RC 21508 (96944) 21MAY99 Computer Science/Mathematics

Research Report

Tandem Repeat Detection Using Pattern Discovery, with Applications to the Identification of Yeast Satellites

G. Stolovitzky, Y. Gao, A. Floratos, and I. Rigoutsos
IBM T. J. Watson Research Center
P. O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - T. J. Watson - Tokyo - Zurich

Tandem-repeat detection using pattern discovery, with applications to the study of yeast satellites.

G. Stolovitzky[1], Y. Gao, A. Floratos and I.Rigoutsos
Computational Biology Center, IBM T.J. Watson Research Labs,
Yorktown Heights, New York 10598

Abstract

A rigorous definition of Tandem Repeats (TR) is proposed, and a sensitive method is presented for TR detection in genomic data. The heart of our method lays on organizing the information provided by a pattern discovery tool, TEIRESIAS. The latter is a model-less algorithm, and thus our method inherits the feature of not relying on a predetermined model.
Using this technique, we made a systematic study of the TR in the yeast S. Cerevisiae. We studied the distribution of periods and copy-numbers as well as the proportion of TR in coding *vs* non-coding regions. With respect to the later, we found that for periods greater than 10, 60% of the TR lay on coding regions. We also discuss the phenomenon of nesting of TR (which occurs when a TR is contained within another TR locus). Finally we study patterns that are common to the yeast TR.

---

[1] corresponding author; phone: (914) 784-7552; FAX: (914) 784-6307; e-mail: gustavo@us.ibm.com.

1 Introduction

All eukaryotic genomes contain regions that consist of a basic sequence that repeats itself (albeit with slight variations), head to tail, a number of times. These arrangements are called tandem repeats (TR). Both length of the basic sequence that forms the TR, hereafter called the period of the repeat, and the number of repetitions, hereafter called the copy number, vary from TR to TR within the same genome. The overall amount of repeated genomic material is different in different genomes, ranging from approximately 10% in mammalian genomes to 50% in some arthropods [Lewin 1997].

Even though the functional origin of the repeated sequences is still a matter of debate [Charlesworth 1994] (the prevailing conjecture being that short repeats might have a structural function), it is clear that their very presence has important consequences during chromosome pairing, allowing for misalignments to occur, and thus causing repeat clusters to be highly polymorphic [Lewin 1997]. This polymorphism is used in the forensic technique known as DNA fingerprinting [Inman 1997]. Longer repeats tend to have a very complex and hierarchical internal structure, whose details might give hints about DNA molecular mechanisms such as misaligned recombination and saltatory replication. For a more extensive discussion of different biological aspects of TR see [Lewin 1997, Primrose 1995].

Despite the fact that the discovery of these repeating sequences took place about 25 years ago, a detailed characterization of the main parameters of the repeats, such as distribution of lengths, distribution of periods, phylogeny of its hierarchical structure, how often they belong to coding as opposed to noncoding regions, etc, is missing. The reasons for this are twofold. On the one hand only recently enough data to address these questions became available. On the other, and in part due to the first reason, fast algorithms to interrogate the genomes (see below) came into existence only recently.

In this paper we pursue two main goals. The first one is to report on a new algorithm that detects TR in biological sequences. The second one is to characterize in detail, and systematically, the TR in the 16 nuclear chromosomes of the yeast S. Cerevisiae.

2 The Algorithm

2.1 Antecedents

The first attempts to systematically find repeats in sequences dealt with exact repeats [Karp 1972, Milosavljevic 1993]. The constraint of exact repetition was relaxed in [Landau 1993, Kannan 1993, Benson 1994], where approximate repeats appearing exactly twice either contiguously [Landau 1993] or in a nonoverlapping fashion [Kannan 1993, Benson 1994] were considered. An alternative strategy has also been followed. If the length T of the repeating unit is short enough that all words of that length [Rivals 1995] or all the possible ordered k elements in a window of size T ($k < T$) [Guan 1996] can be generated and each used as candidate for the repeating unit, then these candidates can be used as probes and "fitted" to the sequence under interrogation. The same approach can be followed if the repeating unit is given [Fischetti 1992]. In [Benson 1994] and algorithm that detects TR of a given size was proposed.

All of the above algorithms are practical for a specific range of input parameters. For example, some of them are exhaustive algorithms [Kannan 1993, Benson 1994], with time complexity being $O(n^2 polylog(n))$ where n is the length of the full sequence, and thus n cannot be too large. This is a serious constraint if entire chromosomes are interrogated (for example, n is about 1.5 Mbp for chromosome 4 of S. Cerevisiae). For other algorithms the constraint of exact repeats [Karp 1972, Milosavljevic 1993] is untenable for biological sequences of DNA, where mutations, insertions and deletions constitute an important ingredient. TR appear in variable copy numbers, and therefore the algorithms that limit the search to only two repeats are too restrictive [Landau 1993]. In the same vain, the size of the repeating unit can be of a few hundreds, and thus the algorithms that are practical only for short repeats [Rivals 1995, Guan 1996] will be useful only in a restringed range of sizes. The assumption of the repeat [Fischetti 1992] (or its length [Benson 1994]) being given results in limitations when no a priori knowledge about the repeats is available.

More flexible algorithms that find TR of arbitrary period and copy number have recently been reported [Benson 98, Sagot 1998]. In [Sagot 1998], the algorithm exhaustively searches for all possible consensus models whose edit distance to the real repeated units is no bigger than an upper bound; it works by extending prefix models. Benson's algorithm [Benson 98] works by finding k-tuples of consecutive characters from the sequence with identical contents (matching k-tuples) at a common distance. A region containing a sufficiently large number of matching k-tuples is accepted as a TR region if it passes a number of statistical tests based on a probabilistic model.

A cursory comparison of the results reported in [Benson 98] and [Sagot 1998] shows that the sensitivity of both algorithms is different, in that the number of TR reported in both works for the same chromosomes are substantially different. For example, for yeast chromosome 8, no TR model of length greater than 6 was reported in [Sagot 1998], whereas 16 TR of lengths between 10 and 45 were reported in [Benson 98]. (Only TR of size 10 or more were considered in [Benson 98], whereas the repeat sizes covered in [Sagot 1998] ranged between 2 and 45.)

It seems that the heuristic algorithm of Benson is very sensitive, being able to detect TR in a wide range of sizes and copy numbers. In this paper we will explore a variation on Benson's algorithm, that will allow for a considerable improvement in sensitivity, paid for by longer times of run. The idea is to replace Benson's contiguous k-tuples with patterns of not necessarily contiguous

k characters. Because the statistics of matching patterns (with the properties discussed in the next Section) are not as well understood as the statistics of contiguous k-tuples, the statistical criteria used in [Benson 98] are not easily transplanted to our scheme. Thus we use an alternative set of criteria based on our proposed definition of TR. At the heart of our approach resides a fast pattern discovery algorithm, TEIRESIAS, which is described next.

2.2 TEIRESIAS

TERIRESIAS [Rigoutsos 1998a, Rigoutsos 1998b] is a novel combinatorial algorithm for the discovery of rigid motifs contained in a set of input sequences. The motifs, or patterns, discovered by TEIRESIAS are guaranteed to be maximal in both length and composition. In order to understand the problem solved by TEIRESIAS, it is appropriate to introduce some definitions and notation.

Given and alphabet Σ (e.g., $\Sigma = \{A, C, G, T\}$), we define a pattern P as any string of the form $\Sigma(\Sigma \cup \{'.'\})^*\Sigma$. That is, we define P to be any string that starts and ends with a character in Σ and contains an arbitrary combination of "don't care" characters $('.')$ and of characters in the alphabet. Any substring of P that is itself a pattern will be called a *sub-pattern* of P. A pattern P is called an $\langle L, W\rangle$ pattern if every sub-pattern of P with length W or less contains at least L characters. We shall denote by $|P|$ the length of the pattern P, that is the number of characters and don't care characters defining P.

P is really a regular expression and as such it defines a language $G(P)$, whose elements are all the strings that can be obtained from P by substituting each don't care character for an arbitrary character from Σ. A string of characters over Σ is said to *match* a given pattern P if it contains at least one substring that belongs in $G(P)$.

Given a pattern P and a sequence s, we define the offset list of P with respect to s to be the set of integers $L_s(P) = \{i |$ sequence s matches P at offset $i\}$. The cardinal of $L_s(P)$, denoted by $|L_s(P)|$, will be called the support of P by s.

A pattern P' is said to be more specific than a pattern P if P' can be obtained from P by changing one or more don't care characters to actual characters of Σ, or by appending and/or prepending patterns to P. For example, the patterns "AFCH..E", "A.CHL.E.K" and "SA.CH..E" are all more specific than the pattern "A.CH..E". Notice that if a pattern P' is more specific than a pattern P, then the support of P' is smaller than or at most equal to the support of P, i.e., $|L_s(P')| \leq |L_s(P)|$.

A pattern P is called *maximal* with respect to a sequence s if for any pattern P' that is more specific than P the strict inequality $|L_s(P')| < |L_s(P)|$ holds. In other words, if P is maximal, it cannot be made more specific without simultaneously reducing its support.

