem repeat sequence given as input. Given the tandem repeat sequences at the same genomic location in two individuals and a cost function for amplifications, contractions, and mutations, the purpose of Tandem Repeat Allele Alignment is to find an alignment of the sequences having minimal cost. We present a survey of these two problems that allow to investigate evolutionary mechanisms at work in tandem repeats.

1. Introduction

A striking genetic difference between species is the size of their genome. Relatively simple organisms, like the protist *Amoeba dubia*, may have much larger genome than *Homo sapiens* for instance. These dramatic differences are due to the presence of repeats. In general, in eukaryotes, organisms whose cells bear a kernel, duplicated genetic material is abundant and can account for up to 60% of the genome. Although some of the mechanisms that generate these repeats are known, from the point of view of evolution, the reasons for such redundancy remain an enigma.

Repeats whose copies are distant in the genome, whether or not located on the same chromosome, are called distant repeats. In this review, we focus on repeats whose copies are adjacent on a chromosome. Because of this characteristic, they bear the name of **tandem repeats**. Among those, biologists distinguish **micro-satellites**, **mini-satellites**, and **satellites**, according to the length of their repeated unit: between 1 and 6 base-pairs, between 7 and 50 base-pairs, and above

50 base-pairs[a], respectively. These names are mainly used for repeats located in regions that do not contain genes. In addition to these sub-classes, numerous groups of similar genes that originate from the same ancestor gene are organized in tandem. They are termed **tandemly repeated genes**.

Local repeats in the DNA arise, grow or disappear through molecular events that copy a contiguous segment on the DNA and insert one or many copies of it next to the original segment, or perform the dual operation. We name these two types of events **amplification** and **contraction**. Like any other segment of the genome, the repeated copies also change through **point mutations**: insertion, deletion or substitution of one base. Point mutations give rise to approximate tandem repeats. The pattern of point mutations along the tandem array of copies informs us on the parent-child relationships between copies. In other words, it gives access to the history of the tandem repeat.

The relatively high frequency of these events let these local repeats evolve rapidly. For a given species and at a precise location on the chromosome, a **locus**, the repeat varies in sequence and/or length in different individuals. Hence, such a locus is said to be **polymorphic** and each different sequence encountered at this locus is called an **allele**.

1.1. *Approximate Tandem Repeats*

In biology, local repetitions in DNA are called "tandem repeats" irrespectively of the number of copies. In computer science, a local repetition is dubbed a **square** if it contains two copies, a **cube** if it contains three, and so on.

An amplification creates a substring that is an **Exact Tandem Repeat**, ETR for short. An ETR is a power of the original pattern: for an integer m, it equals u^m if the pattern is u. When later in the course of evolution point mutations affect this ETR, they let identical positions in adjacent copies differ and the ETR becomes an **Approximate Tandem Repeat**, ATR for short. Note that any sequence is an ATR of some motif. In practice, only repeats whose copies are similar enough receive attention. The level of internal similarity that distinguishes any random sequence from a sequence of true repeats, i.e., that is created by some amplifications, is defined from a statistical view-point (for example in the software TRF[5]) or by an information theoretical measure ([43, 42]). The problem of detecting significant ETR or ATR is an active area of research (see for instance [42, 46, 14, 5, 35, 36, 44]). In the sequel of the paper, by ATR we mean a tandem repeat with sufficient internal similarity. An example of an ATR is given Fig. 1 under the form a multiple alignment of its copies.

Point mutations could cause two adjacent copies to diverge so far that their common ancestry is not recognizable anymore from sequence similarity. In this

[a] Chromosomes are made of a double-stranded Deoxyribonucleic Acid (DNA) helix, whose basic building block is a pair of bases. The unit of a DNA sequence is thus called a **base-pair** and is abbreviated by bp.

c	t	g	a	g	c	t	c	**A**	a	**C**	c	t	t	g	c	t	c	**T**	g	a	g	c	**A**	**T**	c	a	t	c	t	t	-	c	t
c	t	g	a	g	c	t	c	c	a	t	c	t	t	**A**	c	**A**	c	**T**	g	a	g	**A**	**A**	**G**	c	a	**C**	c	t	**G**	-	c	t
G	**C**	**A**	a	g	c	t	c	c	a	t	c	t	t	g	c	t	**T**	**G**	g	a	g	c	t	c	c	**T**	t	c	t	t	g	c	t
c	**C**	**A**	a	g	c	t	c	**T**	a	t	c	-	t	**A**	c	t	c	c	**A**	a	g	c	t	c	c	a	t	c	t	t	g	c	t
c	**A**	g	a	g	c	t	c	c	a	t	c	-	t	g	c	t	c	c	**A**	a	g	c	t	c	c	a	t	c	t	t	g	c	t
c	**G**	**A**	a	g	**T**	**G**	c	c	a	-	**A**	t	**C**	g	c	t	c	c	**A**	a	g	c	**A**	c	**T**	a	t	c	t	t	g	c	t
G	t	g	a	g	c	**A**	**A**	c	a	t	c	-	t	g	c	**A**	**T**	**A**	g	a	**C**	**A**	t	**T**	c	a	t	c	t	t	a	c	t
c	**A**	g	a	g	c	t	c	c	a	t	c	t	**A**	g	-	t	c	**A**	g	a	g	**A**	t	c	c	a	t	c	**C**	**A**	-	c	t

Fig. 1. A multiple alignment of the 8 copies of a tandem repeat found on the human chromosome 22. The lines of the alignment contain the copy in the same order than on the chromosome. Symbols in bold uppercase mark differences between the current copy and a 34 bp consensus motif. On the third column from the right, the copies 3 to 6 all have an extra g character suggesting that they may have arisen through an amplification of arity 4 after the g was inserted in the original copy.

case, it is not a repeat anymore. A major hypothesis is that amplification is favored by the similarity of adjacent patterns, and that when copies have diverged for a long time such former repeat does not undergo amplification anymore. In highly polymorphic loci, like some minisatellites, amplifications and contractions are more probable than point mutations. On the contrary, tandemly repeated genes can accumulate hundreds of mutations and still undergo some amplifications; in this case, amplifications and contractions are less frequent than point mutations.

When one wishes to establish the common ancestry of any two genes, one first searches for sequence similarity. The similarity is quantified through sequence alignment. The Alignment is a weighted version of the Longest Common Subsequence problem and, in the classical setup, considers only point mutations. An exact solution is based on dynamic programming [25, 45]. Dealing with tandem repeat requires to consider also amplifications and contractions. We do not report on other algorithmic and combinatorial problems on local repetitions and refer the reader to numerous textbooks on the subject, among which [39, 13, 25].

1.2. *Interest in Tandem Repeats*

In this section, we summarize theoretical, technical, and medical interests in tandem repeats.

Theoretical Interests.

The abundance of tandem repeats rise some theoretical questions concerning their role in the structure and evolution of the genome. How and why do they appear and evolve? Are they correlated to other local characteristics of the DNA? How frequently do new genes appear through tandem amplification? Already in the 70's, Ohno [40] argued that gene duplication is a major force in the evolution of genomes. For more information on these topics, the reader may refer to textbooks on molecular evolution like [41, 38].

Technical Interests.

Tandem repeats, especially polymorphic micro- and mini-satellites, have proven useful in many areas of molecular biology. Polymorphic markers are used since the beginning of the 90's to construct low resolution genetic maps. A well-known example is the first genetic map of the human genome built with more than 5000 microsatellites markers [12]. These microsatellites also serve in linkage analysis and positional cloning to detect and locate molecular variations causing disorders [37, Chap. 3]. Linkage analysis looks for inheritance correlations between a trait and genetic markers within a pedigree. Polymorphic tandem repeats are markers of choice for Mendelian diseases because the discriminative power of linkage analysis increases with the number of alleles.

In population genetics, polymorphic markers enable biologists to trace the propagation of genetic traits in populations. For instance, highly polymorphic minisatellites allow to confirm the "Out of Africa" hypothesis, i.e., that our species originated in Africa and invaded afterwards the rest of the world [2]. Differences between alleles of highly polymorphic markers, like the minisatellite MSY1 on the human Y chromosome (see Section 3), give us access to recent populations history.

Because of their level of variability, some polymorphic tandem repeats distinguish any two individuals from the same population and enable the technique of DNA fingerprinting [33, 34]. Such markers serve as genetic identifiers in forensic studies for the identification of dead corpse, in paternity testing, and so on [24, 27]. In 1992, the skeletal remains exhumed in 1985 in Brazil were identified, through testing of bone DNA, to be those of Dr Josef Mengele, the Auschwitz 'Angel of Death' [29].

Medical Interests.
At last, tandem repeats are involved in several diseases. Variable minisatellites are known to influence the development of type-1 diabetes, epilepsy and some cancers [6]. Some microsatellites are known to play a role in the regulation of some genes. The most well-known examples are the dozen of severe neurodegenerative diseases caused by large amplifications of CAG/CGG microsatellites either inside or near a gene: fragile X mental retardation, myotonic dystrophy, Huntington's disease, etc (see references in [48, 38, 26]). In healthy individuals, the tandem repeat size varies around a few tens of copies, while in affected individuals the number of copies at the same locus reaches hundreds or a thousand in some cases.

For all these reasons, there are some needs to understand the evolution of tandem repeats. The two problems surveyed in this article should help to establish which mechanisms are responsible for amplification or contraction, to estimate how fast the copies of a repeat change over time, to investigate hypotheses on the disease development or to recover recent evolutionary relationships.

1.3. *Two problems of interest*

Let us first introduce a notation for strings. Let Σ be a finite alphabet of size σ. A sequence of n letters of Σ indexed from 1 to n is called a *word* or a *string* of length n over Σ. We denote the *length* of a word $U := U[1] \ldots U[n]$ by $|U|$. For any $1 \leq i \leq j \leq n$, $U[i,j] := U[i] \ldots U[j]$ is called a *substring* of U. For $U, V \in \Sigma^*$, $U.V$ denotes the concatenation of U and V. For any integer $h > 0$, U^h denotes the h-th power of U, i.e., the concatenation of h times U. We denote by Σ^*, respectively by Σ^n, the set of all finite words, resp. of all words of length n, over Σ. d_L, d_H denote respectively the Levenshtein and the Hamming distance on Σ^*.

Definitions.
We first define the events a sequence can undergo. The classical point mutation events are **substitution** of a symbol by another, **insertion** or **deletion** of a symbol. A synonym of a substitution is a change. When considered together, insertions and deletions are called **indels**.

