JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 5, 2004

Pp. 933–944

The Role of Unequal Crossover in Alpha-Satellite DNA Evolution: A Computational Analysis*

CAN ALKAN,[1] EVAN E. EICHLER,[3] JEFFREY A. BAILEY,[2] S. CENK ŞAHINALP,[4] and ERAY TÜZÜN[3]

ABSTRACT

Human DNA consists of a large number of tandem repeat sequences. Such sequences are usually called satellites, with the primary example being the centromeric alpha-satellite DNA. The basic repeat unit of the alpha-satellite DNA is a 171 bp monomer. Arbitrary monomer pairs usually have considerable sequence divergence (20–40%). However, with the exception of peripheral alpha-satellite DNA, monomers can be grouped into blocks of k-monomers ($4 \leq k \leq 20$) between which the divergence rate is much smaller (e.g., 5%). Perhaps the simplest and best understood mechanism for tandem repeat array evolution is unequal crossover. Although it is possible that alpha-satellite sequences developed as a result of subsequent unequal crossovers only, no formal computational framework seems to have been developed to verify this possibility. In this paper, we develop such a framework and report on experiments which imply that pericentromeric alpha-satellite segments (which are devoid of higher order structure) are evolutionarily distinct from the higher order repeat segments. It is likely that the higher order repeats developed independently in distinct regions of the genome and were carried into their current locations through an unknown mechanism of transposition.

Key words: alpha-satellite DNA, tandem repeat evolution, unequal crossover.

1. INTRODUCTION

A CONSIDERABLE PORTION OF THE HUMAN DNA SEQUENCE consists of tandemly repeated sequences which are generally called satellites. The primary example of satellite sequences is the alpha-satellite DNA which is located in the centromeric regions of human chromosomes. Alpha-satellite sequences are composed of tandemly repeated *monomers*, basic repeat units of size approximately 171 bp. Arbitrary pairs of alpha-satellite monomers usually exhibit considerable sequence divergence (up to 40%). However, it is usually possible to partition the alpha-satellite sequence into blocks of some k monomers ($4 \leq k \leq 20$) between which the sequence divergence is much lower (5% or less) (Willard and Waye, 1987).

[1]Department of EECS, Case Western Reserve University, Cleveland, OH 44106.
[2]Department of Genetics, Case Western Reserve University, Cleveland, OH 44106.
[3]Department of Genome Sciences, University of Washington, Seattle, WA 98195.
[4]Department of Computing Science, Simon Fraser University, Barnaby, BC V5A1S6.
*A preliminary version of this paper appeared in the Proceedings of GIW'02.

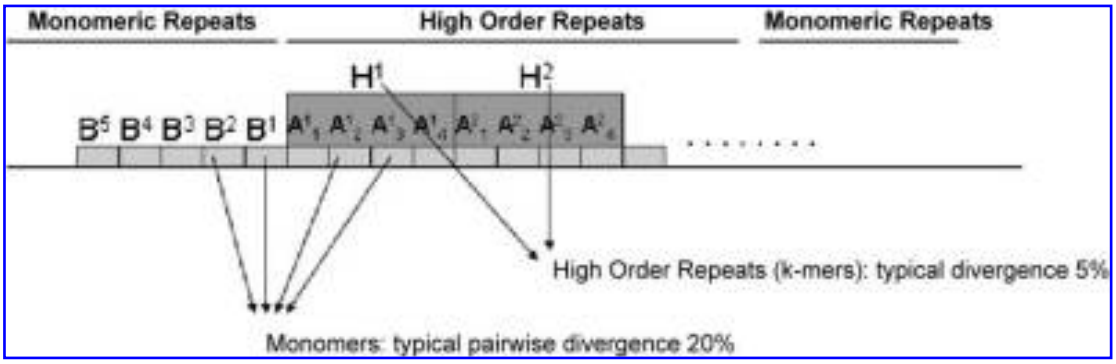

FIG. 1. The composition of human centromeric DNA. Here, *A* and *B* represent monomers, *A* being in higher-order, and *B* in monomeric structure. The higher-order monomers marked with the same subscript are closer to each other (divergence $\leq$ 5%).

In addition to higher order repeats, large tracts of alpha-satellite DNA that are devoid of any higher order repeat structure have been observed recently (Mashkova *et al.*, 1998; Horvath *et al.*, 2000) at the periphery of human centromeric DNA (Schueler *et al.*, 2001; Horvath *et al.*, 2000; Mashkova *et al.*, 1998). These are usually called *monomeric alpha-satellite DNA*; subsequently, the alpha-satellite segments with higher order repeat structure are usually called *higher order alpha-satellite DNA* (see Fig. 1 for the composition of the alpha-satellite DNA).

1.1. Satellite DNA evolution

One possible explanation for the amplification (i.e., replication/duplication) of satellite DNA is through random unequal crossover events, either between sister chromatid pairs during meiosis, or between homologous chromosomes (Charlesworth *et al.*, 1994). The potential role of unequal crossover in tandem DNA array amplification was investigated by Smith (1976) and others; it can be argued that DNA segments that are not maintained by natural selection may acquire short segments which are highly similar at nearby locations due to random mutations, and random unequal crossover events between regions containing such segments will result in deletion or tandem replication of these segments.

Subsequent unequal crossovers between pairs of tandem array blocks either tandemly duplicate or delete an integral number of blocks. The number of duplicated or deleted blocks ℓ is simply equal to the number of unpaired blocks at either end of the tandem array. Typically, duplication events occur with $\ell = 1$, but if a duplication with some $\ell > 1$ blocks occurs, the next duplication or deletion event may have a better chance of involving exactly ℓ blocks again, giving rise to a *higher order* repeat structure.