Using the above definitions, we can pose the problem addressed by TEIRESIAS in the context of this paper: " *Given a sequence s , and parameters L, W*

and K, find all maximal $\langle L, W\rangle$ patterns that have support at least K."

It should be noticed that the statement of the problem solved by TEIRESIAS has been tailored in the above discussion to the purposes of the present paper. For the more general formulation of TEIRESIAS, the interested reader is referred to [Rigoutsos 1998a, Rigoutsos 1998b]

2.3 Definition of Tandem Repeats

Even though there is an intuitive notion of what a TR should be, we still need a formal definition in order to make any progress. Benson [Benson 98] argued that the best definition of a TR is a sequence resulting from a duplication event that occurred in the past, but cautioned about the difficulties of elucidating the history of a sequence. TR must then be judged according to the sequence appearance today.

Our definition of a TR will depend upon an scoring system. It could likewise be defined in term of a metric or a similarity measure. Let the integer T be given, and let $\mathbf{A}(o, T) = A_o A_{o+1} \ldots A_{o+T-1}$ be the subsequence of length T starting at offset o in a longer sequence of length L (typically $L >> T$), for example a chromosome in an organism, a protein sequence, etc. In this notation, the longer sequence will be $\mathbf{A}(1, L)$. The letters A_k are taken from a given finite alphabet. Let $S(\mathbf{A}(o_1, T), \mathbf{A}(o_2, T))$ be the alignment score between $\mathbf{A}(o_1, T)$ and $\mathbf{A}(o_2, T)$). (See Ref. [Waterman 1995] for a complete account of the subject of alignment scores.) Let $P(S; T)$ be the probability density function of the scores between two subsequences of length T taken at random from the sequence $\mathbf{A}(1, L)$. Finally, let $\alpha(p, T)$ (to be called the score threshold) be the value of S such that the probability that S is greater or equal than $\alpha(p, T)$ is equal to p, i.e.,

$$\int_{\alpha(p,T)}^{\infty} dS\, P(S; T) = p \tag{1}$$

We shall say that the array of sequences

$$\mathbf{A}(o, T)\mathbf{A}(o + T, T) \cdots \mathbf{A}(o + (k-1)T, T) \tag{2}$$

is a (p, μ)-tandem repeat of period T and copy number k, with respect to the scoring system S and its statistics $P(S)$, if the following two conditions are verified:

- 1) μ criterion: The average of the scores of the $k-1$ adjacent repeats is greater than μ:

$$\frac{1}{k-1}\sum_{j=0}^{k-2} S(\mathbf{A}(o + jT, T), \mathbf{A}(o + (j+1)T, T)) \geq \mu. \tag{3}$$

- 2) p criterion: There exists an $n \in \{0, 1, \ldots, k-1\}$ such that for all $0 \le j \le k-1$, the alignment scores between $\mathbf{A}(o+nT, T)$ and $\mathbf{A}(o+jT, T)$ is greater or equal than $\alpha(p, T)$:

$$\exists n : \forall j, \quad 0 \le j \le k-1, \quad S(\mathbf{A}(o+nT, T), \mathbf{A}(o+jT, T)) \ge \alpha(p, T) \quad (4)$$

The sequence $\mathbf{A}(o+nT, T)$ in the p-condition will be called a *pivot* of the TR. A (p, μ)-TR can have more than a pivot. Notice that while the first condition is a constraint on the mean of the adjacent scores, the second one is a condition on the variability of the different repeats around a common theme. The pivot (the n-th repeat that the p criterion refers to) can be though of as a valid *model* for the repeats, very much in the spirit of Ref. [Sagot 1998].

Several inclusion relations can be easily derived from the above definition. For example for any $p_0 \le p$, and $\mu_0 \le \mu$, a (p, μ) TR is also a (p_0, μ_0) TR.

For $k = 2$, the p criterion and the μ criterion set the same constraint, and the definition converges to only one condition:

- $S(\mathbf{A}(o, T), \mathbf{A}(o+T, T)) \ge \max\{\alpha(p, T), \mu\}$.

2.4 Detecting potential tandem-repeat regions

The main heuristics we are going to exploit in our algorithm is that the number of equispaced patterns reported by TEIRESIAS is much larger in TR regions than elsewhere in the sequence.

To get a quantitative grasp of this fact, let us first consider the simple example of a sequence over the alphabet $\Sigma = \{A, C, G, T\}$ of length 2,100. The sequence is assumed random (i.e., the letter at an arbitrary location in the sequence is chosen with equal probability within the letters of the alphabet) except that there is a stretch of length 100 in the middle of the sequence that consists of 5 consecutive repetitions of exactly the same tuple of length 20. If we set the TEIRESIAS parameters to $L = 10$ and $W = 10$ (basically contiguous 10-tuples), we expect in the mean 2 patterns from the random part of the sequence ($2 \approx C_2^{2000} p_m (1 - p_m)^{1999}$, where $p_m = (1/4)^{10}$ is the probability that two randomly chosen 10-tuples exactly match). Thus most 10-mers will appear exactly once in the sequence, except, perhaps, for two 10-mers which will appear twice. However, the 20-mer motif will be detected as a pattern and will appear 5 times. Furthermore, the length between subsequent occurrences of that 20-mer will be constant and equal to 20. Thus the TR region will be that region that contains the highest number of equispaced patterns.

There is an obvious caveat to the previous discussion, which can be easily taken care of. It might happen that for a given sequence, one pattern appears more than twice at equispaced offsets. The likelihood of that event is small, and consequently there will be only a few spurious regions (spurious in the sense of not constituting TR). Thus each time a TR-candidate region appears, a verification step has to be taken to assure that we are indeed in the presence of a TR. Such verification step will be done on a small set of candidates, most of which are honest to goodness TR.

More precisely, let the TEIRESIAS parameters L and W and the sequence s be given. The first step is to use TEIRESIAS to find all the $\langle L, W\rangle$ maximal patterns P_j of s with support 2 or more. Let $L_s(P_j) = \{o_1^j, o_2^j, ..., o_n^j\}$ ($n = |L_s(P)| \geq 2$) be the offset list of P_j, that is the list of sites in the original sequence where the pattern P_j matches s. For later use, we denote the differences between consecutive offsets as $T_i^j = o_{i+1}^j - o_i^j$, and by $\Delta_s(P_j)$ the list of differences $\{T_1^j, \ldots, T_{n-1}^j\}$. For each pattern P_j reported by TEIRESIAS, we shall consider as *candidates* for periods all those T_i^j's such that $T_{min} \leq T_i^j \leq T_{max}$, where $T_{min,max}$ are the minimum and maximum periods that we are interested in considering, and are thus input parameters of our algorithms. (In the analysis of the yeast TR to be discussed later on, T_{min} and T_{max} were chosen as 10 and 300 respectively.)

For each T in $[T_{min}, T_{max}]$, we shall call the T-candidate list $\Pi(T)$ to the set of patterns P_j such that there exists at least one i in $\Delta_s(P_j)$ with $T_i^j = T$. It is clear that $\Pi(T)$ could be the empty set. For each T we define a function $h_T(m)$ as

$$h_T(m) = \sum_{P\in\Pi(T)} \sum_{o\in|L_s(P)|} I_{[o,o+|P|-1]}(m) \tag{5}$$

where $I_{[a,b]}(m \in [a,b]) = 1$ and $I_{[a,b]}(m \notin [a,b]) = 0$. In Eq.(5), m ranges from 1 to the length $|s|$ of the sequence s. In words, the value of h_T at location m is equal to the number of patterns in $\Pi(T)$ whose extremes flank the site m. Because the difference list of each pattern will have a variety of T's, each pattern may belong to more than one T-candidate list.

Figure 1 shows $h_T(i)$ for $T = 27$ [Fig.1(a)] and $T = 135$ [Fig.1(b)], for a stretch of $20Kbp$ in yeast chromosome 1. The insets in each figure are enlargements of the highest peak. The parameters chosen for TEIRESIAS were $L = 10$ and $W = 12$. It can be seen in Fig.1(b) that there is a clear peak where the patterns accumulate. The fact that h_{135} reaches a value of 102 means that there are 102 patterns voting for period 135 at that location. The situation depicted in Fig.1(a) is less clear. In the tallest peak there are 5 patterns voting for period 27, which is a "marginal" case. Notice also that the secondary peaks to the right of the main peak in Fig.1(a) are located in the same region where the peak of period 135 in Fig.1(b) is. This is an indication that the TR contain some sort of hierarchical structures: within a region with period 135, there are subregions with period 27 (notice that $135 = 27 \times 5$). However the number of patterns corresponding to period 27 is much smaller than that for period 135.

A third possible behavior for the function $h_T(i)$, not shown in Figs. 1(a) and (b), is that in which not more than one pattern votes for a given period at each location. Typically, those situations do not correspond to TR.