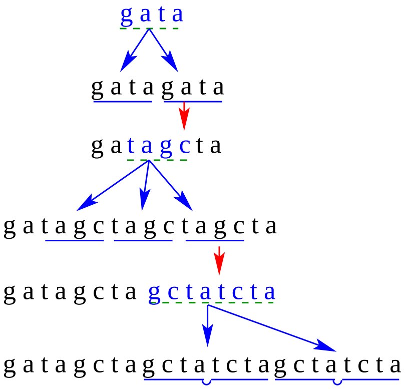
Definition 1 (Hamming and Levenshtein distances). *Let $r, s \in \Sigma^*$ with $n := |r|$ and $m := |s|$. If $m = n$ then the* ***Hamming distance*** *between s and t, denoted $d_H(r,s)$, equals the minimum number of substitutions needed to transform r into s. When m is not necessarily equal to n, the* ***Levenshtein distance*** *between r and s is denoted $d_L(r,s)$ and equals the minimum number of substitutions, insertions, and deletions required to change r into s.*

Examples: Let $r := gata$ and $s := gctt$. $d_H(r,s) = 2$ because changing r into s requires substituting the first a by c and the second a by t. Now let $r := gatag$ and $s := tagcta$. $d_L(r,s) = 4$ because to transform one into the other one needs to mute the first g into t, to insert g and c in s, and to delete the last g of r.

Let k be the pattern length, i.e., the minimum size of substrings that can be copied. We term **amplification** the general event that generates copies in tandem of a substring and **contraction** the dual event. The **order** of an amplification is the number of patterns that are copied at a time and its **arity** is the number of copies produced by the amplification plus one.

Definition 2 (Amplification - Contraction) *Let T be a text over Σ and $k, i, m > 0$ be integers such that $ik \leq |T|$ and $m \geq 2$. An* ***amplification of order*** **i** ***and arity*** **m** *on T replaces a substring u in T of length ik by u^m of length mik. In other words if $u := T[j, j+ik-1]$, it creates a new text $T' := T[1, j-1].T[j, j+ik-1]^m.T[j+ik, |T|]$. We say u is the pattern of the amplification. A* ***contraction of order*** **i** ***and arity*** **m** *on T' is the dual event of the amplification, that is, it replaces u^m by u and yields a new text $T'' := T$.*

Figure 2 gives an example of the history for the ATR sequence *gatagctagctatctagctatcta* with a basic motif size of $k := 4$. The sequence on the last line gives the sequence observed nowadays on the chromosome, while other lines give putative ancestral sequences in order of appearance. The sequence on the first line is assumed to contain a single copy of the motif, i.e., to be of length k, and this copy is the putative ancestral motif of the whole ATR. This implies that



$$\begin{bmatrix} gata \\ gcta \\ gcta \\ tcta \\ gcta \\ tcta \end{bmatrix}$$

Fig. 2. On the left: An example history of an approximate tandem repeat with a basic motif size of $k := 4$. The following events occurred in order: an amplification of order 1 and arity 2 (pattern is *gata*), a substitution of an *a* by a *c*, an amplification of order 1 and arity 3 (pattern is *tagc*), a substitution of a *g* by a *t*, an amplification of order 2 and arity 2 (pattern is *gctatcta*). The patterns and the copies produced are respectively underlined with dashed and straight lines; the number of arrows of an amplification equals its arity. On the right: The multiple alignment of the motifs of length k of the observed sequence. A majority consensus motif is *gcta* although the ancestral motif of the history is *gata*.

the history begins with an amplification of order 1. In this example, the second amplification (*tagc* replaced by $(tagc)^3$) is of order 1 since the pattern copied has length k and of arity 3 since two new copies have been inserted after the original pattern. The last amplification (*gctatcta* replaced by $(gctatcta)^2$) has order 2 since the pattern copied has length $8 = 2k$ and arity 2 because only one copy has been inserted after the original.

When amplifications are of arity 2 only, we use the word **duplication** instead. Point mutations and amplifications are illustrated in Figure 2. It is natural to extend this definition by allowing i and m to be rationals instead of integers. Unfortunately, none of the works reported in literature consider this case. Often the terms duplication, triplication, and m-duplication are used instead of amplification. We choose amplification to avoid ambiguity. Also in [4], contraction is used with another meaning that we will give later on. In the biological literature, long amplifications are also called expansions.

Given these basic definitions, we can informally state our two problems:

Tandem Repeat History : Given the sequence of one approximate tandem repeat and its minimal pattern size, recover its history of amplifications and mutations.

The actual input is given under the form of a multiple alignement with k columns and as many lines as there are copies in the repeat. If the sequence has lengt n, the alignment contain n/k lines. When only substitutions are allowed in the alignment, the alphabet of T is the DNA alphabet and k is the motif length. It is the case of the alignment of Figure 2. When indels are also considered like in the alignment of Figure 1, then the alphabet is the original alphabet Σ plus the symbol $-$ that denotes insertions or deletions in some copies and k is the consensus size of the motif including indels columns. In that case the sequence T is the concatenation of the lines of the alignment and is written over $\Sigma \cup \{-\}$. In both cases, the notions of amplifications and contractions remain the same.

Tandem Repeat Allele Alignment : Given two allele sequences of the same polymorphic tandem repeat locus, compute an optimal alignment between the two sequences considering point mutations, as well as amplification and contraction events.

2. History of a Tandem Repeat

Given the sequence of an approximate tandem repeat and a minimal pattern size, we want to recover the series of events that led to the present sequence. This problem resembles the one of Phylogenetic Reconstruction. For this problem, given a set of sequences of the same gene or protein from different species, one wants to recover the evolutionary tree that led to the apparition of the actual species from an ancestral one by a series of speciations (i.e., division of a species in two). The present sequences are associated with leaves and ancestral species with internal nodes of the tree. In our setup, amplifications replace speciations and can have an arity m greater than two (thereby creating m-ary branching). A natural question arises: can one represent the history of a tandem repeat by a tree? It is in general not the case since subsequent amplifications/contractions can act on sequence segments that are not in phase according to the minimal pattern size (cf. the second amplification in Figure 2). In other words for a given pattern size k, all amplifications may not start at position $1 + jk$ for some integer j. To restrict to histories that can be depicted by a tree, one has to limit the starting positions of amplifications to those that respect the pattern phase. This is the **Fixed Boundary** constraint [4]. Most researchers envisaged the history problem, which was first proposed by Fitch [21], with this restriction, since they consider the case of tandemly repeated genes where it seems to apply.

Logically, researchers harbored their formalizations from the field of phylogeny reconstruction and considered the problem as an optimization problem with two different criteria. The first criterion is the **maximum of parsimony**; one searches for a tree with sequences labeling internal nodes that minimize the number of evolutionary events. The second is **minimum evolution** and gives rise to distance-based approaches which search for a tree that minimizes the distance between leaves, but do not compute ancestral sequences. [21, 4, 47, 19, 31] followed the first line and [47, 16, 17] gave algorithms for the second. Note that when one considers tree-like histories as in phylogeny, the order of events is only partially known. Without the help of a molecular clock, it is impossible to order events that lie in different branches of the tree.

All researches achieved on this subject assume that the minimum pattern size, denoted k, is known and that the copies of the repeat have been aligned in a multiple alignment (see Figure 1 for an example). In the case of the distance based approach, the multiple alignment serves to compute a distance between any pair of copies; the method takes as input the resulting distance matrix. This assumption is real restriction since in a complex history there might be several patterns whose sizes are different and not necessarily multiples of another. The problem of multiply aligning sequences is in general NP-hard and thus, the history reconstruction relies on an approximate solution. Finally, this requirement introduces circularity since the multiple alignment itself and the determination of the pattern size depend on the history of the repeat.

The last restriction imposed by all attempts made so far is that apart from point-mutations only amplifications, but no contractions, occurred in the development of a tandem repeat. The history depicts an always increasing repeat which gains new copies at each amplification. From a biological point of view, it is known that tandem repeats undergo multiple amplifications and contractions, but this restriction allows not to consider infinite sequences of events. However for some loci, the reconstructed history seems robust to deletions in the sense that after the deletion of one copy, it still satisfies the constraint of a tandem repeat history. Robustness is also proven when only amplification of order 1 occurred in the history.

After commenting on the different assumptions, we can state the problem more formally. Unfortunately in the literature, most authors do not tackle with exactly the same problem. Often the formalization is expressed in different ways and each work investigates specific versions of the problem. Here, we introduce an unambiguous terminology and a unifying definition such that each version of the prolem based on the parsimony criterion is an instance of this definition.

From now on, let k denote the minimum pattern size, T be the approximate tandem repeat sequence, and M be the multiple alignment of the copies of T. Let d be a metric on strings over Σ. Figure 1 shows such a multiple alignment with $n := 8$ copies and $k := 34$ columns. We want to find backwards in time the series of events that led to T from a single copy of an unknown pattern. The history is the repetition of the following process: new identical and adjacent copies were added by amplification and then diverged by point mutations. Recovering the history backwards requires to reduce the number of copies to one. We introduce the notion of **reduction**. A reduction rewinds the process described above. Indeed, it chooses adjacent but (in general) not identical copies, rewinds the differences between copies by point mutations, and contracts the now identical copies. To a **reduction of order i and arity m**, denoted $r_{i,m}$, that transforms $u_1 \dots u_m$ into u is associated a cost function of the form: $C(r_{i,m}) := \sum_j d(u_j, u) + \mathcal{A}(i, m)$ where $\mathcal{A}(i, m)$ is a cost function for an amplification of order i and arity m. The choice of an additive function could be discussed, but seems natural since the genetic events do not occur at the same time and are based on different and independent molecular mechanisms. The reconstruction of a history by successive reductions for

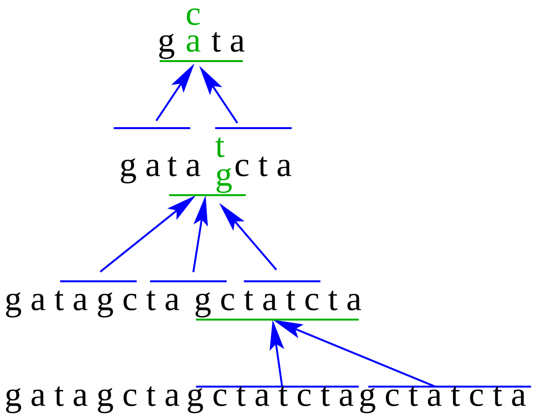

c
g a t a
t
g a t a g c t a
g a t a g c t a g c t a t c t a
g a t a g c t a g c t a t c t a g c t a t c t a

Fig. 3. Reconstruction of a tandem repeat history by successive reductions. The basic motif size is $k := 4$. The history contains three reductions, each combines contractions and mutations if the reduced copies are not identical. The merged copy resulting from the second reduction has a subset of symbols at the first position, while the one of the last reduction has a subset of symbols at the second position. The first ambiguity is resolved in the next reduction and does not appear anymore at the root.

the sequence of Figure 2 is illustrated in Figure 3. In the first reduction from the bottom, the contracted copies have length $2k$ and are identical ($u_1 = u_2$). Thus, they are replaced by the same motif u (i.e., $u = u_1 = u_2$) and the term $\sum_j d(u_j, u)$ equals zero in the reduction cost. In the second reduction, three non identical copies are replaced either by *tagc* or by *tatc*. In the first case, $\sum_j d(u_j, u)$ equals one since two among the reduced copies are *tagc*, while in the second case, only one reduced copy is *tatc* and thus $\sum_j d(u_j, u) = 2$.