Although it is possible that some of the higher order repeat segments in the alpha-satellite DNA appear as a result of unequal crossover events (see, for example, Mashkova *et al.* [1998]), recent studies (Alexandrov *et al.*, 2001) suggest the possibility of an unknown mechanism, complementing unequal crossover in this task. More specifically, it is suggested that the higher order alpha-satellite DNA in certain chromosomes emerged elsewhere in the genome and was transposed into the already existing monomeric repeat sequence by an unknown mechanism, overtaking the function of the existing monomeric structure associated with the centromere. After the establishment of higher-order arrays, the monomeric arrays became inactive.

1.2. Summary of contributions

The main goal of our paper is to develop an algorithm to assess whether unequal crossover is solely responsible for the evolution of alpha-satellite DNA or was complemented by an unknown mechanism as suggested by Alexandrov *et al.* (2001). As a first step of our experimental approach, we construct the phylogenetic trees of the monomers which are extracted from sequenced clones from Human Genome Project (HGP) (International Human Genome Sequencing Consortium, 2001) databases that involve large tracts of alpha-satellite DNA (as per Tang *et al.* [2002] and Zhang *et al.* [2002]). Because many of the

monomers are extracted from unfinished draft sequences, we make no assumptions on the relative ordering of the monomers on their respective clone. Therefore, we associate with each monomer only the accession number of the clone it is extracted from. Thus, the phylogenetic trees we build involve tandem arrays which have large gaps whose repeat units are unordered. This limits the applicability of available methods such as those of Tang *et al.* (2002) and Zhang *et al.* (2002) in our studies and necessitates the development of a new approach for interpreting the phylogenetic trees obtained.

We build phylogenetic trees of monomers extracted from two or more clones at a time and try to reach conclusions based on the evolutionary relationships between monomeric and higher order repeats. Many of the phylogenetic trees we built exhibit a strong separation in the evolutionary history of monomers from higher order alpha-satellite DNA and monomers from monomeric tracts as per Alexandrov *et al.* (2001). We also observed that the monomers from different monomeric clones *mix well*.[1]

One of our contributions is a simple probabilistic framework for measuring how *surprising* it is that monomers from arbitrary clones remain evolutionarily distinct (or "mix well"). Within this framework, we obtain exact expressions for the probability of *evolutionary distinctness* for pairs of tandemly generated DNA segments. Our analysis is based on the *restricted tandem repeat history* (RTRH) model used in many other studies of tandem DNA evolution (Benson and Dong, 1999; Tang *et al.*, 2002; Berard and Rivals, 2002; Elemento *et al.*, 2002; Jaitly *et al.*, 2002; Zhang *et al.*, 2002). We thus present a method for checking whether the monomers involved in the phylogenetic studies were derived from each other via tandem duplications prescribed by the RTRH model, practically testing the validity of the model itself. Calculations based on our analysis seem to indicate that unequal crossover is not the only mechanism responsible for alpha-satellite DNA evolution.

2. ALGORITHMIC RECONSTRUCTION OF EVOLUTIONARY HISTORY

Evolution of tandem DNA arrays, especially via unequal crossovers, has recently attracted considerable attention (Benson and Dong, 1999; Tang *et al.*, 2002; Berard and Rivals, 2002; Elemento *et al.*, 2002; Jaitly *et al.*, 2002; Zhang *et al.*, 2002). A significant portion of available literature is devoted to reconstructing the duplication history of the tandem DNA arrays. Introduced by Benson and Dong (1999), the *tandem repeat history* (TRH) problem can be described as follows. Given a tandem DNA array, the TRH problem aims to iteratively contract pairs of subsequent equal-length DNA segments involving one or more repeat units/monomers. The goal is to minimize the total cost of contractions where cost is usually defined in terms of the sequence divergence between contracted copies.

TRH problem involves two key parameters: (1) the length of a duplicated segment (i.e., how many basic repeat units are duplicated in a single unequal crossover event) and (2) the location of the crossover boundary (see Fig. 2). The general TRH problem with no restrictions on these two parameters is NP-complete (Jaitly *et al.*, 2002). A greedy heuristic for this version of the problem is described by Benson and Dong (1999) and also by Tang *et al.* (2002). It is possible to impose restrictions on the TRH problem so that (1) only one repeat unit can be duplicated at a time and (2) the crossover boundaries coincide with repeat unit boundaries; we will call this problem the restricted TRH (RTRH) problem as per Benson and Dong (1999). Unfortunately, the RTRH problem is NP-hard as well (Jaitly *et al.*, 2002) although polynomial time algorithms providing approximate solutions are available (a 2-approximation algorithm is given by Benson and Dong (1999) and polynomial time approximation schemes are given by Jaitly *et al.* (2002) and by Tang *et al.* (2002),

A recent related work (Berard and Rivals, 2002) aims to reconstruct the evolutionary history of *mini-satellite* sequences which are much shorter and less divergent. This work further imposes restrictions on the RTRH problem by assuming that the contraction cost of two subsequent repeat units is independent of their sequence divergence—provided that the two copies are not identical. A smaller cost is assumed for contracting two identical repeat units. If the root sequence of the phylogenetic tree is also specified, a polynomial time ($O(n^4)$) algorithm is available for solving this further restricted version of the RTRH problem (Berard and Rivals, 2002).

[1]Monomers from different clones may occur as siblings in the phylogenetic tree.