What is the threshold value η such that if $h_T(i) \geq \eta$ for some i we have an appreciable likelihood of having a TR of period T? The answer depends on a number of things: the values L and W that we run TEIRESIAS with, the size of the period T, the number of copies for the TR, and the percentage of mutations between repeating units. The dependence of the threshold η on these parameters is rather complex. Here, we take the pragmatic approach of choosing $\eta(T)$ as a compromise between sensitivity and runtime. The approach we shall take is the following: If for some period T, the maximum value reached by h_T is greater than η, then we proceed to a further verification step, to answer the question of whether or not we are in the presence of a TR that verifies the requirements of the definition. Alternatively, if $\eta(T)$ is bigger than $h_T(i)$ for all i, we do not process that period any further. It is then clear that if $\eta(T)$ is small, we shall interrogate many regions that will end up not containing a TR, and thus waste time. If we take $\eta(T)$ to be big, we shall process less regions (and thus the program will run faster) but we will miss some TR (false negatives). A little of experimentation suggests that $\eta(T) = 10$ is a good choice for $T > 30$. For $T < 30$ we take $\eta(T) = 1$. Only if a period T passes the test $h_T \geq \eta(T)$, do we keep processing that period.

Before going to the verification step, it is convenient to filter the data contained in the function h_T. It would be desirable to have an indicator that is bigger than zero in the region where the TR is presumed to be, and zero elsewhere. However it is clear from the insets of Fig. 1(a) and (b) that there need not be patterns covering the full TR region. See for example the valley in Fig.1(a), where a depletion of patterns occurs. In some cases, those valleys will be deep enough to touch zero, as can be seen in the inset of Fig.1(b). Fortunately there is (at least) a simple way to fix this problem. The idea is simply to convolve h_T with the indicator function with support of exactly the assumed period. More precisely, we shall compute an additional function corresponding to period T as

$$f_T(i) = \sum_{j=i-T+1}^{i} h_T(j). \tag{6}$$

The result of this operation is illustrated in Figs. 1(c) and (d), where everything is the same as in Figs. 1(a) and (b) except that it is f_T instead of h_T what is being plotted. This operation ensures that the indicator function of the regions that are candidates to contain TR (namely the sign of f) is strictly positive. To that effect, compare the insets of Figs. 1(a) and (b) with those of Figs. 1(c) and (d). The curves in the latter are much smoother than in the former, and are never zero within the region of interest.

The next step is to find those regions for which f_T is strictly positive. For each bump of the kind shown in the insets of Figs. 1(c) and (d), we search for the first values i_{min} and i_{max} to the right an to the left of it such that $f_T(i_{min}) = f_T(i_{max}) = 0$. If $i_{max} - i_{min} < 2T$ (i.e., the support of the considered region spans less than two periods), we do not consider that region any further. However, if $i_{max} - i_{min} \geq 2T$, we proceed to a further verification step (see below). This is exhaustively done for all periods T in the T-candidate list for which h_T is bigger than $\eta(T)$.

2.5 Verification Step

Once we have all the regions that can potentially contain TR, we have to discard the false positives, i.e., those regions that do not verify our definition of TR.

To be precise, let us choose the parameters p and μ, required by the definition. Suppose we have k' consecutive sequences of size T $(k \geq 2)$, $\mathbf{A}(o, T)$, $\cdots$, $\mathbf{A}(o + (k-1)T, T)$. If $k' = 2$, we accept only if $S(\mathbf{A}(o,T), \mathbf{A}(o+T,T)) \geq \max\{\alpha(p,T), \mu\}$. If $k' > 2$, we proceed in the following way. After having all the $k'(k'-1)/2$ pairwise alignment scores, we find the pair for which the score is the highest. If this score is below the score threshold $\alpha(p,T)$, we discard that region and proceed to the next. Otherwise we choose from the highest scoring pair the element of the pair that is closest to the center of the repeating region and label it as the *reference* sequence, which is going to be, by construction a pivot of the TR. We then sweep over the remaining sequences to the right and to the left of the reference sequence. *The extremes of the TR region will be the first* $(\mathbf{A}(o,T))$ *and last* $(\mathbf{A}(o+(k-1)T,T))$ *(with* $k \leq k'$*) sequences whose pairwise score with respect to the reference sequence falls above the score threshold* $\alpha(p,T)$. This ensures that the p-criterion of the definition is verified. If the array $\mathbf{A}(o,T) \cdots \mathbf{A}(o+(k-1)T,T)$ also satisfies the μ-criterion, we accept that array as a valid TR.

2.6 Scoring systems

When dealing with concrete applications it is necessary to fix a scoring system, and know its statistics. In this Section we shall study in some detail a family of global alignment scores (in percent units).

For the results to be presented below, the percent similarity score $S(\mathbf{A}, \mathbf{B})$ between sequences $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{A} = A_1 A_2 \ldots A_T$ and $\mathbf{B} = B_1 B_2 \ldots B_T$ is computed as

$$S(\mathbf{A}, \mathbf{B}) = \max\left\{ \frac{1}{T} \sum_{i=1}^{L} s(A_i^*, B_i^*) \ : \text{all alignments} \right\} \tag{7}$$

where the A^*s and B^*s that are not indels '-', are equal to their respective A's and B's, and

$$\begin{aligned} s(A, A) &= 1 \\ s(A, B) &= 0 \quad \text{if } A \neq B \\ s(A, -) = s(-, A) &= a \quad \text{if '-' are not terminal gaps} \end{aligned}$$

$$s(A,-) = s(-,A) \quad = \quad b \qquad \text{if '-' are terminal gaps}$$

The literature of maximum alignment score statistics is rich. A good survey can be found in [Waterman 1995]. When gaps are not allowed, the main result is that the statistics of ungapped scores follows an extreme value distribution [Karlin 1990]. This result holds when the average score for two random symbols to be aligned is negative, and both the length of the aligned sequences and the score tend to infinity. The same assumptions are used for a more recent work [Mott 1999], where the approximate statistics of gapped alignments is considered.

The assumptions described in the discussion of the previous paragraph will not hold in general in our TR alignment statistics. We are thus forced to resort to numerical simulations to tackle the problem at hand. We do this next.

Our aim is to determine a sensible score threshold $\alpha(p,T)$. The statistics of the family of scores considered lines above (below Eq.(7) will depend upon the choice of a and b. We shall consider two particular cases: $a = b = -\infty$ and $a = -1, b = 0$.

1) **Case** $a = b = -\infty$.

This case corresponds to the situation in which the alignment is given in advance and there are no indels [Waterman 1995]. The scores for this choice of function $s(\cdot,\cdot)$ is simply the percentage of matches of the two sequences. The statistics in this case is well known [Waterman 1995], but we shall discuss it as it will become of importance for the next choice of a and b.

If two random DNA sequences of length T (generated as T Bernoulli trials in which each outcome can be any of the four events A, C, G and T with probabilities p_A, p_B, p_C and p_T) were to be compared, the number of matches m would follow a binomial distribution:

$$P(m) = C^T_m \rho^m (1-\rho)^{T-m}, \qquad \rho = p_A^2 + p_C^2 + p_G^2 + p_T^2, \tag{8}$$

where C^T_m is the binomial coefficient. The expected number of matches is then $\langle m \rangle = \rho T$, and the variance is $\langle (m - \langle m \rangle)^2 \rangle = \rho(1-\rho)T$. Thus the mean value and standard deviation of the number of matches per unit base, which for this choice of a and b is actually the result of the alignment score, are given by

$$\mu_T = \langle \frac{m}{T} \rangle = K_1 \tag{9}$$

$$\sigma_T = \left[\langle \left(\frac{m}{T} - \mu_T \right)^2 \rangle \right]^{1/2} = \frac{K_2}{T^{1/2}} \tag{10}$$

where $K_1 = \rho$ and $K_2 = \sqrt{\rho(1-\rho)}$.

If we use the normal approximation to the binomial, we have that the score threshold $\alpha(p,T)$ is in this case

$$\alpha(p,T) = K_1 + \xi(p) \frac{K_2}{T^{1/2}}, \tag{11}$$

where $\xi(p)$ is the inverse of the function (related to the error function)

$$p(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\xi}^{\infty} dx \, \exp\left(-\frac{x^2}{2}\right) . \tag{12}$$

If we take $p = 0.05$, then $\xi(p) = 1.65$.
2) **Case** $a = -1$, $b = 0$.

When we allow for a and b to take finite values, the problem makes a leap in difficulty. This is because in this case the alignment is not given, but determined by optimality. In this case, however, we can use a few known theoretical results.

A corollary of Kingman's theorem (see Theorem 11.3 in [Waterman 1995]) asserts that for finite a and b, Eq. (9) remains valid asymptotically for $T \rightarrow \infty$, with $K_1 > \rho$.Therefore Eq. (9) remains valid at least for T bigger than some value (yet to be determined) for which the asymptotic trend is untenable.

As to the variance, it follows from Steele's theorem (see Theorem 11.6 in [Waterman 1995]) that the second centered moment of the global alignment score is bounded by $K_2'T$ (for some constant K_2'), and therefore the second centered moment of the global alignment score in 'per base' units verifies

$$\sigma_T \leq \frac{K_2'}{T^{1/2}}. \tag{13}$$

This inequality is close enough to, but not quite the same as, Eq. 10. We shall see next that Eq. 10, with an appropriate K_2 holds as an excellent approximation for $a = -1$ and $b = 0$. To do so we shall resort to numerical experiments to fit the values of K_1 and K_2 in Eqs. (9) and (10), for the applications to yeast to be discussed in the second part of this paper.