Definition 3 (Maximum Parsimony Tandem Repeat History (MP-TRH)). *Let T be a tandem repeat sequence containing n approximate copies of a motif, k be an integer, M be a multiple alignment of k columns and n lines, such that the i-th line contains the i-th copy of T with possibly some indels, and C be a cost function for reductions. The Maximum Parsimony Tandem Repeat History problem is to find the minimum cost series of reductions that convert M into a single copy of length k. (The multiple alignment is reduced until a single line is left.)*

This series of reductions gives a putative reverse history for T and an associated ancestral pattern. k is also called the repeat unit length and the size of the problem is kn. We denote the n copies of the tandem repeat $s_1, \ldots, s_n$.

Variations Around a Theme.
Some works [4, 31] consider a multiple alignment with insertions and deletions. This means that in each line k is the number of columns but not the copy's length. Other authors consider that copies differ only by mismatches/substitutions; in their case,

k is the number of columns in the alignment as well as the copy's length. In the first and second cases respectively, the authors use the Hamming and the Levenshtein distances on Σ^*. When one considers the Hamming distance, one usually discards all columns of the alignment that contain indels. In the case of divergent genes, the number of columns left, k, may be much less than the lengths of the genes that were amplified. In that case, k is not the biological pattern size.

Even if Tang and coworkers [47] include a general cost function in their formalization, all algorithms published so far including theirs only account for differences between copies, but not for amplifications (i.e., in the reduction cost, the function $\mathcal{A}(i, m)$ is always equal to 0).

A major constraint is the **Fixed Boundary** constraint which specifies that amplifications start only in the first column of the alignment. From the computational view-point, it constrains strongly the problem, since many alternative contractions starting at different positions in T, i.e., columns in M, would yield the same result. Also, when boundaries are not fixed, the history cannot be represented by a tree; one needs a more complex structure. From the biological point of view, many tandem repeats do have not an integer, but rather a truly rational number of copies, showing that boundaries of amplifications vary. Nevertheless, when considering tandemly repeated genes, the currently accepted biological model enforces fixed boundaries for amplifications.

Definition 4 (Fixed Boundary Maximum Parsimony Tandem Repeat History (FBMP-TRH)). *This problem is identical to FBMP-TRH except that reductions start at a position* $1 + jk$ *for some* $0 \leq j \leq n - 2$.

In each section, we will precise exactly which version of the problem is examined. Under the criterion of Minimum Evolution, the problem is called Minimum Evolution Tandem Repeat History and is defined in Section 2.6, where two greedy and one exact algorithm are presented. Its complexity class is unknown. In summary, in the Maximum Parsimony case we face a NP-hard problem for which a 2-approximation, a Polynomial Time Approximation Scheme (PTAS), and several greedy algorithms have been described.

2.1. *Benson and Dong's Approximation and Greedy Algorithms.*

[4] is the only reference where the general MP-TRH problem with variable boundaries is investigated; the paper also deals with the restricted version with fixed boundaries and amplifications of arity 2 (these are termed "binary duplications" in the paper). Their cost function accounts for the Levenshtein distance between copies and charges no cost for amplifications; that is $d := d_L$ and for any order i and arity m, $\mathcal{A}(i, m) := 0$. The authors provide a greedy algorithm for the general problem, a 2-approximation, and two ways of computing lower bounds. They also report that the greedy algorithm performs in general better than the approximation

and is close to the lower bounds. In their paper, the word "contraction" means a reduction.

We describe the greedy algorithm and the 2-approximation which is based on an ordered spanning tree built on the tandem repeat copies. Before that, we give a little correction of the cost function and of the reduction rule[b].

When a reduction of arity m is applied to the multiple alignment M, a merged copy replaces the m reduced copies. At a column j, differences between the contracted copies may suggest several characters for j-th position of the merged copy. In a reduction step, Benson and Dong authorize such a position to store a set of putative ancestral symbols. To compute optimal ancestral characters, one needs to count for each possible symbol its number of occurrences in the j-th column of the contracted copies and to store the set of all symbols having the majority at the j-th position of the merged copy. If one then chooses any symbol having the majority, say x, for the j-th position of the merged copy, the number of point mutations that must be accounted for is given by: $m - \mathit{count}[x]$ where the vector *count* stores the counts mentioned above. Indeed, each contracted copy whose j-th character differs from x requires a single point mutation. The cost function in [4] does not consider the relative counts of the characters at a given position; this leads to an incorrect number of mutations, i.e., an incorrect cost.

2.1.1. *A Greedy Algorithm for MP-TRH.*

The greedy algorithm iteratively applies the reduction with the lowest cost ratio; it prefers the reduction with highest arity and breaks other ties arbitrarily. The cost ratio of a reduction is defined as: $\frac{C(r_{i,m})}{k(m-1)}$, i.e., the reduction cost over the arity minus one times the unit length k.

When the arity of reductions is restricted to 2, the greedy algorithm takes $O(kn^3)$ time and when any arity is allowed it takes $O(kn^3 \log(n))$. Let us first explain the complexity for the restricted case. The cost for each possible reduction is computed for increasing order between 1 and $\lfloor \frac{n}{2} \rfloor$. For order i, there are reductions starting at all columns on line j such that $1 \leq j \leq (n - 2ik)$ and at column one on line $(n - 2ik) + 1$. In the left to right order, the cost of reductions whose substrings overlap by all but the first and last characters can be deduced in $O(1)$. Thus, all costs can be obtained in $O(kn)$ time. For each reduction step, the computation takes $O(kn^2)$ and as there are at most $n - 1$ steps the complexity in $O(kn^3)$. In the case of unrestricted arity, the algorithm is more complex since for a given order i and a given position, the costs for arity m can be deduced from those of arity $m - 1$ in $O(1)$. At most n/m costs are updated; the complexity for each step and the total complexity become $O(kn^2 \log(n))$ and $O(kn^3 \log(n))$, respectively.

[b] G. Benson told us he also noticed this error and corrected it in his algorithms.

H	1	2	3	4
1	0	1	0	0
2	0	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0

Fig. 4. Table H for the sequence *gatagctagctatctagctatcta* with $k := 4$ and $n := 6$.

Example : Computation of the reduction costs for reduction of order 1 and arity 2. The greedy algorithm computes a table H with k columns and $n-1$ lines, such that for a position j of T, the entry $H[j/k+1, j, k]$ equals 0 if the character at position j equals the one k position further (i.e., if $T[j] = T[j+k]$), and 1 otherwise. The table H for the sequence of Figure 2 is given in Figure 4. The cost $C(j)$ of the reduction starting at position j in T is given by $C(j) := \sum_{l=0}^{k-1} H[(j+l)/k+1, (j+l), k]$. The cost for the reduction starting at position $j+1$, which overlaps the one starting at position j by all but the first and last positions, can be deduced by substracting and adding the corresponding cells of H in constant time. Thus, the whole computation takes $O(nk)$ time.

2.1.2. *A 2-approximation for Fixed Boundaries Order-1 Amplifications.*

The approximation presented in [4] is for what the authors call the "restricted problem": when amplifications have a fixed boundary and their order is 1. In the restricted case, the history can be represented by a leaf and edge ordered and labeled tree whose leaves represent the copies of the tandem repeats. Such a tree is called a **duplication tree**. Internal nodes are labelled by ancestral copies. To an edge (t_j, t_l), where $t_j, t_l \in \Sigma^k$ are the labels of the linked nodes, is associated the cost $d(t_j, t_l)$.

The proof relies on the concept of **ordered spanning tree** (OST). Given an ordered set of nodes $V := \{1, \ldots, n\}$, an OST is a spanning tree on V such that for any two edges (b_1, e_1) and (b_2, e_2), $b_1 < e_1$, $b_2 < e_2$, we have $(b_1 - b_2)(b_1 - e_2)(e_1 - e_2)(e_1 - b_2) \geq 0$. In other words, if the nodes are placed on a line, the edges are on the same side of the line on the plan and do not cross each other. An example of an OST is given in Figure 5. We will consider OSTs on the ordered set of copies, $s_1, \ldots, s_n$ and assign $d(s_j, s_l)$ as the distance between nodes j and l, $1 \leq j, l \leq n$. For both types of trees, the cost of a tree is the sum of the costs of its edges.

An ordered spanning tree B on $s_1, \ldots, s_n$ describes the structure of a unique duplication tree. It is possible to construct the latter from the former in linear time; this is illustrated in Figure 5. The leaves of the two trees are the same. To construct the duplication tree, one starts from a node of degree one in the OST and follow the edges. From each edge one reaches a new leaf, if edge entering that leaf go on

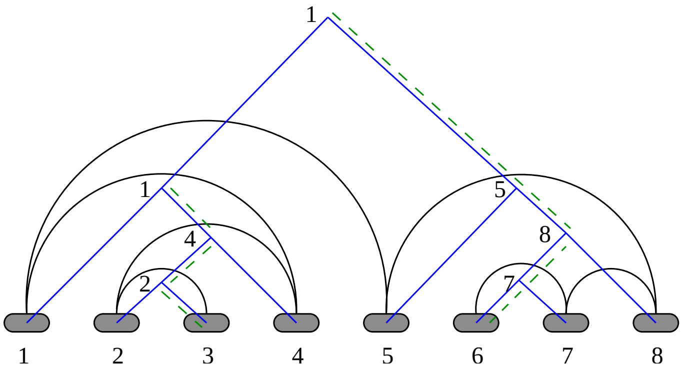


Fig. 5. An ordered spanning tree on $n := 8$ nodes (in curved lines) and its associated duplication tree drawn (in straight lines). The number at an internal nodes represents the leaf label that has been lifted up to that node. In the lifted duplication tree, only edges mark by dashed lines have non-zero cost; all other edges have the same label at each extremity and cost zero.

the same side of the line, one creates an internal node, branches the leaf on one side and the subtree constructed so far on the other side. Otherwise, the branching of the leaf depends on the weigths of the edges. For leaf number seven for instance, the internal node links it to leaf 6 because $d(s_6, s_7) < d(s_7, s_8)$. If $d(s_6, s_7) > d(s_7, s_8)$ one would have created two internal nodes and put leaves 7 and 8 in the same subtree, and leaf 6 at the end of the other branch. This situation happens for leaf 5 and at this step the second internal node is the root of the tree.