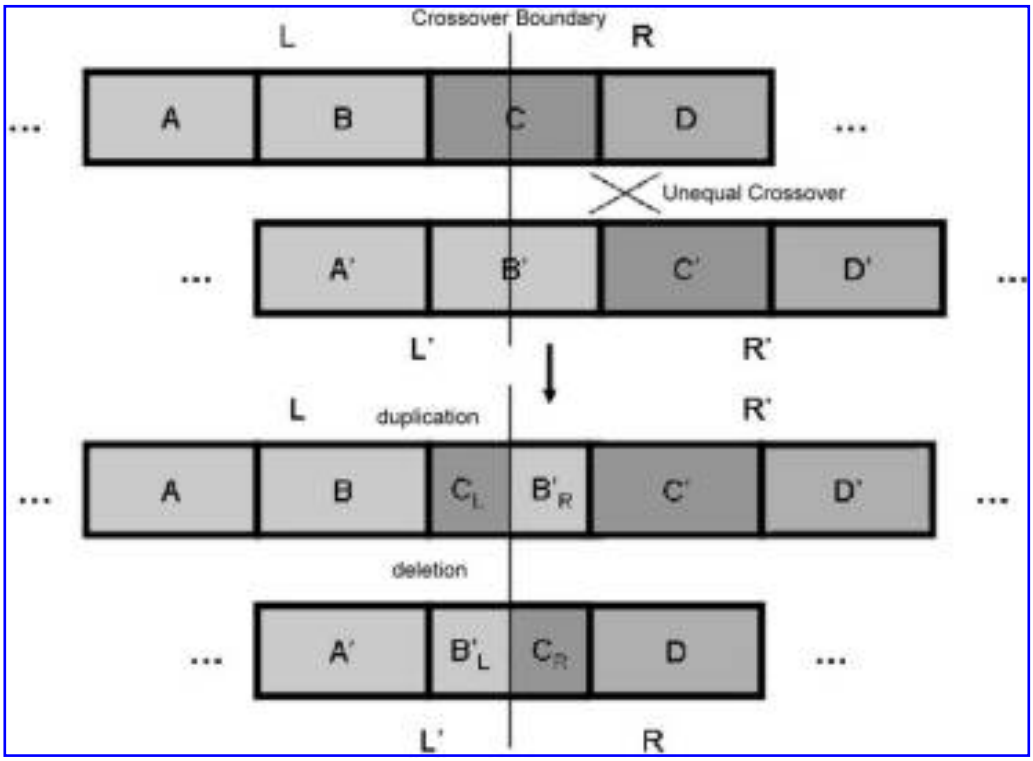

FIG. 2. Unequal crossover leading to tandem repeat structure.

3. PHYLOGENETIC STUDIES OF TANDEM DNA ARRAY EVOLUTION

As explained earlier, our main concern in this paper is not directly inferring the evolutionary history of the alpha-satellite DNA. Methods for solving this problem *assume* that tandem duplications are the only source of tandem DNA evolution. We rather focus on how likely it is that the alpha-satellite sequences evolve solely through unequal crossover events resulting in tandem duplications and deletions. One recent work (Tang *et al.*, 2002) investigates a similar *reconstruction of a duplication model from phylogeny* (RDMP) problem. Here, given as the input (1) the complete sequence of a tandem DNA array and (2) the (binary) phylogenetic tree of its basic repeat units, the goal is to check whether there exists a sequence of tandem duplication events that can produce the complete tandem array from a single repeat unit. A quadratic time algorithm for this problem is given by Tang *et al.* (2002), and an optimal linear time algorithm is given by Zhang *et al.* (2002).

Unfortunately, the nature of the sequence data we use and the specific problem we need to address limits the applicability of the available methods for our purposes. In particular, they usually require the tandem array sequence in question to be *complete and correctly assembled*. Furthermore, the basic question we consider is different. The algorithms of Tang *et al.* (2002) and Zhang *et al.* (2002) aim to find whether it is *possible* that tandem duplication can be the sole mechanism behind the evolution of a tandem DNA array. We would rather like to compute how *probable* it is that the DNA array was produced via tandem duplications only.

One more issue that we had to consider is related to input data reliability. Although heterochromatic regions have not been a target of HGP (International Human Genome Sequencing Consortium, 2001) there is now a large number of sequenced clones that contain large blocks of contiguous alpha-satellite DNA. Unfortunately, many of the clones that contain these monomers are unfinished draft sequences. They contain large gaps and the order of the monomers is suspect. Furthermore, the relative locations of these clones in their respective chromosomes are largely unknown. These issues are not unique to alpha-satellite

sequences; however, input data reliability has never been considered in the context of tandem repeat history reconstruction.[2]

4. PROBABILISTIC TESTING OF THE ROLE OF UNEQUAL CROSSOVER IN TANDEM DNA ARRAY EVOLUTION

As mentioned earlier, our main goal is to derive information about the evolutionary history of the alpha-satellite DNA, whose sequence information is known only indirectly as follows: We are given a collection of monomers whose relative locations in the tandem array are unknown. The only information we have about the monomers is that each of them comes from one of a number of distinct clones extracted from the tandem array and it is known to which clone each monomer belongs.

In the following section, we show that one can derive clues about the evolutionary history of the alpha-satellite DNA sequence if clones from both the higher-order region and the monomeric tracts are considered: Under the assumption that the alpha-satellite sequences were generated by uniformly random tandem duplications (as per the RTRH model), we calculate the probability that the monomers from any two clones have distinct ancestors. Our experiments indeed indicate that monomers from higher-order regions of the alpha-satellite DNA and those from monomeric tracts have distinct ancestors. Within the parameters associated with our experimental setting, we show that the likeliness of this observation is quite low; this leads us to the conclusion that unequal crossover is not exclusively responsible for alpha-satellite sequence evolution.