In order to have realistic values for K_1 and K_2, we performed the following numerical experiment. We took independently two randomly chosen subsequences of length T, with $5 \leq T \leq 500$, from yeast chromosomes 1, 6 and 8. (Actually, these samples were taken from those chromosomes, after masking all the poly dA:dT tracts of length bigger than 10.) For each chromosome we computed the scores given by the global alignment when used to align 50 different pairs, and computed the mean and standard deviation of the scores. The results are shown in Fig. 2, where we also plot (thick solid lines) the ansatz given by Eqs. (9) and (10). The best fitted values for K_1 and K_2 were coincidentally the same for both, and equal to 0.37. Fig. 2 shows that our ansatz for μ_T and σ_T fit the experimental data extremely well.

The distribution of these scores (not shown) for large values of T is Gaussian-like, somehow expected from the Central Limit Theorem. Therefore a sensible score threshold would be the same as in Eq. (11) with the appropriate values for K_1 and K_2.

2.7 Algorithm overview

The pseudocode corresponding to the algorithm described in this Section can be found in Fig. 3. When given a sequence whose TR we wish to detect, we need to provide the following input parameters: L and W (the TEIRESIAS parameters), T_{min} and T_{max} (the minimum and maximum periods we are interested in detecting), $\eta(T)$ (the minimum number of patterns below which we do not bother to explore a region any further), a scoring system, and p and μ, the parameters that determine the similarity between elements of the TR.

The algorithm gets from TEIRESIAS all the maximal $< L, W >$ patterns that appear twice or more and for each T in the interval $[T_{min}, T_{max}]$ whose T-candidate list is non-empty, we compute the function h_T. If h_T gets bigger than $\eta(T)$ at some point, we further compute for that T the function f_T. We explore all the connected regions for which f_T is strictly positive. If the length of this region is smaller than $2T$, we move to the next region. Otherwise, we check whether there is a subsequence contained in that region that qualifies as a (p, μ) TR according to the definition.

At the output of this process, we report the following information for all the valid TR: start and stop location of TR, number of copies, length of the repeating unit (period of the repeat), whether it is in a coding region or not (if that information is available), adjacent-match score, adjacent-indel score, the percentage of total matches (which coincides with the adjacent-match score if the copy number is 2), the consensus length of the clustalw [Thompson 1994] multiple alignment, and the clustalw score for the multiple alignment. Examples are given later on, when we analyze in detail the TR detected in S.cerevisiae.

2.8 Performance

Both the sensitivity and the runtime performance of the algorithm depend on the set of parameters used in the run. Table 1 and 2 show respectively instances of the runtime performance in an IBM $RS6000$ workstation and the number of TR detected by the algorithm, for the following set of parameters: $L = 10$, $10 \leq W \leq 15$, $T_{min} = 10$, $T_{max} = 300$ $\eta(T \leq 30) = 1$, $\eta(T > 30) = 5$ or 10 $p = 0.05$ and $\mu = 0.7$. The sequence under interrogation was the first stretch of $22,000$ nucleotides of yeast chromosome 1.

The runtime (Table 1) will typically depend on the parameters L and W for the TEIRESIAS stage of the algorithm, and on the threshold η that is used to determine which periods are going to be subject to further exploration, in the verification stage. Table 1 shows that for fixed $L = 10$, the runtime roughly doubles for each increment of W in one unit. To get a rough estimate of the run-times for a yeast chromosome of length L one has to multiply the previous estimates times $L/22,000$ (for yeast chromosome 1 this number is 10.5).

Table 2 shows the number of repeats detected for the same range of parameters considered in Table 1. It is clear from Table 2, that not much is gained by increasing W above 12, for there is a saturation of the number of TR. Thus depending on the type of application envisioned, one could sacrifice speed by sensitivity, or vice versa. In the applications to yeast discussed below, a compromise between speed and sensitivity suggested the choice of $W = 12$ and $\eta = 10$.

Table 1. Runtime dependence on parameters.

runtime [sec]	W =10	W =11	W =12	W =13	W =14	W =15
η =10	8.3	16.1	36.2	94.4	223.7	552.4
η =5	9.8	17.6	41.0	132.7	340.0	789.1

Table 2. Sensitivity dependence on parameters.

No. repeats	W =10	W =11	W =12	W =13	W =14	W =15
η =10	9	13	21	25	28	28
η =5	9	13	21	26	30	32

3 Analyses of the Tandem Repeats in yeast

In this Section we shall apply the algorithm described above to make an exhaustive study of the TR in the 16 nuclear chromosomes of yeast [Yeast Genome] and its statistics. All the results to be shown in this section correspond to the output of our algorithm run using the following parameters: $L = 10$, $W = 12$, $\eta(T \leq 30) = 1$, $\eta(T > 30) = 10$, $T_{min} = 10$, $T_{max} = 300$, $p = 0.05$, $\mu = 0.7$. The scoring system is the one shown is Section 2.6, with $a = -1$, $b = 0$.

3.1 Statistics of Tandem Repeats in yeast

In order to have a feeling for the kinds of features arising in the TR of a yeast chromosome, we tabulated all the (p, μ) TR with $p = 0.05$ and $\mu = 70\%$ of yeast chromosome 1 in the table of Appendix A. The first column in that table shows the existence of a nesting of TR of different periods. The "less than" $<$ symbol indicates the first of the TR in a nest. The member of the nest which starts the latest (but does not necessarily end the latest) is indicated by a "greater than" $>$ symbol. In between these extremes, the other members of the nest are indicated by the <> symbol. The second and third columns in Table 3 are the beginning and the end of a TR. The fourth column contains the copy number k of repetitions of the basic unit. In the fifth column we tabulate the period T of the repeat, i.e., the number of nucleotides in the repeating unit. In column 7 it is indicated whether the TR belongs in a coding region (in which case the ORF name is provided), or to a non-coding region. Column 7 and 8 contains the average score of adjacent units and the indel percentage respectively in a TR. Column 9 provides the percentage of global matches (specially relevant for copy number 3 or more). Column 10 gives the consensus length of the multiple alignment produced by Clustalw [Thompson 1994], and the score of this multiple alignment is given in column 11.

The asterisks to the left of the start column in some of the entries correspond to the TR reported by Benson [Benson 1994]. Our algorithm finds all the TR Benson finds and many more, which is an indication of the high sensitivity of our algorithm.

The following two observations can be made from a quick glance at the table in the Appendix:

- Most of the TR have copy number 2.
- Many TR appear in the coding regions.

We shall devote the following Section to a discussion of those two features.

3.2 Most of yeast Tandem Repeats have two copies

In Figure 4 we have plotted the number of TR appearing in each chromosome as a function of the chromosome length. For easy reference, the number of each chromosome appears near the corresponding solid circle.

There are several features to discuss with respect to the table in the Appendix. The first one is that the number of TR is quite large. This is indeed true for all the chromosomes, as is shown in Fig.4. It can be seen there that 90 % of all the TR detected consist of only two copies, whereas only the remaining 10 % contains 3 or more copies. Also, the number of TR, (regardless of the number of copies) tends to increase linearly with the chromosome size. The slopes of these curves give an estimate of the number of TR *per Kbp* expected in yeast. For copy numbers of 3 or more, one has 0.05 TR *per Kbp*, i.e., one could expect one TR of copy number 3 or bigger every 20 Kbp. The TR of copy number exactly 2 appear much more often: once every 2 *Kbp* in the mean.

In order to have a rough estimate of what is to be expected in a non-biological sequence, we run our algorithm over a random DNA-like sequence of base composition similar to that of yeast (30 % of A's and T's and 20% of C's and G's), and of length $1,500,000$ *bp*, similar to that of Chromosome 4 (the longest of yeast chromosomes). The result is plotted in Fig. 4 with a plus symbol. There appeared 246 TR. Except for one of them that had copy number 3, the remaining had copy number 2. The estimate in a random sequence with the same base composition as yeast is of one TR of copy number 2 every 7 Kbp, and about 200 times less for copy number greater than 2. We shall discuss more about the random sequence in due course.

The number of TR discussed above is a partial description of the relative abundance of TR in a genome. The reason for this is that we are not specifying neither the period nor the copy number of the repeats. A more descriptive measure of the relative abundance of TR is the total length occupied by the repeats relative to the total length of each chromosome. The total length of TR includes of course the information of the number of copies and the period size. In the computation of the total length, the nesting of different repeats must be taken into consideration. If several TR are nested (see the first column of the table in Appendix A), the total length to be consider is the difference between the stop of the TR that ends the latest in the nest and the start of the TR that starts the earliest in the nest. In other words, the length of a nest is the size of the maximal segment that encompasses all the TR, and within which, each base belongs to at least one TR of the nest. For example, the nest contained in the ORF YAR044W (see the table in Appendix A) begins at location $193,956$ and ends at $194,030$.