If in the duplication tree, one assigns to each internal node the label of one of the leaves in its subtree, one obtains a duplication tree B whose cost equals the one of the ordered spanning tree. We propose to call B a **lifted duplication tree**. The duplication tree given in Figure 5 with the leaf labels associated to internal nodes is a lifted duplication tree. The same idea of lifting the leaf labels up the tree was proposed by Wang and coworkers [51] to compute a 2-approximation for the Phylogenetic Multiple Alignment problem and their approximate tree was termed a **lifted tree**.

Proof. [of the 2-approximation.] Let P be the optimal duplication tree, and R be a depth-first traversal of P. As each edge of P is visited twice in R we have $cost(R) = 2 \times cost(P)$. Let SR be the cycle obtained by visiting the leaves in order and cycling from leaf s_n to leaf s_1. As d satisfies the triangle inequality, $cost(SR) \leq cost(R)$. Now, removing any edge in SR yields an ordered spanning tree TR. Thus, we have $cost(\mathrm{TR}) < cost(SR) \leq cost(R) = 2 \times cost(P)$; TR is a 2-approximation of P. □

An example of an optimal duplication tree with $n := 8$ copies, its depth-first traversal and a trivial cycle (i.e., corresponding in the proof to P, R, and SR, respectively) are given in Figure 6. Note that the topology of TR varies very little with the input (since TR equals SR minus one edge) and presents no biological interest

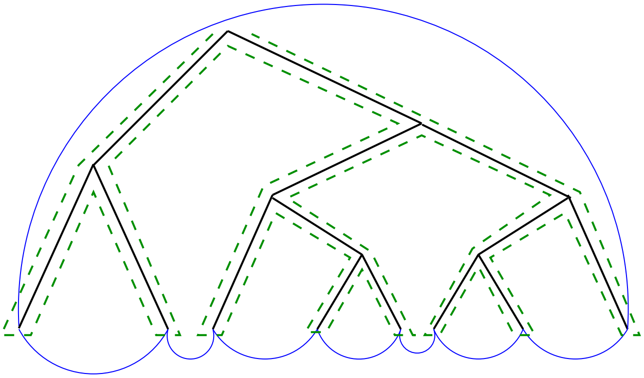


Fig. 6. An optimal duplication tree on $n := 8$ leaves (straight lines), a depth-first traversal of it (dashed lines), and a simple cycle on the leaves (curved lines).

in itself. On the opposite, a true minimum OST and its associated duplication tree provide an informative approximate solution. Benson and Dong exhibit a dynamic programming recurrence to compute the minimum OST (on all possible intervals of lines in M) in $O(kn^2 + n^3)$. We will see below that Tang et al. [47] also use the lifting technique to obtain a 2-approximation for this problem in $O(n^2(k + n^3))$.

Open question:

— Is MP-TRH with variable boundaries approximable?

2.2. *Tang et al. Dynamic Programming Approach.*

In their framework, Tang and coworkers [47] allow only duplications, that is amplifications of variable order but of arity 2, and fixed boundaries. They generalize the duplication tree introduced by [4] in a **duplication model** to enclose information relative to duplications of order higher than one. We first describe this model. We present their dynamic programming scheme for what they call the Single Gene Duplication problem, that is FBMP-TRH restricted to arity 2 and order 1. (Actually, the authors do not really tackle with their general model.) In Benson and Dong's vocabulary, this corresponds to the "restricted" FBMP-TRH. The dynamic programming scheme is flexible and by combining it with the lifting technique, they obtain a 2-approximation and claim a PTAS can be achieved with the same technique. The authors also give an algorithm for the distance-based problem, but we will delay its presentation until Section 2.6.

2.2.1. *The Duplication Model.*

The duplication model is a duplication tree (from [4]) with higher order duplications. A duplication of order i duplicates i adjacent copies, say $s_1 \ldots s_i$, into

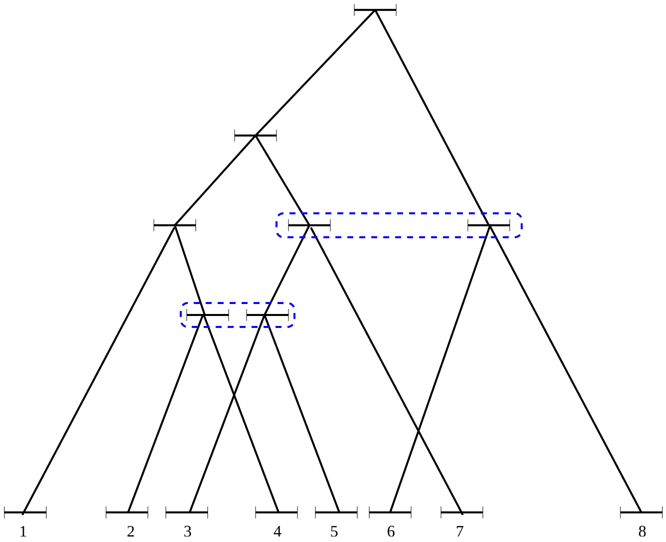


Fig. 7. An example of a duplication model or of an unrooted duplication tree with $n := 8$ leaves. Each node represents a copy of the tandem repeat, either an ancestral one at an internal node or a present one at a leaf. Subsets of internal nodes that are duplicated together by a duplication of order 2 are surrounded by a dashed line. These correspond to blocks in the terminology of [47].

$s_{1,l} \dots s_{i,l} s_{1,r} \dots s_{i,r}$[c] where $s_{j,l}, s_{j,r}$ denote the right and left children of s_j. As a duplication tree is ordered, the left to right order of nodes represents the sequence order of the copies; this is valid for both internal nodes and leaves. If we place a duplication of order i in such a tree, it must take i nodes representing adjacent copies and create i left and i right children. Therefore for $j < h \leq i$, the edge connecting s_j and its right child $s_{j,r}$ crosses all edges connecting s_h and its left child $s_{h,l}$, i.e., $(s_h, s_{h,l})$. This the only way two edges can cross in the model and it requires that $i \geq 2$. Such a model is said to be **consistent** with the ordering of the leaves/copies on the sequence. This is the main constraint that a tree must satisfy to be a duplication model. Nodes involved in higher order duplications form a **block** and all blocks are memorized in the duplication model. Figure 7 illustrates the notion of duplication model.

For a given duplication model, it is possible change the order of nodes to uncross all edges and obtain a unique planar tree topology. Tang et al. name it the associated **phylogeny**. The term phylogeny means a binary tree without order constraint that represents the evolution of genes/species associated with its leaves. A way to answer our problem is to use a phylogenetic reconstruction method on the set of copies without the order constraint, and check if the returned phylogeny has an associated duplication model. There exist efficient heuristics for phylogenetic

[c] Note that there are no commas in the notation; it means that the copies are adjacent on the chromosome.

reconstruction (for instance [22]). Therefore, Tang et al. investigate what we call the **Recognition problem**, that is to find the duplication model associated with a rooted phylogeny if it exists, and they give a $O(n^2)$ algorithm for it. Note that the problem size is n, i.e., the number of leaves of the phylogeny. We do not describe their solution since better algorithms have been published since then in [23, 52] (see Sections 2.4 and 2.5).

2.2.2. *The Dynamic Programming Scheme.*

In [47], the authors define the ME-TRH with fixed boundaries with an arbitrary additive cost function for the reductions. However, the dynamic programming scheme they present is further restricted to the FBMP-TRH with order 1 duplications only. In this case, the duplication model has no edges crossing each other and is simply a tree on an ordered set of leaves. In the remainder of this section, we use the word "tree" instead of "duplication model" for simplicity. The authors consider the Hamming distance between words, and not the Levenshtein distance as in [4].

Let s belong to Σ^k, i, j be integers such that $1 \leq i \leq j \leq n$ and S, S_l, S_r be subsets of Σ^k. A tree for an interval subset of the copies, $s_i, \ldots, s_j$, is said to **span** the interval $[i, j]$. With the Maximum Parsimony criterion, each internal node is labelled with a putative ancestral copy. We introduce a notation used below in the recurrence. Let $D([i, j])$ be the cost of an optimal tree spanning $[i, j]$. Let $D([i, j], s)$ be the optimal cost of a tree spanning $[i, j]$ and whose root is labelled by s. If $i < j$, let an integer m satisfy $i \leq m < j$ and $D([i, j], s, m)$ be the minimum cost of tree spanning $[i, j]$ whose root is labelled by s, and whose left and right subtrees span $[i, m]$ and $[m + 1, j]$, resp. For $s \in \Sigma^k$, $D([i, i], s)$ is initialized to 0 is $s = s_i$ and to infinity otherwise. It follows from these definitions that:

$$D([i,j]) = \min_{s \in S} D([i,j], s) \tag{1}$$

$$D([i,j], s) = \min_{i \leq m < j} D([i,j], s, m) \tag{2}$$

$$\begin{aligned} D([i,j], s, m) = & \min_{v \in S_l} (D([i,m], v) + d(s,v)) \\ & + \min_{w \in S_r} (D([m+1,j], w) + d(s,w)) \end{aligned} \tag{3}$$

A tree spanning the whole repeat can be recovered by backtracking through the matrices from entry $[1, n]$. If one chooses S, S_l, S_r equal to Σ^k the returned cost and tree are optimal, but the running time is exponential in k. Indeed, it is in $O(|\Sigma|^{2k}(k+n^3))$ since computing all Hamming distances between any two strings of length k is done in $O(k \times |\Sigma|^{2k})$ and filling the matrices for each of the n^2 intervals takes $O(|\Sigma|^{2k} n)$ time.

Now, if one restricts S, S_l and S_r in such a way that to an internal node is associated the label of either its left or of its right child, the result is a lifted duplication tree. By a lemma from [51], the cost of such tree is at most twice the cost of an optimal tree. This gives a 2-approximation algorithm in $O(n^2(k + n^3))$

time. The authors also report that the PTAS developed by Wang and Gusfield [49] can be adapted to this problem.

The 2-approximation of [47] yields a lifted tree as the 2-approximation of [4] obtained from the trivial cycle. The latter is more effective since it requires linear time. Its improvement using the minimum OST is computed in only $O(kn^2 + n^3)$ time. Another remark is that the dynamic programming scheme delivers an optimal solution in a time that is only exponential in k. Thus, FBMP-TRH with order 1 duplication is Fixed Parameter Tractable for parameter k. We refer to [15] for details on Parameterized Complexity.

2.3. *An Exhaustive Exploration Approach*

Elemento and coworkers [18, 19] investigate the same problem as in [47]: recovering the duplication tree of tandemly repeated genes. They define independently the concept of duplication model and call it a **partially ordered duplication history**. They name such an unrooted history a **tandem duplication tree**. The authors argue that the Fixed Boundaries restriction applies here, because the main mechanism of gene amplification is unequal recombination. They exhibit an exponential algorithm, named *DTExplore*, that exhaustively searches the space of duplication histories. The algorithm works for a limited number of copies, in the order of $n = 10$. In addition to some applications, the algorithm is used to compare empirically the number of phylogenies with n leaves with the number of duplication models with n copies. The authors conclude that, although the number of duplication models seems exponential, it represents a small fraction of the number of phylogenies. This question is addressed later on in [23] and is detailed in Section 2.4. We also present their $O(n^2)$ solution for the RECOGNITION problem as defined in Section 2.2.1.