4.1. Evolutionary distinctness problem

We first describe a probabilistic framework to analyze how likely it is that monomers extracted from two long substrings of a tandem array have distinct ancestors.

Let S and R be two substrings of a tandem repeat array A. It is assumed that (1) A has been generated by restricted tandem duplications as per the RTRH model (i.e., each unequal crossover event results in the duplication of exactly one monomer, and the crossover boundaries coincide with monomer boundaries) and (2) the probability of which monomer gets duplicated (and to which direction) in any event is uniform over all monomers and both directions, independent of previous events. Consider the actual evolutionary tree T of the monomers in A; let the lowest common ancestor (LCA) of all monomers in S be $a(S)$ and the LCA of all monomers in R be $a(R)$. The evolutionary distinctness problem is as follows: given the distance between S and R, what is the probability that $a(S)$ and $a(R)$ are two distinct nodes in T such that none is an ancestor of the other? In other words, what is the probability that the lowest common ancestor of any pair of monomers m_S from S and m_R from R is identical for all pairs m_S, m_R?

We provide an exact expression for evolutionary distinctness of S and R as follows: Let the lengths (in terms of monomers) of S and R be w and v respectively. Assuming that S and R do not overlap in A, let P be the substring of A that stretches between S and R; let k be the length (in terms of monomers) of P. Given any $0 \leq h \leq k$ in P, let $P_h(R)$ be the length h substring of P that is closest to R; similarly, let $P_h(S)$ be the length h substring of P that is closest to S. Clearly, the concatenation of $P_h(S)$ and $P_{k-h}(R)$ gives P itself. During the construction of array A through uniformly random duplication events, let E be the event that $a(S)$ and $a(R)$ are distinct and none is an ancestor of the other; thus, $Pr(E|k, v, w)$ is what we want to evaluate.

Lemma 1. $Pr(E|k, v, w) = \frac{k+1}{k+v+w-1}$.

Proof. In order for E to take place, there must exist an $0 \leq h \leq k$ for which the following event E_h must take place. Given h, let $a_h(S) = a(S, P_h(S))$ be the LCA of all monomers in S and $P_h(S)$, and

[2]For example, the experimental study of Tang *et al.* (2002) examines only a few (16) repeat units with considerable divergence and thus input reliability was not really an issue; the study of Benson and Dong (1999) is performed on synthetic data. Other studies (Jaitly *et al.*, 2002; Zhang *et al.*, 2002) focus on performance evaluation under the implicit assumption that the sequence data is reliable.

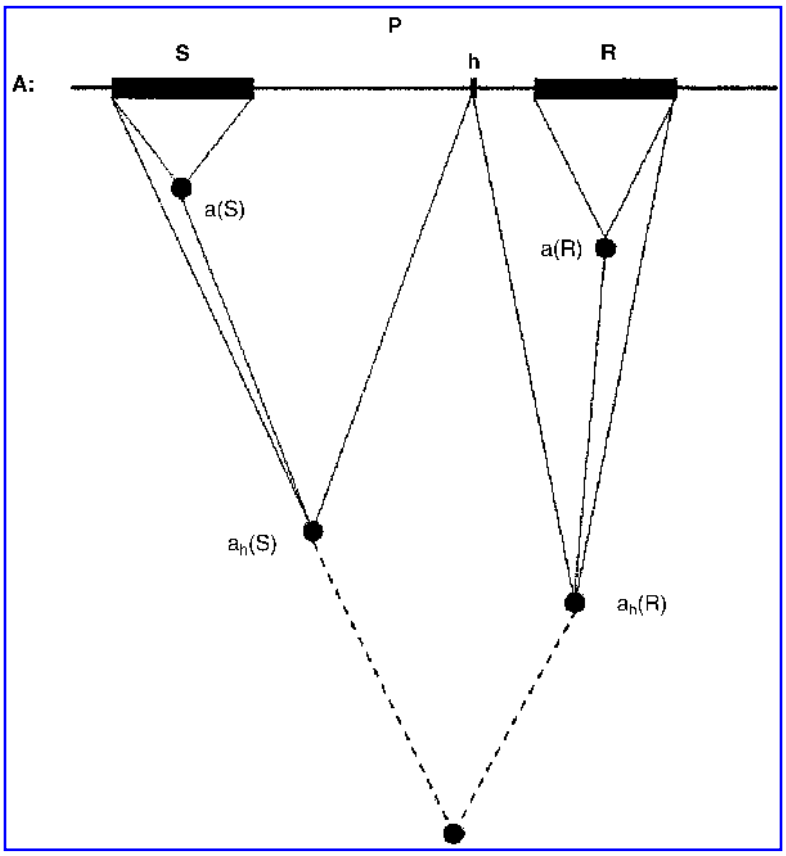

FIG. 3. Distinct evolution of substrings R and S.

$a_h(R) = a(R, P_{k-h}(R))$ be the LCA of all monomers in R and $P_{k-h}(R)$. Event E_h is the event that $a_h(S)$ and $a_h(R)$ are distinct and none is an ancestor of the other (see Fig. 3).

Notice that for any two $h \neq h'$, the events E_h and $E_{h'}$ cannot take place simultaneously. Thus, $Pr(E_h \cap E_{h'} | h \neq h', k, v, w) = 0$. Furthermore, $E = \cup_{0 \leq h \leq k} E_h$; thus, $Pr(E|k, v, w) = \sum_{0 \leq h \leq k} Pr(E_h|k, v, w)$, which reduces our job to the computation of $Pr(E_h|k, v, w)$.