Figure 5 shows the percentage of total chromosome length taken by TR, both within coding regions (plus symbol) and outside coding regions (filled circles). It can be seen that these percentage is approximately constant across chromosomes for the non-coding regions, with a mean of about 0.7 % and standard deviation of 0.2 %. For coding regions the mean is about 1.6 % and a standard deviation of 0.7 %, but there are two conspicuous outliers (chromosome 1 and 9) for which the TR represent over 2.4 % of the total chromosome length. This indicates that in the mean, the length taken by the TR in the coding regions of yeast is about twice as long as that corresponding to the non-coding regions. This is somehow unexpected, given that the conventional wisdom is that TR appear preferentially in non-coding regions. That is not the case for yeast. A graphical portrayal of this feature is depicted in the inset of Fig. 5 where we have plotted the TR length within coding regions versus the TR length outside of coding regions: all the points (except for chromosome 4) fall above the diagonal (dashed line).

We discussed before that most of the TR correspond to a copy number of two, and thus it is sensible to explore the way in which the previous results are altered if TR with copy number equal to 2 are excluded. This is done in Fig. 6, where we plot the percentage of the total length occupied by TR of copy number greater than 2. The results of Fig. 6 are qualitatively similar to those of Fig. 5, except for the expected decrease in the mean percentage of total length taken by the TR in both coding and non-coding regions which now is 0.5 % and 0.1 % respectively. TR of copy number 3 or more tend to occupy more often coding regions than non-coding regions, as shown in the inset of Fig. 6.

3.3 Distribution of copy numbers

We have seen in the previous section that the number of TRs whose copy number is 2, by far outweighs the other copy numbers. In this section we shall take a look at the frequency of occurrences of copy numbers $k \geq 3$.

In Fig. 7 we show the number of occurrences of different copy numbers in different chromosomes. Fig. 7(a) corresponds to coding regions, and Fig. 7(b) corresponds to non-coding regions. Vertical bars of different shades or patterns correspond to different chromosomes. The details of this shade coding is showed in the inset of Fig. 7(a), where chromosomes are sorted in order of increasing length.

Figure 7 shows that small copy numbers occur more often than large copy numbers, but this trend saturates (of course around one occurrence) towards copy numbers in the range of 10. Coding regions contain many more TR with copy numbers larger than ten than non-coding regions do. Indeed the largest copy number observed in yeast corresponds to a TR that repeats 33 times. (This TR has 23 % of indels, and it does not consist of a simple minded repetition of 33 basic units, but to a complex developmental history. Its present structure, however, fulfills our definition of TR. See Section 3.8 for a more detailed account of this particular example). This observation suggests that the high copy number might serve some biological function.

The insets of Fig. 7 show that there seems to be a very weak trend for the number of occurrence of a given copy number to increase with the length of the chromosome. This is true for copy number 3 (depicted in the insets) as well as higher copy numbers.

3.4 Distribution of periods

Figure 8 shows the number of occurrences of different periods and all yeast chromosomes for coding (Fig. 8(a)) and non-coding (Fig. 8(b)) regions. The shade coding for the chromosomes is the same as the one shown in the inset of Fig. 7(a). For improved resolution, the main figures show periods between 10 and 35 (abscissae are in nonlinear scale for better visualization), whereas the insets show periods in the range of 36 to 300.

The number of occurrences of a given period tends to be larger for longer chromosomes. This trend is stronger than that observed in Fig. 7 where the number of occurrences of a given copy number was studied.

The distribution of periods is broad: no period range is particularly excluded, but periods in the range between 10 and 20 appear more often than longer periods. Interestingly, TR of longer periods are more frequent in coding regions than outside of them (please compare the insets of Figs. 8(a) and (b)).

The longest period detected was 288, appearing 4 times in the ORF YBL113C of chromosome 2 and 6 times in the ORF YIR019C of chromosome 9. These TR can be subsumed by repeats of the smaller period 36; still it is a TR of period 288 according to our definition. This kind of ambiguities are expected: a periodic function of period T is also a periodic function of period kT, for any k. When the "signals" being dealt with are strictly periodic, then one chooses the smallest possible period and the ambiguities are resolved. However, when there are slight distortions between periods and the periodicity is not strict (caused in our case by mutations or indels), then the operation of taking the "minimum" period becomes cumbersome: reducing the period in our definition might make the copy number to change, or the range covered by the TR to be modified. Thus our algorithm reports all the periods found. The different uses of the data so organized will dictate whether one keeps only one period or a multitude of them. We shall come back to this discussion in Section 3.8.

An interesting, though somehow expected, observation is that period distribution in coding regions is multi modal, with the modes being at periods that are an integer multiple of three. The absolute maximum in the period range from 10 to 300 studied in this paper is at $T = 12$. No multi-modality is observed in the period distribution outside of coding regions (Fig. 8(b)). A less obvious observation is that there is a non-negligible number of periods that are *not* multiple of three in the coding regions. An example of this is the TR of period 25 appearing in the ORF YGR083C (coding for GCD2, a translation initiation factor) of chromosome 7:

$$TTCTTGTTTCTTTAGTTCTTTTAGC$$
$$TCCTTGTTTGTTAATTTCTTTTCTC$$

However, most of these periods correspond to a copy number of 2, and may very well be in a proportion corresponding to what is expected randomly (see Section 3.6).

3.5 Is there a correlation between copy number and period?

In Fig. 9 we have plotted for each of the detected TR, its copy number *versus* its period. Most of the data points correspond to small copy numbers and small periods. The data becomes sparser as either the copy number or the period is increased.

The overall trend seems to be that the longer the period of a TR, the smaller its copy number. This is certainly related to the fact that TR of small overall length occur more frequently. The total length L of a TR is clearly the period T times the copy number k: $L = kT$. Therefore, the points of Fig. 9 lying on the same hyperbola

$$k = \frac{L}{T} \tag{14}$$

will share the same overall length. The dot-dashed line in Fig. 9 corresponds to $L = 250$. Thus many of the TR in yeast accumulate in the region corresponding to $L < 250$. The dashed curve corresponds to $L = 2,500$. Only one TR spans more than 2,500 nucleotides: it corresponds to the triangle at $T = 192$ and $k = 14$, with a total span of 2688. It is in chromosome 4, in ORF YDR150W (protein NUM1, involved in nuclear migration control).

3.6 Tandem Repeats in a random sequence

We have seen that yeast chromosomes contain a considerable number of TR of different periods and copy numbers. However, it would be desirable to know what to expect for a random sequence whose base composition is similar to that of yeast. We devote this Section to such study.

Let us revisit in slightly greater detail some of the results found so far. This time however, we shall look at a particular chromosome. We choose chromosome 4, for it is the longest and thus we can gather more statistics. Any other chromosome behaves likewise.

Figure 10 shows the period (Fig. 10(a)) and the copy number (Fig. 10(b)) of the TR as a function of their location in chromosome 4. The thick line in both Figs. 10(a) and (b) corresponds to copy numbers equal to 3 or more. The fainter line in both figures corresponds to TR with copy number exactly equal to two.

The first observation to be made with respect to Fig. 10 is that the distribution of both periods and copy numbers along the chromosomes is reasonably homogeneous, both for copy numbers equal to 2 or larger. Notice from Fig. 10(a) that for copy number equal to 2, the periods of the corresponding TRs in chromosome 4 go to rather high values. Also from Fig. 10(b) it is clear that there are quite a few TR with copy number grater than 2. These results are made more quantitative in Fig. 11, where the histograms of periods for all copy numbers (Fig. 11(a)) and for copy numbers greater than 2 (Fig. 11(b)) are plotted. (Obviously, the difference between Fig. 11(a) and Fig. 11(b) would yield the histogram of periods for copy number exactly equal to 2.) We find here features that are similar to those already discussed in the context of the data presented in Fig. 8, namely the multi-modality of the histograms of T, with the modes corresponding to integer multiples of 3, and periods reaching values larger than 250. The histograms of the copy numbers of the TR of chromosome 4 are shown for all copy numbers in Fig. 11(c) and for copy numbers greater than 2 in Fig. 11(d). There were a total of 939 TR detected, of which 853 (the big peak in Fig. 11(c)) correspond to copy number 2. Fig. 11(d) is an enlarged version of the remaining copy numbers ($k > 2$), not resolved in the scale of Fig. 11(c). We have 45 TR that repeat 3 times, 15 that repeat 4 times, and so on. This will be contrasted below with a realization of a random sequence.

Next, we consider a biased random sequence of length 1.5 *Mbp* (approximately the same size as chromosome 4). In this random sequence each letter was independently chosen at each position, from a probability distribution that corresponds to the same frequency of appearance in the yeast genome: A and

T appear each with a frequency of 30 % and C and G appear each with a frequency of 20 %. The results shown in Figs. 12 and 13 are the equivalent to those presented in Figs. 10 and 11 for chromosome 4. The main observations to be drawn from Fig. 12(a) is that no TR of period greater than 27 was found in the random sequence. More importantly, (but somehow expected) of the 214 TR found in the random sequence, all but one had copy number 2, and the one that didn't had copy number 3 (Figs. 12(b) and 13(b)). Fig. 13(a) shows that the histogram of the periods do not show the increase of frequency at periods that are multiple of 3, as was found for the real chromosome 4. The peak of the histogram for the random sequence is at period 12, interestingly the same value found in the real chromosome 4.