Like in [47], authorized amplifications are duplications (i.e., of arity 2) of variable order. In their vocabulary, a duplication of order x is denoted an x-duplication. The distance measure between gene copies is the Hamming distance and no cost is charged for amplifications (i.e., $d = d_H$ and $\mathcal{A}(i, m) = 0$ for all i, m).

2.3.1. *An Exhaustive Search Strategy.*

The algorithm *DTExplore* simulates the duplication process to explore in a depth-first search manner the space of all duplication histories of n leaves. It starts with a rooted tree with two leaves and applies a duplication of order x to obtain a tree with $2 + x$ leaves. The process is iterated until the number of leaves reaches n. The current phylogeny is given as input to Fitch's algorithm [20] which computes its Maximum Parsimony score (i.e., the optimal ancestral copies for all internal nodes). The algorithm backtracks and duplications of any possible order are applied to visit all topologies. *DTExplore* outputs the phylogenies with the minimal score.

This first version of *DTExplore* suffers from redundancy as it generates several time the same rooted duplication model. An improvement is achieved by enabling *DTExplore* to memorize which duplication model has already been visited. This is performed by encoding each model in a character string and storing all codes in a prefix tree. Each generated topology is first encoded, searched for in the prefix tree, and its maximum parsimony is evaluated only if it was not found in this data structure. The speed improvement is drastic.

2.3.2. *A Simple Recognition Algorithm.*

Given the topology of a phylogeny on n leaves, the Recognition problem is to decide if the topology also is a duplication model. Elemento et al. [19] report a simple $O(n^2)$ algorithm for it. It iteratively reduces the phylogeny by replacing the $2i$ leaves of an order i duplication by their fathers. At most n such steps are performed, if all duplications were of order 1. At each step, to identify a possible duplication, it searches for a subset of adjacent $2i$ leaves that are appropriately intermingled. If this search fails, the phylogeny is not a duplication model and the algorithm returns false. Otherwise, it stops with a tree reduced to a root and answers true. The search is done in $O(n)$ time at each step. After investigating the combinatorics of duplication models, Gascuel et al. [23] improve the algorithm's complexity to $O(n)$ (cf. Section 2.4).

2.4. *Combinatorics of Duplication Trees*

In the same framework than [19], Gascuel and coworkers [23] investigate the cardinality, denoted $\mathrm{rdt}(n)$, of the set $\mathrm{RDT}(n)$ of rooted duplication trees with n leaves. They exhibit a recurrence for it and show it is the double of the number of unrooted duplication trees. They deduce an algorithm to uniformly sample duplication trees and a linear time procedure for the Recognition problem. Here again, amplifications are of variable order and of arity 2.

2.4.1. *Counting Duplication Trees.*

We review the main recurrence for rooted duplication trees. To obtain this recurrence, the authors introduce the notion of an **(l,i) duplication** where i is the order and l the number of copies located after the last copy involved in the duplication. So, an (l, i) duplication duplicates i copies $s_{n-l-i+1} \dots s_{n-l}$, where i and l satisfy $1 \leq i \leq n$ and $0 \leq l \leq n-2i$. Given a rooted duplication tree R, an (l, i) duplication is said to be **visible** if none of the $2i$ copies it created has been further duplicated in R. Let $P(n, l)$ be the subset of $\mathrm{RDT}(n)$ whose leftmost visible duplication is an (l, i) duplication for some i such that $1 \leq i \leq (n-l)/2$. Let $p(n, l)$ be the cardinality of $P(n, l)$. By definition, for $l > n-2$, $P(n, l) = \emptyset$, $p(n, l) = 0$ and $p(2, 0) = 1$.

Theorem 1. *Let $n > 2$ and $0 \leq l \leq (n-2)$. $P(n, l)$ and $\cup_{j=0}^{l+1} P(n-1, j)$ are in one-to-one correspondence.*

Proof. Let us denote the substring of $2i$ copies created by an (l, i) duplication in the present sequence by $s_{n-l-2i+1} \dots s_{n-l-i} s_{n-l-i+1} \dots s_{n-l}$. For all $1 \leq f \leq i$,

$s_{n-l-i+f}$ is the twin of $s_{n-l-2i+f}$, i.e., they are offspring of the same father copy. Let $T \in P(n,l)$; deleting s_{n-l-i}, the left child of the rightmost copy duplicated by the (l,i) duplication, in T maps T to T'. If $i = 1$ then the leftmost visible duplication in T' is at the right of s'_{n-1-l} and thus, T' is in $P(n-1,j)$ for some j in $0 \leq j \leq l$. If $i > 1$ the (l,i) duplication in T becomes the leftmost visible duplication in T' and is an $(l+1,i-1)$ duplication; so T' belongs to $P(n-1,l+1)$. As the transformation is reversible the mapping is a bijection. □

As all trees have a leftmost visible duplication, if we denote the cardinality of $\mathrm{RDT}(n)$ by $\mathrm{rdt}(n)$, we have $\mathrm{rdt}(n) = \sum_{l=0}^{n-2} p(n,l)$. Combined with the recurrence of Theorem 1 it provides a way to compute $\mathrm{rdt}(n)$.

Although the root of a duplication tree is necessarily located on the path between the left- and right-most copies ([21]), not all edges on this path are possible root locations. First, the root cannot be below some duplication of order strictly greater than 1. Second, it comes out that in average only two locations are possible. Hence, the surprising result that the number of unrooted duplication trees with n leaves equals $\mathrm{rdt}(n)/2$. Asymptotically, this cardinality behaves like $(\frac{27}{4})^n$ when n tends towards infinity.

2.4.2. *An $O(n)$ Recognition Algorithm.*

The concept of visible duplication allows to improve the recognition algorithm described in Section 2.3.2. It proceeds by iteratively agglomerating leaves of a duplication (such leaves belong necessarily to a visible duplication). The improvement consists in choosing the leftmost visible duplication at each step; an amortized analysis shows that all steps take $O(n)$ time alltogether. For this, the scan for a duplication proceeds from left to right and the endpoints of already encountered blocks are memorized.

Open questions:

— In the case of fixed boundaries, study the number of amplification trees when amplifications of variable arity are allowed.

2.5. *Complexity, Approximability and Other Results*

In [31], Jaitly et al. consider the FBMP-TRH, with a maximum arity 2, with the Levenshtein distance between strings $d := d_L$, and no cost for amplifications. They proved that this restriction of FBMP-TRH is NP-hard by reducing it to the Max-Cut problem. They exhibit a PTAS that uses the lifting technique to partition the topology and exact optimization by dynamic programming to compute optimal labels for subtrees of constant size. A detailed sketch of the proof is given in [31] and it relies on previous difficult approximation results for the Phylogenetic Multiple Alignment problem by Wang and coworkers [51, 49, 50]. As in [4], the authors notice the relation of FBMP-TRH with the Ordered Leaves Steiner Tree problem. They show their algorithm also is a PTAS for the latter.

Zhang et al. [52] describes a $O(n)$ algorithm for the Recognition problem from rooted phylogenies (with duplications of variable order, the distance used over Σ^* is not given). Assume the leaves of the phylogeny are numbered in order from 1 to n. Their method identifies for all leaf $j := 1, \ldots, n$ the internal nodes belonging to duplications of order ≥ 2 (to blocks in the terminology of Tang et al.) that allow leaves j and $j + 1$ to be put next to each other in the duplication model. For this, they associate to each node v the pair $(l(v), r(v))$ where $l(v), r(v)$ denote resp. the smallest and largest leaf numbers in the subtree of v. Fast identification of block's nodes is achieved by comparing pairs $(l(v), r(v))$ of nodes on the path to leaf j to those of nodes at the same level on the path to leaf $j + 1$. Given an unrooted phylogeny, the search for a duplication model has to be performed for all $O(n)$ possible root locations because in practice one infers unrooted phylogeny. Thus, although in linear time, their algorithm is less effective than the one of Gascuel et al. [23] that deals with both case of rooted and unrooted phylogenies.

Zhang and coworkers also proposed a greedy search strategy for the FBMP-TRH. It infers a phylogeny with a traditional reconstruction method, checks if it is associated to a duplication model, and if not attempts a transformation of the topology (like a Nearest Neighbor Interchange). The two last steps are iterated until a duplication model is found. Such a strategy was first developed in the field of phylogeny reconstruction. The authors report it performs better than Benson and Dong's greedy algorithm or Tang et al.'s Window method on three real data cases.

Open questions:

— Does the MP-TRH with variable boundaries admit a PTAS?

— For some parameters, are these problems fixed parameter tractable in the sense of Fellows and Downey's theory of parameterized complexity [15]?

— Improve the Zhang et al.'s greedy strategy by inferring directly a duplication model and invent a transformation operation that allows to visit each possible duplication model on n leaves.

2.6. *Minimum Evolution Approaches*

As mentioned in the introduction, the criterion of Minimum Evolution borrowed from Phylogeny leads to a formalization of the problem that differs from MP-TRH. The data is a matrix D giving the pairwise distance between any pair of copies in the sequence T and the ordered list of copy numbers. The output is a duplication tree whose sum of the branch lengths is minimum. We call this problem the Minimum Evolution Tandem Repeat History (ME-TRH).

This formalization implicitly considers fixed boundaries since the history is represented by a tree. It is thus not as general as MP-TRH. Moreover, as the input is a distance matrix, the sequences are disregarded in the remaining of the algorithms. The link with combinatorics and algorithmics on words resides only in the computation of the input pairwise distances. Nevertheless, we include a section on these methods for the sake of completeness and because such approaches were shown to be reliable in practice [16].

The two methods presented here, *DTScore* from [16] and *Window* from [47], optimize a local criterion on the tree and are based on the same algorithmic scheme. A current list of leaves in the tree is maintained and initialized to the original list of copies. The methods iterate an agglomeration step that replaces a subset of $2i$ adjacent leaves resulting from an order i duplication by new leaves representing their parent copies. Such a subset of $2i$ leaves is called a **window** in [47]. Entries of the distance matrix corresponding to the deleted copies are removed and entries for the parent copies are inserted. Entries for the parent copies are the average of their children distances. The algorithms proceed until 2 or 3 leaves are left. The window is chosen to maximize a score function. *DTScore* and *Window* differ in the choice of this function. *Window* considers the average of the Hamming distances between twin copies j and $j+i$ for all possible j. This is known to produce the correct tree if the data respect the **Molecular Clock** hypothesis, i.e., if evolution proceeded at the same pace in each branch of the tree. In practice, biologists proved it is not often the case. In *DTScore*, the score of a pair $(j, j+i)$ is the number of times j and $j+i$ are next to each other in every possible quartet of leaves $(j, j+i, l, m)$ according to the **Four-Point Condition** ([11]). The score of a window is the minimum of the scores of its pairs; the window with the maximum score is selected. This scheme works even if the Molecular Clock is not respected.