E_h is the event that given $k + v + w$ subsequent monomers, exactly $h + w$ subsequent monomers are descendants of $a_h(S)$ and the remaining $k-h+v$ subsequent monomers are descendants of $a_h(R)$. Consider the reverse random process of building an evolutionary tree out of these $k+v+w$ monomers in iterations. In the first iteration, the probability of any two neighbor monomers being contracted is uniform. Only a contraction between the two boundary monomers between the sets of $h+w$ and $k-h+v$ monomers will contradict with E_h. Similarly, in each subsequent iteration, only one potential contraction will contradict with E_h. Thus,

$$Pr(E_h|k, v, w) = \prod_{i=1}^{i=k+v+w-2} \left(1 - \frac{1}{k+v+w-i}\right)$$

$$= \frac{1}{k+v+w-1}.$$

Because E_h's are independent,

$$Pr(E|k, v, w) = \frac{k+1}{k+v+w-1}.$$

■

Lemma 2. *If k is not fixed but is determined uniformly at random from the range $[0 \ldots (m - w - v)]$, then ignoring some small order ($o(\frac{1}{m-v-w} + \frac{1}{v+w})$) additive terms,*

$$Pr(E|v, w) \cong 1 - \left(\frac{v + w - 2}{m - v - w + 1} \cdot \ln \frac{m - 1}{v + w - 1} \right).$$

Proof. By definition,

$$\begin{aligned} Pr(E|v, w) &= \left(\frac{1}{m - v - w + 1} \right) \cdot \sum_{k=0}^{m-v-w} \frac{k + 1}{k + v + w - 1} \\ &= \left(\frac{1}{m - v - w + 1} \right) \cdot \sum_{k'=v+w-1}^{m-1} \frac{k' - v - w + 2}{k'} \\ &= \left(\frac{1}{m - v - w + 1} \right) \cdot \sum_{k'=v+w-1}^{m-1} \left(1 - \frac{v + w - 2}{k'} \right) \\ &\cong \left(\frac{m - v - w}{m - v - w + 1} \right) - \left(\frac{v + w - 2}{m - v - w + 1} \right) \cdot \ln \frac{m - 1}{v + w - 1} \\ &\cong 1 - \left(\frac{v + w - 2}{m - v - w + 1} \cdot \ln \frac{m - 1}{v + w - 1} \right). \end{aligned}$$

■

Corollary 3. *The probability of observing evolutionary separatedness (of monomeric repeats from higher order repeats) in j out of $l \geq j$ clones each with exactly w monomeric repeats under the assumption that each clone evolved independently is*

$$\begin{aligned} &\binom{l}{j} Pr(E|1, w)^j (1 - Pr(E|1, w))^{l-j} \\ &\cong \binom{l}{j} \left[1 - \left(\frac{1 + w - 2}{m - 1 - w + 1} \cdot \ln \frac{m - 1}{1 + w - 1} \right) \right]^j \\ &\quad \cdot \left[\frac{1 + w - 2}{m - 1 - w + 1} \cdot \ln \frac{m - 1}{1 + w - 1} \right]^{l-j}. \end{aligned}$$

Because the monomeric tracts are assumed to have been completed before the amplification of the higher order tracts, $v = 1$.

5. THE EXPERIMENTS

Our experimental framework involves constructing the phylogenetic trees of monomers extracted from various clones from HGP (International Human Genome Sequencing Consortium, 2001) databases that involve (higher-order and monomeric) alpha-satellite DNA. In this section, we first describe how we obtain the monomer datasets and which datasets we chose to construct the phylogenetic trees. Based on the calculations presented earlier, we then try to compute the probability that random unequal crossover events can result in the evolutionary trees we obtained.

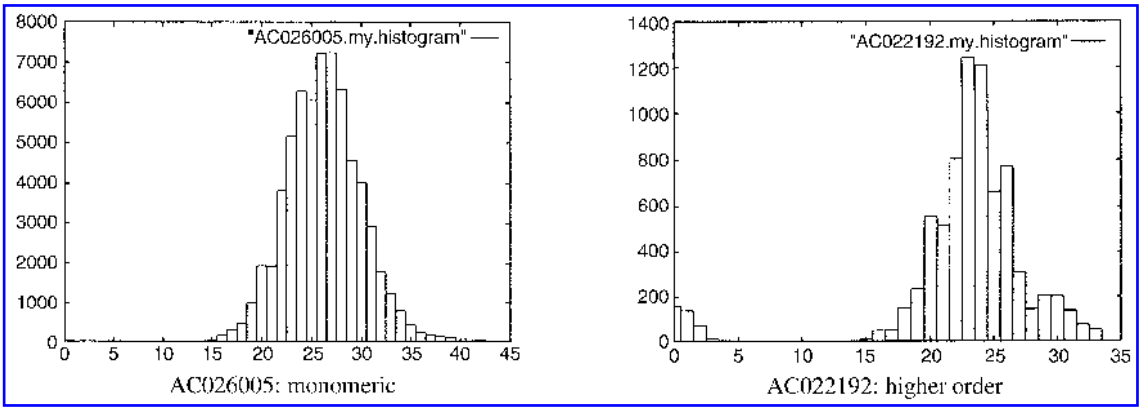

FIG. 4. Classification of clones according to pairwise monomer divergence. The range of percentage divergence between monomers (x-axis) is plotted against the number of monomer pairs with given divergence levels (y-axis).

5.1. Data extraction and classification

As a first step, we established a library of higher-order repeat sequences that have been identified in the literature for each human chromosome (Alexandrov *et al.*, 2001; Lee *et al.*, 1997). These sequences were searched in the entire Human Genome Project (International Human Genome Sequencing Consortium, 2001) sequence data so as to identify large-insert clones that include tracts of higher-order and monomeric alpha-satellite DNA. As a result, a number of large-insert clones were selected whose map location on specific human chromosomes were confirmed to overlap with centromeric DNA.