The previous discussion pertaining to a random sequence is relevant to a proper interpretation of our results for yeast. We have detected 853 TR of copy number 2 in chromosome 4, whereas for the random sequence we have detected 213 TRs of copy number 2. Because the number of TRs grows linearly with length, that means that we can extrapolate to smaller chromosomes in yeast and state that at least 25 % of the TR with copy number 2 are expected in a random sequence. That is, the mere assembly of a sequence of letters produces a non-negligible number of TR. It is likely that the great heterogeneity found within each yeast chromosome [Li 1998] may have as a byproduct the generation of a greater number of "random" TR than that found in our simple biased random sequence. Indeed it has been shown in [Li 1998] that the heterogeneity is the result of the segregation of regions that are homogeneous in base composition. The base composition in each homogeneous region may be more biased than in the overall chromosome. This effectively reduces the alphabet size from 4 to smaller numbers in each homogeneous region, thus increasing the probability of a random match. In short, many TR of 2 copies can hardly be assigned any biological function or origin.

The previous is a negative result. But there is more than we can learn from a comparison with the random sequence. The first thing is the realization that none of the TR detected in the random sequence had period bigger than 27. In yeast chromosome 4 we have detected 13 TR of copy number 2 and period greater than 30 (the longest period of the TR with copy number 2 is 216). These TRs might have some biological cause or effect. On the other hand, only one TR of copy number 3, and none of larger copy number, was found in the random sequence. In chromosome 4, we detected 45 TR with copy number 3, and 41 TRs with copy number greater than 3. Thus these TR might have some biological *raison d'être*.

One might thus enunciate the following conservative rule of thumb: In yeast, a TR is statistically significant if its copy number is strictly greater than 2, or if its copy number is 2 and its period is bigger than 30. While a more detailed statistical analysis would be desirable, the previous estimates are robust and provide a basis for the interpretation of the results.

3.7 Base Composition of Tandem Repeats

In order to address whether the 75% excess of TR with copy number equal to 2 originate from a very biased base composition, we have explored the base composition within the TR. This exercise is useful also as a point of comparison with other genomes. It might be that among the many mechanisms conjectured to create the TR, base composition plays a role. In Fig. 14 we show the base composition of all the TR with copy number of 2 (Fig.14 (a)) and with copy number greater than 3 (Fig. 14(b)), for all the nuclear yeast chromosomes. Figure 14(a) also contains the data (open symbols in the figure) corresponding to the base composition of the TR with copy number 2 corresponding to the random chromosome studied in the previous section. No noticeable difference appears to exist between the base composition in TR for chromosome 4 and the corresponding random sequence. Also, all the TR of copy number 2 seem to keep the same overall base composition in all the chromosomes. This base composition is more biased than that corresponding to the full yeast genome (dot-dashed lines). This is roughly in accord with the idea that random TR "condensate" in regions where the base composition is more biased.

The TR with copy numbers greater than 2 have a more erratic behavior. There is a weak trend for the frequency of A and T to be larger than that of C and G, but if this trend exists at all, it is very faint to make a story out of it. In any case, these frequencies exhibit large fluctuations due to lack of enough statistics: we do not detect as many TR of copy number greater than 2, as we do of copy number equal to 2, for which the statistics is better converged.

3.8 Nesting of Tandem Repeats

We have mentioned the phenomenon of nesting of TR: the situation in which two or more TR overlap. In this Section we shall analyze in detail a few instances of this phenomenon.

Our first example will be taken from chromosome 1, for which the whole list of detected TR is given in the Table of Appendix A.

Table 3. Example of simple nesting of tandem repeats.

nest	start	stop	k	T	coding	% score	% indels	% global matches
<	100356	100404	2	24	YAL025C	75	0	75
<>	*100359	100413	3	18	YAL025C	74	0	55
>	100371	100395	2	12	YAL025C	83	0	83

This nest starts at location 100,356 (the start of the TR with period 24 of the nest), and ends at location 100,413, which is the end of the TR with period 18. This DNA stretch belongs to the ORF YAL025C, whose role might be related to the apparatus concerned with the nuclear events of the cell cycle [Bussey 1995]. Notice that the span of the TR with period 12 is completely embedded within the TR of period 24. This might indicate that the latter can be subsumed by the former, which indeed is the case: The TR with period 24 and two copies

$$TCACTATCGCTGTCGCTCTCGCTT$$
$$TCACTGTCGCTTTCGCTTTGACTG$$

can be interpreted as a TR with period 12 and 4 copies:

$$TCACTATCGCTG$$
$$TCGCTCTCGCTT$$
$$TCACTGTCGCTT$$
$$TCGCTTTGACTG$$

Our algorithm failed in picking the full extent of the number of copies corresponding to period 12 (it only picked 2 instead of 4), because the parameters used in TEIRESIAS, $L = 10$ and $W = 12$ were not permissive enough. However, if we increase W to 13, then exactly the previous TR appears. In any case, one could surmise from the nesting in the output, with the algorithm run with the original parameters $L = 10$ and $W = 12$ that the period 24 could be subsumed by a period 12 with more copy numbers.

What about the nested period 18. Again, one can make a very reasonable hypothesis that if the periods 24, 18 and 12 are there, then the data in that region can be subsumed by a more parsimonious TR of period 6. (We have not detected it, because we run our algorithm for periods between 10 to 300.) Running our algorithm we indeed find that the 8 copies ($p = 0.05$,$\mu = 0.7$) TR of period 6 extends from locations 100,355 to 100,403:

$$ATCACT$$
$$ATCGCT$$
$$GTCGCT$$
$$CTCGCT$$
$$TTCACT$$
$$GTCGCT$$
$$TTCGCT$$
$$TTGACT$$

Why is the remaining part or the TR with period 18 not subsumed by period 6. The corresponding sequence $GGCGCTGGAA$ extending from locations $100,404$ to $100,413$, has an hexanucleotide suffix $GGCGCT$ which does not verify the second p condition of the definition of the TR. However, it does qualify within the context of the TR of period 18.

These observations might give hints towards possible histories for the formation of this TR, but such goal is beyond the scope of this paper. From a pragmatic point of view, the nesting of TRs might suggest the existence of other periods not detected by the algorithm, or of longer copy numbers of periods detected by the algorithm. In either case, when that information is of the essence, it is desirable to postprocess those nested regions in greater detail. Let us give another illustrating example.

Table 4. Example of a complex nesting of tandem repeats.

nest	start	stop	k	T	coding	% score	% indels	% global matches
<	390588	390984	3	132	YIR019C	73	0	48
<>	390748	390796	2	24	YIR019C	70	0	70
<>	390756	392736	22	90	YIR019C	73	15	2
<>	390763	390799	3	12	YIR019C	70	0	50
<>	390888	392778	15	126	YIR019C	70	11	11
<>	390931	392812	11	171	YIR019C	79	10	25
<>	390978	392733	13	135	YIR019C	73	19	24
<>	390989	391019	2	15	YIR019C	73	0	73
<>	390990	392790	8	225	YIR019C	76	15	34
<>	390995	392777	22	81	YIR019C	77	12	11
<>	391006	392950	9	216	YIR019C	74	13	23
<>	391106	392834	6	288	YIR019C	75	16	40
<>	391116	391146	2	15	YIR019C	73	0	73
<>	391271	392756	33	45	YIR019C	71	23	0
<>	391845	391875	2	15	YIR019C	73	0	73
<>	391890	391920	2	15	YIR019C	73	0	73
<>	391971	392001	2	15	YIR019C	73	0	73
<>	392229	392589	10	36	YIR019C	87	7	31
<>	392866	393046	6	30	YIR019C	81	3	36
>	392940	393048	3	36	YIR019C	75	25	50

The nest shown in Table 4, corresponding to ORF YIR019C (a glucoamylase protein) of chromosome 9 [Barrel 1994], spans from location 390,588 to 393,048. A cursory look at that table shows a variety of nested periods (126, 171, 135, 81, 216, etc), all of which satisfy our definition of TR. The picture is a little puzzling: how is it possible for a TR with period 171 to be nested with one with period 126, almost through the entire span of both TRs. To address this question we analyzed the sequence in that region in detail. We run TEIRESIAS on the subsequence spanning the full nest, and found, among others, two different sequences appearing a large number of times. In regular expression format, those sequences where:

$$\begin{aligned}\alpha = \ & [AG]CTGGAG[CT]AGAAGA[GA]CTTTC[\\ & AG]GT[AG]GT[AG]GAGCT[TG]G[AT]T\end{aligned}$$

of length $|\alpha| = 36$, and

$$\beta = GG[AG]GTTGG[G]$$

of length $|\beta| = 9$. Going back to the sequence, those two patterns allowed us to organize the substring spanning from location 390,996 to 392,776 in the following way:

$\cdots AAAGCTTTCAGTAGTAGAGCTGAAT$	$GGAATTGAA$	Σ
$ACTGGTGCTACAGAGCTTTCAGTGGTGGAGCTGGAT$		
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTTGAT$	$GGGGTTGGT$	
$ACTGGAACAGAAGAGCTTTCAGTGCTAGAGCTGAAT$	$GGGGTTGAA$	Σ
$ACTGGTGCTACAGAGCTTTCAGTAGTAGAGCTTGAT$	$GGGGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCGGTAGTAGAGCTGGAT$	$GGAGTTGGC$	
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGAGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT$		
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGAGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGAGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGAGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT$		
$ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT$		
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGAGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGGGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT$	$GGAGTTGGT$	
$ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT$		
$ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTTGAT$	$GGGGTTGGA$	

$$\begin{array}{lll}
GCTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTTGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTTGAT & GGAGTTGGC & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTGGAGCTGGTT & & \\
ACTGGAGTAGAAGAGCTTTCAGTAGTAGAGCTGGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTGGAGCTTGAT & GGGGTTGGA & \\
GCTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTTGAT & GGGGTTGGT & \\
ACTGGAGCAGAAGAGCTTTCAGTGGTAGAGCTGGTT & & \\
ACTGGAGCAGAAGAGCTTTCAGTAGTAGAGCTGGAT & GGAGTTGGT & \\
ACTGGAGCAGAAGAACTTTCAGTAGTAGAGCTTGAT & GGGGTTGGT & \\
ACTGGAGTAGTAGTCTTCTTAGTGCATGTCTTAGAT & GTGGTAGTT & \cdots
\end{array}$$

where

$\Sigma = GATGGAGCGGAGGAAGTGATGTTGCTAGAGGAAGATGGGGTTGGT$

and $|\Sigma| = 45$.

In a strict sense, the previous sequence can not be regarded as periodic, because there is no unambiguous period. Indeed the basic unit could be α, in which case the period is $T_\alpha = 36$, or the basic unit could be $\alpha\beta$, in which case the period is $T_{\alpha\beta} = 45$. All the periods in the previous table that correspond to TR in the substring spanning from $390,996$ to $392,776$ can be explained in terms of T_α and $T_{\alpha\beta}$:

$$\begin{aligned}
36 &= 1T_\alpha + 0T_{\alpha\beta} \\
45 &= 0T_\alpha + 1T_{\alpha\beta} \\
81 &= 1T_\alpha + 1T_{\alpha\beta} \\
126 &= 1T_\alpha + 2T_{\alpha\beta}
\end{aligned}$$

$$\begin{aligned}135 &= 0T_{\alpha} + 3T_{\alpha\beta}\\171 &= 1T_{\alpha} + 3T_{\alpha\beta}\\216 &= 1T_{\alpha} + 4T_{\alpha\beta}\\225 &= 0T_{\alpha} + 5T_{\alpha\beta}\\288 &= 3T_{\alpha} + 4T_{\alpha\beta}\end{aligned}$$

This exercise shows that in many cases the nesting is an artifact: the basic structure that is repeating is a composite of two or more intertwined units. This hints at the fact that the definition of TR could be generalized to include these cases. However, we would like to emphasize the we discovered this nucleotide arrangement using the intuitive definition of TR, and paying special attention at the nesting structure. So that intuitive definition can serve us well in spite of the fact that it is not the most general one. (See the Discussion Section.)

The existence of a rather complex, and yet quasi-regular sequence hints at mechanisms that produced it. Was it replication slippage? Or is there a signal in α that prompts the smaller sequence β to get inserted? The analysis of this sort of structures in the nucleotide sequence may lead to good hypotheses that could eventually be tested in the wet lab.

3.9 Is there a common pattern shared by Tandem Repeats?

Once all the detected TR are collected, an interesting question arises. The TR can be "independent", or correlated in an as yet unspecified manner. One way of addressing this question is to use a pattern discovery tool to find the patterns shared by different TR in a given chromosome. Because of the presumed different nature of TR with copy number $k = 2$ and copy number $k > 2$, we divided the search in those two categories. Table 5 shows the result of such computation for $k = 2$. For each chromosome, (column 1) including the random one discussed in Section 3.6, we tabulated the pattern (column 2) which appeared in the largest number of TRs (column 3). The total number of TR detected for each chromosome appears in column 4, whereas in column 5 we tabulated the total number of occurrences of that pattern in the TRs (which may include multiple occurrences in any given TR). There is a clear emerging picture: the highest ranking patterns for the yeast chromosomes contain only either A's or T's. Notice that in the random sequence, the highest ranking pattern contains both A's and T's. From Fig. 14(a) we see that the frequency of appearance of A's and T's in the TRs with $k = 2$ is similar and around 35 % both for the yeast chromosomes and the random sequence. Therefore, if the patterns detected in yeast were completely random, one would expect that each of these patterns would contain a comparable composition of A's and T's. However, without exception, the highest ranking patterns break this symmetry (a symmetry that is indeed verified in the TR of the random sequence) and contain only one of the two expected nucleotides.

The number of TR in which these patterns appear (column 2) is about 20 % of the total number of TR detected in each yeast chromosome. This number is 14 % for the random sequence. We have discussed in Section 3.6 that we expect that a fraction of the yeast TR with $k = 2$ is expected merely on a random basis. Those TR should have the same characteristics as the TR of a random sequence. However, the remaining fraction might have its own features. We think that this breakage of symmetry is one of those features, intrinsic to the TR generated through some biological mechanism, yet to be elucidated.

Another observation is related with an expected symmetry, which is indeed observed in Table 5. If the mechanism responsible for the creation of TR is as likely to occur in the Watson strand as in the Crick strand, then one should expect that if a pattern is abundant, its complementary will be too. In broad strokes, that is observed in Table 5: Eight of the sixteen chromosomes have patterns that contain only A's. The remaining eight, contain only T's. In some cases, the complementarity is exact: the patterns of chromosome 4 and 16 are the exact complementary of the patterns of chromosome 7 and 15.

The situation pertaining to the TRs with $k > 2$ is not as clear, even though we shall see that some regularities emerge. The organization of Tables 6 and 7 is the same as that of Table 5. In Table 6 we show the patterns that appear the largest number of times in different TR, whereas in Table 7 the patterns that appeared the most often in total are tabulated. Patterns in both Table 6 and 7 appear typically in about 50 % of the TRs (the ratio in percent units between column 3 and 4), which is a considerable fraction. In other words, about 50 % of the TR detected in each chromosome share the same kind of patterns.

These patterns have interesting regularities. All contain at most 2 different nucleotides. Except for the patterns that contain only one nucleotide (all T's, as in the case of TR with $k = 2$, and all C's), if there is a pattern appearing in Table 6 and 7 that contain two bases, then patterns containing the two complementary bases are also in the Tables. More concretely, we have patterns containing only A's and G's, and patterns containing the complementary bases T's and C's, Likewise with A and C on the one hand and T and G on the other. Finally, we have the patterns with only A's and T's, which are self complementary.

As we saw in the case of TRs with $k = 2$, the symmetry between Watson and Crick strands implies that if there is a pattern on a TR in one strand, and it is mechanistically important, then the complementary pattern is likely to occur in another TR on the same strand. But should we expect the complementary pattern on the same TR? This sometimes occurs, and in this case a possible mechanism of generation of TR can be conjectured. But we shall discuss this last issue in a forthcoming publication.

Table 5. Patterns in TR with $k = 2$ that ranked highest in appearances.

Chr	pattern	No of different TR in which pattern occurs	Total No. of TR ($k = 2$)	Total No of occ. of pattern
1	AA.A..AA	41	146	117
2	T..TT.TT	87	402	249
3	T.TT..T.T	43	186	93
4	AA.AA.A	171	853	469
5	T..TT.TT	79	335	243
6	A.AA.AA	33	144	70
7	T.TT.TT	134	604	345
8	AA.A..A..A	62	330	143
9	AA.A..A..A	54	250	141
10	A.A...AAA	60	321	111
11	T..T.TT..T	91	386	329
12	AA.A...A.A	116	586	268
13	A..A..AA.A	103	586	290
14	TT.TT..T	81	404	218
15	T.TT....TT	114	564	278
16	AA....AA.A	88	431	216
rndm	TTA....A.A	30	213	37

Table 6. Patterns in TR with $k > 2$ that ranked highest in appearances.

Chr	pattern	No of different TR in which pattern occurs	Total No. of TR ($k > 2$)	Total No of occ. of pattern
1	TT.TT..T	13	26	48
2	AG.A...G.A	25	52	106
3	A...A.A.AC	16	32	136
4	T.GT..T..T	35	86	166
5	TA.T...A.T	12	22	144
6	A..A..AT.A	14	25	50
7	T.A...T.AT	27	66	117
8	T.G...T.GT	20	51	90
9	G.A..AG.A	27	59	456
10	G..G..G.TG	16	37	182
11	G..G..G.TG	19	54	120
12	A..A..AT.A	25	71	165
13	T.TT....TT	30	61	226
14	TC...TT..T	16	49	44
15	CA.CA.C	27	70	126
16	TT..T..T.T	23	48	188

Table 7. Patterns in TR with $k > 2$ that ranked highest in total appearances.