Empirical tests on pseudo-randomly generated trees and sequences confirm *DTScore*'s improved ability to recover the correct tree compared to *Window* and to maximum parsimony approaches (cf. [16]). For both programs, the time complexities is $O(n^4)$. Elemento and Gascuel [17] consider the ME-TRH under the global criterion of the ordinary least square errors. They propose an exact algorithm in $O(n^3)$ time and $O(n^2)$ space when the duplication order is restrained to 1.

Open questions:

— Find an algorithm to optimize ordinary least square errors when the order is variable.

2.7. *Biological Validations*

Most of the literature referenced for the TANDEM REPEAT HISTORY problems includes tests on real data sets, mainly on clusters of tandemly repeated human genes. For instance, Zinc Finger genes in [47, 52], genes for Olfactory Receptors and the internal tandem repeat of a mucin gene MUC5 in [52], as well as two immunological gene loci, TRGV and IGLC, in [19]. In all cases, the authors found the computed histories were consistent with what is known about the evolution of the gene family and most histories include mainly order 1 duplications.

The most detailed discussion on biological validation is in [19]. The mechanism for gene duplication is hypothesized to be unequal recombination. In the case of the TRGV locus where $n := 9$, the history reconstructed with *DTExplore* is also found by *DNAPENNY*, an exact phylogenetic reconstruction method ([28]), although the probability of finding a duplication model when exploring the space of phylogenies is already low for $n = 9$ (see Section 2.4). Moreover, in the tree for the TRGV locus, the most recent duplication has order 2 and corresponds to a polymorphism observed in some human populations. Indeed, the two additional gene copies are missing in individuals of the Tunisian, Lebanese, French, Black-African and Chinese populations. These two evidences combined with a bootstrap test constitute a strong validation of the model with respect to the mechanism of duplication.

With the availability of complete genomes, numerous tandem gene clusters are discovered and represent potential data for such analysis. We believe biologists will investigate tandem duplication history once the methods described here have been advertised more widely.

3. Allele Alignment

Among polymorphic tandem repeats, hypervariable minisatellites cumulate variations in their number of copies, as well as in the sequence of the repeats. They belong to the most polymorphic markers: at a single locus, one encounters much more alleles than for other loci. Hence, the numerous differences between alleles provide us with detailed information on the evolutionary processes at work. To understand variation in the sequences of a tandem repeat locus, we need to be able at least to compare alleles in a pairwise manner. If one can measure the dissimilarity by a metric, it becomes possible to infer evolutionary relationships of a set of alleles using distance-based phylogenetic methods. Even better would be to simultaneously align several alleles and then use Maximum Parsimony phylogenetic reconstruction to compute a tree with ancestral alleles and to count mutations along the branches. In this section, we consider the problem of Pairwise Allele Alignment. At present, Multiple Allele Alignment remains an interesting future work.

Brard and Rivals [7, 9] introduced this problem and noted that the events are not commutative, which introduces a major difficulty compared to classical Sequence Alignment [45]. In the case of amplifications and contractions of order 1 and arity 2, they describe an exact algorithm in $O(\max(m, n)^4)$ time, where n, m denote the sequences lengths. The alignment distance is a metric and can serve as input for distance-based phylogenetic reconstruction. In [9], the method is applied to a set of alleles from the MSY1 locus, a minisatellite on the human Y chromosome. These biological experiments demonstrate that the approach enables recovering allele relationships. It is shown for instance that phylogenetic alleles having a common origin are segregated according to the population of their bearer. The approach used to compare minisatellite alleles is also valid for other types of markers. Recently, another algorithm has been presented in [10] for the case of amplification and contraction of order 1 and a maximal arity denoted by ρ. It works in $O(max(m,n)^3\rho\sigma)$ time and $O(max(m,n)^2\rho\sigma)$ space.

In the remaining of this Section, we present the notion of minisatellite maps and the hypothesized evolutionary model, we formally define the problem, and show that the non-commutativity forces us to consider the order in which events occur to find an optimal alignment. We then describe the solution of [7, 9] that combines dynamic programming and computations of maximum independent sets in overlap graphs. We present the dynamic programming scheme of [10] and conclude with biological validations.

3.1. *Minisatellites Maps and the Evolutionary Model*

Along the tandem array of a minisatellite, the repeated unit varies in sequence. For a fixed phase, the different adjacent substrings are called **variants** of the repeated unit. In 1991, Jeffreys and colleagues designed a reaction to obtain the sequence of variants of the array. This specific Polymerase Chain Reaction (PCR) is called the **Minisatellite Variant Repeat** reaction or MVR-PCR for short [32]. It yields a sequence in which each variant/substring is encoded by a symbol. Such sequences are called **minisatellite maps**. Note that boundaries of the variants are fixed by the technology. This technology enabled investigations of the processes responsible for minisatellites instability. Here, we deal with the problem of aligning two minisatellite maps (which differs from aligning their DNA sequence). From the computational view-point, maps are still sequences over a finite alphabet, which is not the DNA alphabet. In the sequel, we use the terms string, sequence or map without distinction.

The evolutionary model defines the set of events a map can undergo and associate a real cost with each event. Here, we consider point-mutations, amplifications and contractions of order 1 and arity 2 as defined in Section 1.3. Note that events operate on variants instead of on DNA bases; e.g., a mutation changes a variant into another, a deletion removes a complete variant from the map, etc. To each event is associated a fixed real cost that we denote: A, C, D, I, M for resp. amplification, contraction, deletion, insertion and mutation. We assume our model is symmetrical: dual events have equal costs ($A = C$, $D = I$). To stick to biological conditions, we assume $A, C < M, D, I$. Note that if $I > A + M$ then an amplification followed by a mutation is always preferred to an insertion. Without loss of generality, we make this hypothesis on costs.

3.2. *A Dynamic Programming Approach Combined With Graph Algorithms*

From now on, let r, s be two maps resp. of length n and m over the alphabet Σ. Let $-$ be an additional symbol that does not belong to Σ and denotes an inserted or deleted position in an alignment. An edit script between r and s is a sequence of events that transform r into s. An alignment of r and s is a representation of an

Individual 1

Event	Sequence
	a a a a a
mutation	a **e** a a a

Individual 2

Event	Sequence
	a a a a a
mutation	a a a **b** a
5*amplification	a a a b **b b b b b** a
mutation	a a a b b **c** b b b a
mutation	a a a b b c b **d** b a
amplification	a a a b b c b d **d** b a

Fig. 8. Example of evolution of a minisatellite in two individuals.

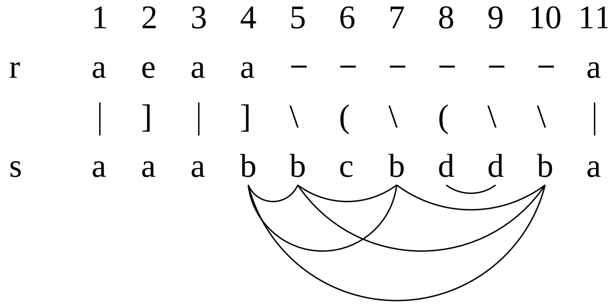


Fig. 9. Alignment of the maps $r := aeaaa$ and $s := aaabbcbddba$. The arch *bbcbddb* and its inner-arches are drawn by curved lines under s. In the middle line, '|', ']', '\', '(' denote resp. a match, a mismatch, an amplification, and an amplification+mutation.

edit script that respects the order of the positions in r and s. The alignment cost is the sum of its operation costs where aligning two identical variants costs zero. It is shown that the alignment cost is a distance metric [9]. An example of evolution of different alleles from the same ancestor allele is given in Figure 8 and an optimal alignment for the two resulting maps is shown in Figure 9.

Definition 5 (Allele Alignment). *Given two maps r and s respectively of length n, m over an alphabet Σ and a alignment scoring scheme, find the alignment between r and s that has minimum cost.*

In Figure 8, the substring of s from position 4 to 10, *bbcbddb*, shows that the final variant at each position does not appear in the order of the sequence: at some stage, position 8 has still the ancestor state b and not its final state d, while position 10 is already a b. The minimal cost series of events to create such a substring is order-dependent. If one computes incrementally alignment for longer and longer prefixes, we cannot find the optimal order of events. This happens when aligning a position in r with a substring of s having identical first and last characters. Such a substring is called an **arch**. The first position of the arch in s is aligned to the position in r, and all other arch's characters are generated in s from the first position. The authors prove that at least in an optimal **generation** of the arch, the last character is obtained by an amplification of the first when the positions are adjacent, and then all other positions in between are generated afterwards. Thus, one needs to consider the arch as whole, and not in order of increasing prefixes. Such an arch generation avoids a mutation per arch. An arch may include other inner-arches, but an alignment cannot include all possible arches. Because of the optimal order of generation, two arches cannot belong to the same alignment if they are **incompatible** with each other, i.e., if they overlap each other by more than one variant (here overlap means that strict inclusions are allowed). It follows that an optimal arch generation should contain the largest number of pairwise compatible arches. The symmetrical situation, when the arch in r is aligned to a single symbol in s, is called an arch **compression**.

Arches represent intervals of a map and incompatibility defines an overlap relationship between these intervals (not an overlap+containment relationship). Consider the graph G whose nodes are arches and whose edges link two nodes if their arches are incompatible. G is an overlap graph. It is shown in [9] that computing the maximum subset of compatible arches is equivalent to finding a maximum independent set in G.

A preprocessing procedure will take as input a map t and compute for each possible interval the cost of the corresponding arch generation or compression (if any) using an adaptation of Apostolico and coworkers's algorithm [1]. This procedure is applied to s and to r and results are stored in two matrices that require quadratic space. Once this is done, one can get the generation cost, resp. the compression cost, of any arch of s, resp. of r, in constant time.

The alignment problem is solved by filling a dynamic programming matrix $\mathcal{A}$ whose entry $\mathcal{A}(i,j)$ is the optimal alignment cost between the prefixes of s and r of length i and j resp. Five dependencies between $\mathcal{A}(i,j)$ and its direct neighbors account for the five possible evolutionary events: amplification, amplification+mutation (A_M), mutation/match, contraction, and mutation+contraction (M_C). (Note that there are no insertion, nor deletion because of our hypothesis on costs.) Moreover, up to i, resp. j, dependencies with non adjacent entries on the same line, resp. on the same column, account for arch generations, resp. arch compressions. These costs are denoted $G_{l'}$ and K_l. These dependencies can be evaluated in $O(1)$ time thanks to the preprocessing. The dynamic programming recurrence is illustrated in Figure 10. The algorithm requires $O(\max(m,n)^4)$ time and $O(\max(m,n)^3)$ space.