In the second step we extracted all monomers from each selected clone through the Repeatmasker program (Smit and Green) using the higher-order monomer library.[3]

As a third step, we classified each clone as higher-order or monomeric based on known sequence similarity distributions between monomers in higher-order and monomeric tracts (Alexandrov *et al.*, 2001; Lee *et al.*, 1997; Mashkova *et al.*, 1998). Accordingly, a clone was identified to be *higher order* if each of its monomers were highly similar ($> 95\%$) to at least one other monomer extracted from that clone. We obtained only four such clones. Similarly, a clone was identified to be *monomeric* if its monomers exhibited significant divergence against all other monomers in the same clone ($> 10\%$). Eleven monomeric clones were obtained as a result of our search. (Figure 4 demonstrates typical distributions of pairwise divergence between monomers from both higher-order and monomeric clones.)

5.2. Phylogenetic tree construction

Our primary goal in phylogenetic analysis was to understand the evolutionary relationships between the monomeric tracts and the high-order tracts of alpha-satellite DNA evolution. Because very few monomeric and higher-order clones were extracted from the same chromosome, we applied the following strategy. For each monomeric clone, we built a phylogenetic tree of its monomers together with the monomers from the higher-order repeat library. This library includes a reasonably comprehensive set of monomers from higher-order alpha-satellite DNA. Our premise is that because the percentage divergence between repeats from the same higher-order tandem array is 5% or less, every potential higher-order clone will have a number of representative monomers in the library. To confirm this, we first constructed the phylogenetic trees of monomers from each higher-order clone against those from our high-order repeat library. A sample phylogenetic tree involving monomers from clone AC022192 from chromosome 5 is given in Fig. 5. In all of the trees built, we observed that monomers from the higher-order clone clustered into distinct families

[3]Existing tandem repeat identification and extraction programs such as that of Benson (1999) could not be trivially employed for our purposes. As mentioned earlier, this is due to the fact that many of the clones we analyzed were unfinished draft sequences whose sequence information is not always reliable. Furthermore, there was a need for filtration of common repeat sequences in the clones we extracted.

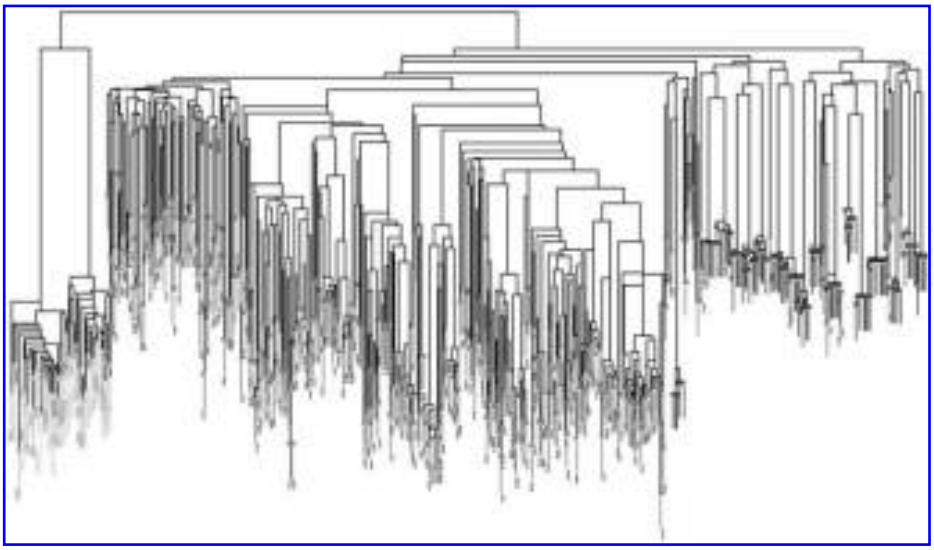

FIG. 5. The phylogenetic analysis of all monomers from higher-order clone AC022192. Color coding: (1) monomers extracted from the clone: *black*; (2) higher order monomer families from our library: *light gray*; (3) Old World monkey sequences: *dark gray* (baboon and macaque). Although the monomers are highly clustered among themselves, they also mix well with some of the monomers from the higher order repeat library.

as expected. Moreover, most of these families appeared to be strongly related to monomers from the repeat library, which seems to indicate that the library is fairly comprehensive.

Our main tests involved construction of a phylogenetic tree for each monomeric clone identified. Each experiment involved all monomers from a given monomeric clone as well as all monomers from the higher-order repeat library. In addition, a sample of monomers from higher-order alpha satellite sequences of *Old World* (macaque and baboon) and *New World* (brown capuchin, titi monkey, etc.) monkey species were added for outgrouping.

The phylogenetic trees are constructed as follows. We used the Clustal-W program (based on the neighbor join (NJ) method) to construct first a *guide tree* (i.e., *dendogram tree*) and then the multiple alignment of the monomers involved. We used the multiple alignment of monomers to reconstruct their phylogenetic tree with a bootstrap support of *100*. In order to do this efficiently, we modified the Clustal-W code (version 1.82) (Thompson *et al.*, 1994) so as to reduce its space usage by a factor of four. On a DELL Pentium-IV PC with 2 Ghz speed and 1 Gbyte memory, the construction of a tree took between 36 hours (for 700 monomers) to 6 days (1,000 monomers). To visualize the phylogenetic trees obtained, we modified the ATV (a tree viewer) program (Zmasek and Eddy, 2001) to color code specified nodes.

One of the phylogenetic trees involving clone AC026005 from chromosome 8 is given in Fig. 6. In the figure, one can observe that the evolutionary relationship between monomers extracted from the respective clone are quite weak suggesting a lack of recent duplication activity. Nevertheless, these relationships are much stronger than those between the monomers extracted from the clone and the higher-order repeats. This suggests the possibility that the higher-order repeats did not emerge from the existing monomeric structure of the respective chromosome, but rather were brought in through some unknown mechanism as suggested by Alexandrov *et al.* (2001). Among the eleven phylogenetic trees we constructed, five of them displayed such characteristics. The division of monomers into higher-order and monomeric clusters was supported by 70%–90% of bootstrap replicates. We also built a phylogenetic tree that involves a large sample of monomers from each of these five clones as well as monomers from the higher-order repeat library and the monkey sequences (see Fig. 7). Note that although these monomers are extracted from five different chromosomes (2,8,9,10,19), they still seem to mix well while staying distant from the higher-order repeats. We observed some level of mixing between monomeric and higher-order repeats in the phylogenetic trees we built for the remaining six monomeric clones.

It is possible to compute the probability that 5 out of 11 monomeric clones stay evolutionarily distinct from the higher-order monomers by using the simple expressions we provided earlier. Although estimates

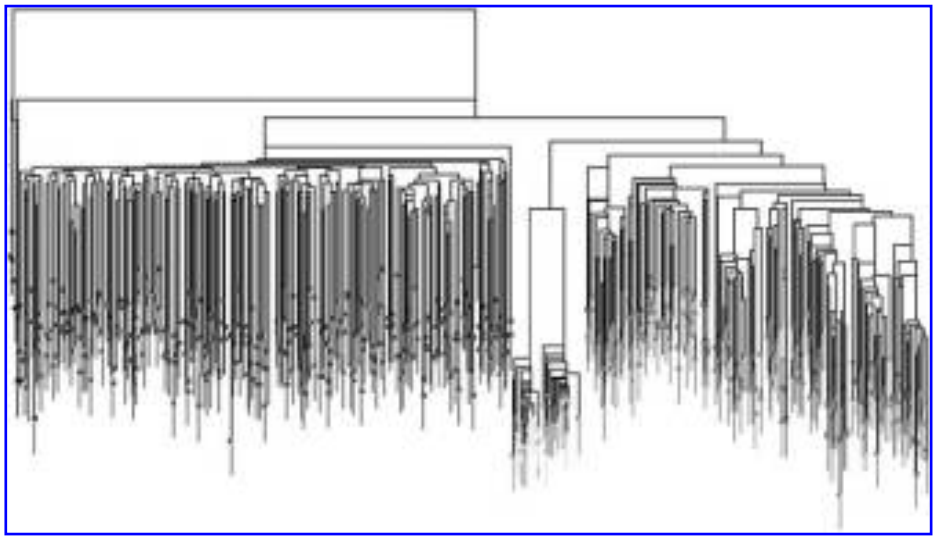

FIG. 6. The phylogenetic analysis of all monomers from another monomeric clone AC026005. Color coding: (1) monomers extracted from clone: *black*; (2) higher order monomer families from our library: *light gray*; (3) monomers from Old World monkey sequences: *dark gray* (baboon and macaque, located in the middle of the tree); and (4) New World monkey monomers: *dark gray*. Once again, the phylogenetic relationship among the monomers from the clone are weak; yet they are still well separated from the monomers from our higher order repeat library.

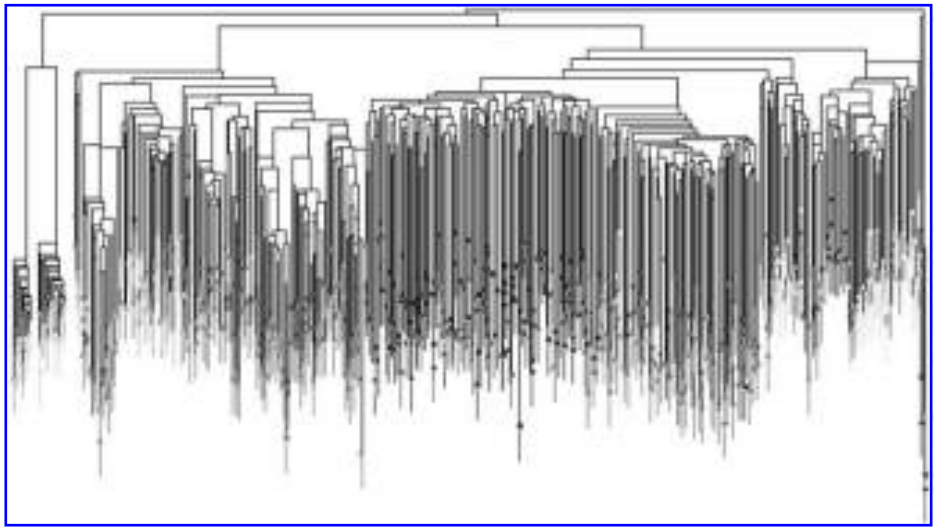

FIG. 7. The phylogenetic analysis of monomers from five monomeric clones. Color coding: (1) monomers extracted from the clones: various tones of *black and dark gray*; (2) higher order monomer families from our library: *light gray*; (3) monomers from Old World monkey sequences: *gray* (baboon and macaque, located on the left end of the tree); and (4) New World monkey monomers: *gray* (located on the right end of the tree). The monomers (colored black and tones of gray) from different clones do mix well; however, they are well separated from the monomers from the higher-order repeat library (with 31% bootstrap support).

on the length of monomeric alpha-satellite DNA vary, 200 Kbp seems to provide an agreeable upper bound. All the monomeric clones used in our experiments were longer than 150 Kbp—which we take to be the size of the clone. Our calculations make the following assumptions: (1) unequal crossover is solely responsible for the evolution of the alpha-satellite DNA, (2) once the higher-order tracts emerged, the monomeric repeat units became inactive, and (3) our higher-order repeat library involves at least one higher-order monomer from the tandem array involving each monomeric clone.