Open questions:

— What is the complexity class of PAIRWISE ALLELE ALIGNMENT when higher order amplifications and contractions are allowed? Find a practical solution.

— An algorithm for MULTIPLE ALLELE ALIGNMENT under any of the assumptions mentioned above.

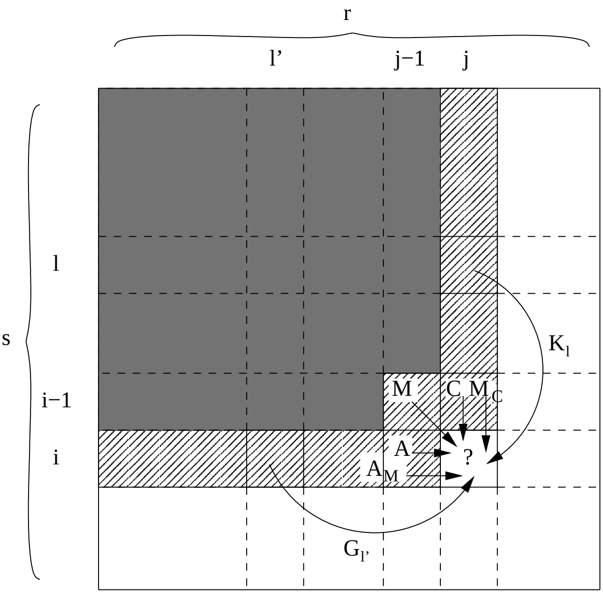


Fig. 10. Dependencies in the dynamic programming matrix. To compute cell $\mathcal{A}(i, j)$, we need at most all cells in the striped patch but not the ones in the dark patch. Dependencies are shown by arrows. For arch generations and compression the arrows are indexed by the beginning position of the arch to show that there are multiple dependencies.

3.3. *A Second Dynamic Programming Scheme*

Between the submission and acceptance of this survey, another work on Allele Alignment has been published, showing that this field is active. In [10], the authors consider Allele Alignment with amplifications and contractions of order 1 and variable arity, which is a more general model than in [9] where arity is limited to 2. Moreover in their setup, the cost of events may differ according to the symbols involved (E.g., an insertion of a variant a may cost more than that of a b). Despite this generalization, their solution is based solely on dynamic programming and requires $O(max(m, n)^3 \rho\sigma)$ time and $O(max(m, n)^2 \rho\sigma)$ space where ρ denotes the maximal arity of amplifications and contractions. Note that the vocabulary used in [10] may confuse the reader, since what we defined as the arity is called "order" and the alignment distance is termed "Transformation Distance". We stick to our terminology.

In their approach, the authors consider the alignment of source map s with a destination map r as a series of **reductions** of segments in s and of **generations** of segments from r. This is illustrated in Figure 11. In s, diamonds represent **generative** symbols from which the corresponding segment of r is generated (the segment at the opposite face of the triangle), while black circles are **vanishing** symbols. All vanishing symbols between two consecutive diamonds are reduced into the leftmost diamond, the associated segment in r is generated from the latter. The same process

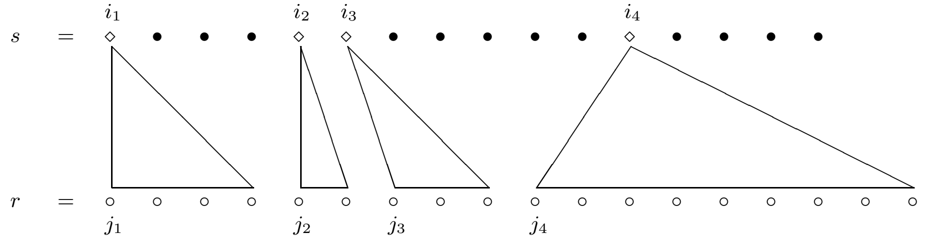


Fig. 11. Alignment of minisatellite maps as in [10]. A •, a ◇ denote respectively a vanishing or a generating symbol of s, and a ○ represents a symbol of r.

is repeated for each diamond in order of appearance in s until r has been generated. The main dynamic programming reccurrence optimizes the decomposition of s and r to minimize the alignment cost. Note that in this approach, the first symbol of the map necessarily is a generative symbol.

The correctness of this process resorts on two lemmas. The "independence" lemma asserts that any reduction may surely be performed before any generation. The "generation" lemma states that an optimal generation is a non-decreasing, i.e., does not include deletions nor contractions. This implies that the length of string being generated never decreases. Such a generation is made of amplfications, insertions and mutations and, although it is not stated in [10], can be modeled by a regular grammar. The authors associate a derivation tree to all non-decreasing generations that yield the same string and have equal cost. Such equivalent generations differ by the order in which operations are applied and this is because some operations commute.

In a preprocessing phase, the algorithm computes by dynamic programming the tables $T[i,j,x]$ and $S[i,j]$ where the former gives the optimal generation cost of $r[i,j]$ from symbol x and the latter the minimum reduction cost of $s[i,j]$ into $s[i]$. This computation of each table involves two other dynamic programming matrices we do not detail here. Then, the computation of the alignment cost is done in the table TD, where $TD[i,j]$ equals the cost aligning $s[1,i]$ with $r[1,j]$, with the following recurrence:

Initialisation

$$TD[0,0] = 0 \quad \text{et} \quad TD[0,j] = \infty \quad \forall j > 0\,,$$

Recurrence

$$\forall i > 0 \quad TD[i,j] = \min \begin{cases} TD[i-1,l] + R[l+1,j,s[i]] & \forall 0 \le l < j \\ TD[k,j] + S[k,i] & \forall 0 < k < i \end{cases} .$$

The description of the algorithm is incomplete and several details are lacking for reproducibility. In their setup, the insertion event is not allowed at the first position of the sequence, but no explanation justifies this restriction. To circumvent this, an extra $ symbol that does not belong to the alphabet is added in front of the maps.

The authors align $\$s$ with $\$r$ and consider that any event involving a \$ has infinite cost except the match. Hence, the dollar serves as a generative symbol. Let us look at an example where $\Sigma = \{a, b\}$ and we want to align $s := \$ba$ with $r := \$a$ with the costs $A = C = 1$, $M = 10$, $D = I = 20$. With this algorithm, there are two generative symbols since the dollar is systematically one of them and the best symbol to generate the a in r is the a in s. It comes out that the alignment matches the dollars, deletes the b and then matches the a's. This costs 20. With this scheme, there is no way to mutate the b in a, contract an a, and then match the dollars and the a's. The latter costs $10 + 1 = 11$. The corresponding alignments are:

Alignment of [10]

$$\begin{pmatrix} \$ & b & a \\ | & & | \\ \$ & - & a \end{pmatrix}$$

Optimal alignment

$$\begin{pmatrix} \$ & b & a \\ | &] & \backslash \\ \$ & a & - \end{pmatrix}$$

This counter-example of S. Brard shows the algorithm is not correct [8]. To fix it, one could allow the initial block of symbols in s to be reduced into the next generative symbol. Another detail: with the main reccurrence equation it is not possible to compute $TD[1, 0]$.

In conclusion, the algorithm combines a dynamic programming scheme and the point of view of formal languages, which may be appropriate when considering further extensions of the model, e.g., with variable order amplifications. Note that both the independence and generation lemmas are not valid anymore in that case. Hence the open questions stated at the end of previous section remain. In the model considered in [10], the cost of events vary with the variants involved; unfortunately, this model has not been validated on real biological data.

3.4. *Biological Validations*

The algorithm of [9] described in Section 3.2 is implemented in a program named *MS_ALIGN* and has been applied to alleles of the human minisatellite MSY1 [30]. MSY1 is a hypervariable minisatellite locus on the human Y chromosome. Its repeat unit is 25 base pairs long and five different variants, which differ from each other by at most 3 substitutions, have been observed. Amplifications and contractions of order 1 and arity 2 were shown to be the most probable events experimentally [3]. This is in agreement with the model. The data set comprises 609 alleles from all over the world, with the corresponding map, as well as the population of origin and the Y-chromosomal genetic group (the technical term is **haplogroup**) if known. All pairwise distances between alleles were computed. The experiments consisted in reconstructing phylogenetic trees with a distance-based method for all or a subset of the alleles using the corresponding distance matrix. The resulting trees are confronted to other experimental data.

First, an evolutionary tree of the haplogroups was obtained from MSY1 with average distances and found to resemble strongly the tree reconstructed from other

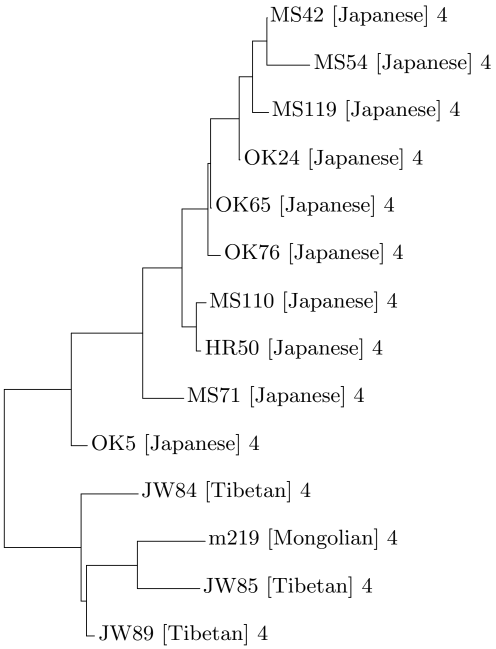


Fig. 12. Phylogenetic tree of haplogroup 4 built from a distance matrix produced by *MS_ALIGN*. Each individual is represented by its code, its population origin and its haplogroup.

less polymorphic markers. However, the MSY1 tree gives a higher resolution. Second, the authors looked at the trees of all alleles from a given haplogroup. In many such trees, alleles are grouped by populations of origin when these are geographically distant. This is consistent with the already observed geographical specificity of the Y chromosome [30]. An example of tree for the fourth haplogroup is given in Figure 12. In the tree of this haplogroup, the Japanese alleles are perfectly separated from the Tibetan and Mongolian alleles.

4. Conclusion

Organisms have the possibility to locally duplicate, triplicate, etc. a segment of their genome, and also to remove one or more copies among adjacent identical segments. This creates repeats with varying numbers of copy units, called tandem repeats. As point mutations also alter the copy units, tandem repeats display variation in copy sequence and in length. It follows that a tandem repeat has a history and that any two individuals may have different tandem repeat sequences, alleles, at the same genomic location. For many reasons, biologists are interested in tracing back the history of a tandem repeat and to compare different alleles of a tandem repeat. In this paper, we surveyed these two problems, TANDEM REPEAT HISTORY and TANDEM REPEAT ALLELE ALIGNMENT, which are related to the evolution of tandem repeats. We gave an overview of algorithmic and combinatorial results on these topics, as well as detailed biological motivations. We provided a unified framework for their formalization. Among the algorithms mentioned in this survey, some have been implemented and are available on the Internet at `http://degas.lirmm.fr/REPSEQ`: *DTScore* and *MS_ALIGN*.

Surely, the solutions for Tandem Repeat History would improve in biological significance, if the algorithms could consider both amplifications and contractions. In order to avoid considering infinite histories, a relevant constraint would be to limit the number of contractions: not all copies created by an amplification can later be removed by a contraction. The Tandem Repeat Allele Alignment problem is exactly solved when amplifications and contractions are limited to order 1. Relaxing these constraints represents future lines of research. Combinatorial properties of tandem repeat histories also require more investigations and could lead to algorithmic improvements. Although probabilistic approaches based on Maximum Likelihood are recognized as the most reliable methods in phylogeny, they were neglected until now for Tandem Repeat History. The design probabilistic methods for this problem also seems a promising direction of research.

Acknowledgements

I would like to thank Franois Nicolas, Sverine Brard, Denis Bertrand, Sylvie Pinloche and Olivier Gascuel for their help and suggestions on the manuscript. This work is supported by: the French Inter-EPST program for Bioinformatics, the Gnopole of Montpellier, the Specific Actions "Algorithms in Biology" and "Algorithms and Sequences" of the STIC section of CNRS.

References

1. Alberto Apostolico, Mikhail J. Atallah, and Susanne E. Hambrusch. New clique and independent set algorithms for circle graphs. *Discrete Applied Mathematics*, pages 1–24, 1992.
2. J. A. Armour, T. Anttinen, C. A. May, E. E. Vega, A. Sajantila, J. R. Kidd, K. K. Kidd, J. Bertranpetit, S. Paabo, and A. J. Jeffreys. Minisatellite diversity supports a recent African origin for modern humans. *Nat Genet*, 13(2):154–60, 1996.
3. R. Andreassen, J. Lundsted, and B. Olaisen. Mutation at minisatellite locus DYF155S1: allele length mutation rate is affected by age of progenitor. *Electrophoresis*, 23(15):2377–83, Aug 2002.
4. Gary Benson and Len Dong. Reconstructing the Duplication History of a Tandem Repeat. In *ISMB*, pages 44–53, Heidelberg, Germany, 1999.
5. G. Benson. Tandem Repeats Finder: a Program to Analyze DNA Sequences. *Nucleic Acid Research*, 27(2):573–80, 1999.
6. J. Buard and A. J. Jeffreys. Big, bad minisatellites. *Nat Genet*, 15(4):327–8, 1997.
7. Sverine Brard and Eric Rivals. Comparison of Minisatellites. In G. Myers, S. Hannenhalli, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proc. of the Sixth Annual International Conference on Computational Molecular Biology*, pages 67–76, Washington DC, USA, 2002. ACM Press.

8. Sverine Brard. *Comparaison de squences rptes en tandem et application la gntique.* PhD thesis, Univ. des Sciences et Technologies du Languedoc, Montpellier, France, 2003.
9. Sverine Brard and Eric Rivals. Comparison of minisatellites. *J Comp Biol*, 10(3-4):357–372, 2003.
10. B. Behzadi and J.-M. Steyaert. An Improved Algorithm for Generalized Comparison of Minisatellites. In Ricardo Baeza-Yates and Maxime Crochemore, editors, *Combinatorial Pattern Matching, 14th Annual Symposium*, volume 2676 of *Lecture Notes in Computer Science*, pages 32–41. Springer, 2003.
11. P. Buneman. A note on metric properties of trees. *J of Combinatorial Theory series A*, 17:48–50, 1974.
12. D. Cohen, I. Chumakov, and J. Weissenbach. A first-generation physical map of the human genome. *Nature*, 366(6456):698–701, Dec 1993.
13. M. Crochemore, C. Hancart, and T. Lecrocq. *Algorithmique du texte.* Vuibert, 2001.
14. O. Delgrange, M. Dauchet, and E. Rivals. Location of Repetitive Regions in Sequences By Optimizing A Compression Method. In Russ Altman, editor, *Proc. of the 4th Pacific Symposium on Biocomputing*, volume 4, Hawai, Jan 4-9 1999.
15. R. G. Downey and M. R. Fellows. *Parameterized Complexity.* Springer, 1999.
16. Olivier Elemento and Olivier Gascuel. An efficient and accurate distance based algorithm to reconstruct tandem duplication trees. In Thomas Lengauer and Hans-Peter Lenhof, editors, *Proceedings of the European Conference on Computational Biology 2002 (ECCB-02)*, volume 18, 2 of *Bioinformatics*, pages 92–99, London, October 6–9 2002. OXFORD University Press.
17. Olivier Elemento and Olivier Gascuel. An exact and polynomial distance-based algorithm to reconstruct single copy tandem duplication trees. In Ricardo Baeza-Yates and Maxime Crochemore, editors, *Combinatorial Pattern Matching, 14th Annual Symposium*, Lecture Notes in Computer Science. Springer, 2003.
18. Olivier Elmento, Olivier Gascuel, and Marie-Paule Lefranc. Reconstruction de l'histoire de duplication de gnes rpts en tandem. In *Actes de JOBIM*, pages 9–11, Toulouse, France, 2001.
19. Olivier Elmento, Olivier Gascuel, and Marie-Paule Lefranc. Reconstructing the duplication history of tandemly repeated genes. *Mol. Biol. Evol.*, 19(3):278–288, 2002.
20. W. M. Fitch. Towards defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.
21. W. M. Fitch. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein A-I. *Genetics*, 86(3):623–44, Jul 1977.
22. O. Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14(7):685–95, Jul 1997.
23. Olivier Gascuel, Michel D. Hendy, Alain Jean-Marie, and Robert McLachlan. The Combinatorics of Tandem Duplication Trees. *Systematic Biology*, 52(1):110–118, 2003.
24. P. Gill, A. J. Jeffreys, and D. J. Werrett. Forensic application of DNA 'fingerprints'. *Nature*, 318(6046):577–9, Dec 1985.
25. Dan Gusfield. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, 1997.
26. J. Hardy and K. Gwinn-Hardy. Genetic classification of primary neurodegenerative disease. *Science*, 282(5391):1075–9, Nov 1998.
27. A. V. Hill and A. J. Jeffreys. Use of minisatellite DNA probes for determination of twin zygosity at birth. *Lancet*, 2(8469-70):1394–5, Dec 1985.

28. M. Hendy and D. Penny. Branch and bound algorithm to determine minimal evolutionary trees. *Mathematical Biosciences*, 59:277–290, 1982.
29. A. J. Jeffreys, M. J. Allen, E. Hagelberg, and A. Sonnberg. Identification of the skeletal remains of josef mengele by dna analysis. *Forensic Sci Int*, 56(1):65–76, 1992.
30. 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). *Hum Mol Genet*, 7(4):643–53, 1998.
31. D. Jaitly, P.E. Kearney, G. Hui, and B. Ma. Methods for reconstructing the history of tandem repeats and their application to the human genome. *J of Computer and System Sciences*, 2002.
32. 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*, 354(6350):204–9, 1991.
33. A. J. Jeffreys, V. Wilson, and S. L. Thein. Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006):67–73, Mar 1985.
34. A. J. Jeffreys, V. Wilson, and S. L. Thein. Individual-specific 'fingerprints' of human DNA. *Nature*, 316(6023):76–9, Jul 1985.
35. R. Kolpakov and G. Kucherov. On maximal repetitions in words. *J of Discrete Algorithms*, 1(1):159–186, 2000.
36. Roman Kolpakov and Gregory Kucherov. Finding approximate repetitions under Hamming distance. In *ESA: Annual European Symposium on Algorithms*, volume 2161 of *Lecture Notes in Computer Science*, pages 170–181. Springer, 2001.
37. Thomas Lengaeur, editor. *Bioinformatics - From Genome to Drugs*, volume II: Applications of *Methods and Principles in Medicinal Chemistry.* Wiley-VCH Verlag, Weinheim, 2002.
38. Wen-Hsiung Li. *Molecular Evolution.* Sinauer Associates, Inc., Sunderland Massachusetts, U.S.A., 1997.
39. M. Lothaire. *Algebraic Combinatorics on Words.* Cambridge University Press, Cambridge, 1999. URL: http://www-igm.univ-mlv.fr/˜berstel/Lothaire/index.html.
40. S. Ohno. *Evolution by Gene Duplication.* Springer Verlag, New York, 1970.
41. Roderick D. M. Page and Eward C. Holmes. *Molecular Evolution: a Phylogenetic Approach.* Blackwell Science Ltd, Osney Mead, Oxford, 1998.
42. É. Rivals, O. Delgrange, J-P. Delahaye M. Dauchet, M-O. Delorme, A. Hnaut, and E. Ollivier. Detection of significant patterns by compression algorithms: the case of Approximate Tandem Repeats in DNA sequences. *Comp. Appl. in Biosciences*, 13(2):131–136, 1997.
43. É. Rivals, M. Dauchet, J-P. Delahaye, and O. Delgrange. Compression and genetic sequences analysis. *Biochimie*, 78(4):315–322, 1996.
44. J. Stoye and D. Gusfield. Simple and Flexible Detection of Contiguous Repeats Using a Suffix Tree. *Theorical Computer Sciences*, 27(1-2):843–856, 2002.
45. David Sankoff and Joseph B. Kruskal, editors. *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison.* CSLI Publications, second edition, 1999.
46. M. F. Sagot and E. W. Myers. Identifying satellites and periodic repetitions in biological sequences. *J Comp Biol*, 5(3):539–53, 1998.
47. Mengxiang Tang, Michael Waterman, and Shibu Yooseph. Zinc Finger Gene Clusters and Tandem Gene Duplication. *J Comp Biol*, 9(2):429–446, 2002.
48. R. D. Wells. Molecular basis of genetic instability of triplet repeats. *J Biol Chem*, 271(6):2875–8, Feb 1996.
49. Lusheng Wang and Dan Gusfield. Improved approximation algorithms for tree alignment. *Journal of Algorithms*, 25(2):255–273, November 1997.

50. Lusheng Wang, Tao Jiang, and Dan Gusfield. A more efficient approximation scheme for tree alignment. *SIAM Journal on Computing*, 30(1):283–299, February 2001.
51. Lusheng Wang, Tao Jiang, and E. L. Lawler. Approximation algorithms for tree alignment with a given phylogeny. *Algorithmica*, 16(3):302–315, September 1996.
52. Louxin Zhang, Bin Ma, and Lusheng Wang. Efficient methods for inferring tandem duplication history. *Lecture Notes in Computer Science*, 2452:97–111, 2002.