To obtain a relatively generous estimate on the evolutionary distinctness of monomeric tracts and higher-order tracts, we set $m = 200K/171$, $w = 150K/171$ ($v = 1$ under the assumption that higher order tracts were developed only after the completion of the monomeric tracts). According to the expression, we obtained earlier, this gives a probability of

$$1 - \frac{w}{m-w} \ln \frac{m}{w} = 1 - \frac{150}{200-150} \ln \frac{200}{150} = 0.14$$

for a single clone. Out of the 11 clones we tested, 5 of them turned out to be "evolutionarily distinct." Based on the expression obtained earlier, the probability of observing 5 evolutionarily distinct clones out of 11 is thus

$$\binom{11}{5} 0.14^5 (1-0.14)^{11-5} \cong 0.01.$$

This is a very small figure indicating that the distinctness of monomeric and higher-order repeats is very likely not be by chance.

Another interesting observation is that the monomers from different chromosomes *mix well* with each other while remaining distinct from the higher-order monomer set, as in Fig. 7. This clustering further supports the hypothesis that the monomeric tracts and the higher-order tracts have distinct origins and thus unequal crossover may not be the only mechanism responsible for the evolution of alpha-satellite DNA.

CONCLUSION

The phylogenetic trees we obtained seem to support a distinct evolutionary relationship between monomeric and higher-order alpha-satellite DNA. The division between monomeric and higher-order sequence sets in the phylogenetic trees constructed were supported by 70–90% of bootstrap replicates—which suggests no derivative phylogenetic relationship between higher-order and monomer repeat classes. This implies that unequal crossover may not be the sole mechanism determining the composition of large tracts of alpha-satellite sequence within the vicinity of the human centromeric DNA.

ACKNOWLEDGMENTS

This research is supported in part by the Charles B. Wang foundation, Ohio Board of Regents, and an NSF Career Award.

REFERENCES

Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V., and Yurov, Y. 2001. Alpha-satellite DNA of primates: Old and new families. *Chromosoma* 110, 253–266.

Benson, G. 1999. Tandem repeats finder—A program to analyze DNA sequences. *Nucl. Acids Res.* 278, 573–580.

Benson, G., and Dong, L. 1999. Reconstructing the duplication history of a tandem repeat. *Proc. 7th Int. Conf. on Intelligent Systems for Molecular Biology*, 44–53.

Berard, S., and Rivals, E. 2002. Comparison of minisatellites. *J. Comp. Biol.* 10, 357–372.

Charlesworth, B., Sniegowski, P., and Stephan, W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371, 215–220.

Elemento, O., Gascuel, O., and LeFranc, M. 2002. Reconstructing the duplication history of tandemly repeated genes. *Mol. Biol. Evol.* 19, 278–288.

Horvath, J.E., Viggiano, L., Loftus, B.J., Adams, M.D., Archidiacono, N., Rocchi, M., and Eichler, E.E. 2000. Molecular structure and evolution of an alpha satellite/non-satellite junction at 16p11. *Human Mol. Genet.* 9, 113–123.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Jaitly, D., Kearney, P., Lin, G., and Ma, B. 2002. Methods for reconstructing the history of tandem repeats and their application to the human genome. *J. Comp. Sys. Sci.* 65, 494–507.

Lee, C., Wevrick, R., Fisher, R.B., Ferguson-Smith, M.A., and Lin, C.C. 1997. Human centromeric DNAs. *Human Genet.* 100, 291–304.

Mashkova, T., Oparina, N., Alexandrov, I., Zinovieva, O., Marusina, A., Yurov, Y., Lacroix, M., and Kisselev, L. 1998. Unequal crossover is involved in human alpha satellite DNA rearrangements on a border of the satellite domain. *FEBS Lett.* 441, 451–457.

Schueler, M., Higgins, A., Rudd, M., Gustashaw, K., and Willard, H. 2001. Genomic and genetic definition of a functional human centromere. *Science* 294, 109–115.

Smit, A., and Green, P. Repeatmasker. at *www.ftp.genome.washington.edu/RM/RepeatMasker.html*.

Smith, G.P. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science* 191, 528–535.

Tang, M., Waterman, M., and Yooseph, S. 2002. Zinc finger gene clusters and tandem gene duplication. *J. Comp. Biol.* 9, 429–446.

Thompson, J., Higgins, D., and Gibson, T. 1994. Clustal-w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22, 4673–4680.

Willard, H., and Waye, J. 1987. Chromosome-specific subsets of human alpha satellite DNA: Analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* 25, 207–214.

Zhang, L., Ma, B., and Wang, L. 2002. Efficient methods for inferring tandem duplication history. *Proc. Int. Workshop on Algorithms in Bioinformatics*, 97–111.

Zmasek, C., and Eddy, S. 2001. ATV: Display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17, 383–384.

Address correspondence to:
S. Cenk Şahinalp
Department of Computing Science
Simon Fraser University
ASB 9971, 8888 University Drive
Barnaby, BC
Canada V5A1S6

E-mail: cenk@cs.sfu.ca

This article has been cited by:

1. Mary G. Schueler, Beth A. Sullivan. 2006. Structural and Functional Dynamics of Human Centromeric Chromatin. *Annual Review of Genomics and Human Genetics* **7**:1, 301. [CrossRef]