Chr	pattern	No of different TR in which pattern occurs	Total No. of TR ($k > 2$)	Total No of occ. of pattern
1	TG..G.T.T	10	26	114
2	T.G...T.GT	23	52	392
3	C.C.C.C.C	13	32	739
4	T.AT..T..T	33	86	251
5	TA.T...A.T	12	22	144
6	TA.T...A.T	10	25	88
7	T..T.AT..T	25	66	144
8	C..C.TC..C	16	51	235
9	G.AG.AG	22	59	971
10	G..G..GT.G	12	37	549
11	T.CT..T..T	17	54	317
12	T.AT..T..T	24	71	177
13	T.TT....TT	30	61	226
14	TC..C.TC	12	49	178
15	C..CA.C..C	25	70	208
16	T..T..T.TT	21	48	218

4 Summary and Discussion

The availability of complete genome sequences has allowed in the past few years to develop a feeling for the main features that characterize genomes. One of these features is the structure of repeating sequences. In this paper we have described an algorithm which has a remarkable sensitivity to detect TR. A first step in that direction had to be, by necessity, an unambiguous definition of TR. We advanced a definition that is based on the basic intuition of what we expect TR to be. Armed with that definition, and the algorithm to find those structures, we analyzed in detail the TR in the 16 nuclear chromosomes of S. Cerevisiae.

For yeast, we found a number of novel results. Most (90 %) of the yeast TR have copy number equal to 2. The remaining 10 % of the TR have copy number greater than 2. By way of a control, we compared our results with a random sequence with similar base composition. Interestingly enough, a surprisingly high number of TR nucleate out of pure chance in that random sequence. Comparing the results for the random sequence with the equivalent results for yeast, we came to the following rule of thumb to determine some sort of significance to the TR detected in biological sequences: we (typically) do not expect TR of copy number greater than 2 in a random sequence. For copy number equal to 2, we expect that the periods are smaller than say, 30 nucleotides.

We study the distribution of periods (restricted to those between 10 and 300 nucleotides) and copy numbers. We showed that the yeast genome contains TR with copy numbers larger than 30, and periods larger than 280.

More importantly, we found that a large proportion (about 60 %) of the TR in yeast with periods between 10 and 300 are within ORFs. This observation goes in a way contrary to the conventional wisdom that TR are located mainly in non-coding regions. This intuition might still hold for micro satellites (TR with periods smaller than 10), but definitely does not apply to yeast for periods of 10 and longer. Indeed, the longest of the TR are located within coding regions.

It has been observed in the literature [Lewin 1997] that TR have a hierarchical organization, presumably as a product of the mechanism through which they were produced. In our results that hierarchy shows up in the way of a nesting of TR in particular regions. Thus, the phenomenon of nesting in the output of our algorithm is indicative of a hierarchical structure within a TR.

However, there are other circumstances in which nesting occurs. As we showed in an example in Section 3.8, there are TR that depart from the norm, in the sense that they have more than one repeating unit. Suppose the repeating unit is α. The usual way of thinking of TR, is of a stretch of sequence of the form $\alpha\alpha\cdots\alpha$. Our definition of TR was indeed based on this intuitive idea, which corresponds to the majority of cases. However, we showed one structure in yeast chromosome 9 where a stretch of sequence was of the form $\cdots\alpha\beta\alpha\alpha\alpha\beta\beta\alpha\cdots$. The definition we proposed captures this structure as a set of nested TR, whose period is an integer linear combination of the periods of the basic units α and β. However, it might be necessary, as the analyses of TR unfolds, to generalize the definition of TR in order to include these kinds of structures.

The final observation we have made is that there seems to be a common motif to the TR. Dividing the set of repeats into two subsets: the set $S_{k=2}$ of TR with copy number 2, and the set $S_{k>2}$ of TR with copy number greater than 2, we observed that in the first set, as many as 20 % of the TR shared motifs consisting only of A's or motifs consisting only of T's. If we guided our intuition on the computed base composition of $S_{k=2}$, we might conclude that A's and T's should appear in the motifs with similar frequency. However, this symmetry is broken in the actual motifs. We can infer that this motifs are a kind of backbone to the TR. However it is still unclear whether this backbone plays a role in the genesis of the repeats.

The second subset $S_{k>2}$ also exhibits some regularities. However, these regularities are not as clear as the ones described above for $S_{k=2}$. There is a clear preference for the use of only two bases in the motifs. And these two letters follow the symmetry principle that Watson and Crick strands are, at a molecular level, indistinguishable. Among the patterns occurring most often in $S_{k>2}$ are the patterns with only T's that we discussed above in the context of

$S_{k=2}$. It is our conjecture that these motifs might have something to say with respect to the way the TRs are created. However we are not aware of any wet lab experiment that would substantiate this conjecture.

We expect that a characterization of the TR of yeast at the level of detail that was done in this paper will be useful in the future as a point of comparison when the TR of other organisms are studied. This comparison will yield light on possible molecular mechanisms responsible for the generation of the repeats.

References

[Barrel 1994] Swiss-prot accession number: P08640. Barrell, B.G, Badcock, K., Bankier, A.T., Bowman, S., Brown, D., Churcher, C.M., Connor, R., Copsey, T., Dear, S. Devlin, K., Fraser, A., Gentles, S., Hamlyn, N., Horsnell, T.S., Hunt,.S., Jagels, K., Jones, M., Louis E., Lye, G., Moule, S., Moule, T., Odell, C., Pearson, D., Rajandream, M.A., Riles, L., Rowley, N., Skelton, J., Smith, V., Walsh, S.V., Whitehead, S., submitted to EMBL/GENBANK/DDBJ Data Banks (Dec-1994).

[Benson 1994] Benson, G. 1994, A space efficient algorithm for finding the best non-overlapping alignment score, 1-14. *In* Crochemore, M., and Gusfield, D., eds., *Combinatorial Pattern Matching*, volume **807** of *Lecture Notes in Computer Science*, Springer-Verlag, New York.

[Benson 1994] Benson, G., and Waterman, M., 1994. A method for fast database search for all k-nucleotide repeats, *Nucelic Acids Res.*, **22** 4828-4836.

[Benson 98] Benson, G.,1998. An Algorithm for Finding Tandem Repeats of Unspecified Pattern Size, 20-29. *In* Istrail, S., Pevzner, P., and Waterman, M., eds., *Proc. 2nd Annual ACM Intl Conference on Comp. Mol. Biol, RECOMB 98.*

[Bussey 1995] Swiss-prot accession number: P10862. Bussey, H., Kaback, D.B., Zhong, W., Vo, D.T., Clark, M.W., Fortin, N., Hall, J., Ouellette, B.F., Keng, T., Barton, A.B., *et al.* 1995. The nucleotide sequence of chromosome I from Saccharomyces cerevisiae, *Proc. Natl. Acad. Sci. U. S. A.* 92, 3809-3813.

[Charlesworth 1994] Charlesworth, B., Sniegowski, P.,and Stephan, W, 1994, The Evolutionary dynamics of repetitive DNA in Eukaryotes, *Nature* **371** 215-220.

[Fischetti 1992] Fischetti, V., Landau, G., Schmidt, J., and Sellers, P., 1992. Identifying periodic occurrences of a template with applications to protein structure, 74-86. *In* Galil, Z., Apostolico, A., Crochemore, M., and Manber, U., eds., *Combinatorial Pattern Matching*, volume **644** of *Lecture Notes in Computer Science*, Springer-Verlag, New York.

[Guan 1996] Guan, X., and Uberbacher, E.C., 1996. A Fast Look-up Algorithm for detecting Repetitive DNA Sequences, 718-719. *In* Hunter, L., and Klein, T.E., eds., *Pacific Symposioum on Biocomputing*, World Scientific, Singapore.

[Inman 1997] Inman, K., and Rudin, N. 1997, *An Introduction to Forensic DNA Analysis*, CRC Press, Boca Raton, Florida.

[Kannan 1993] Kannan, S.K., and Myers, E.W. 1993. An algorithm for locating non-overlapping regions with maximum alignment score, 74-86. *In* Galil, Z., Apostolico, A., Crochemore, M., and Manber, U., eds., *Combinatorial Pattern Matching*, volume **684** of *Lecture Notes in Computer Science*, Springer-Verlag, New York.

[Karlin 1990] Karlin, S., and Altschul, S.F., 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2264-2268.

[Karp 1972] Karp, R.M., Miller R.E. and Rosenberg A.L 1972, Rapid identification of repeated patterns in strings, trees and arrays, 125-136. *In* Proc. 4th Annu. ACM Symp. Theory of Computing.

[Landau 1993] Landau, G., and Schmidt, J. 1993, An algorithm for approximate tandem repeats, 120-133.*In* Galil Z., Apostolico, A., Crochemore, M., and Manber, U., eds., *Combinatorial Pattern Matching*, volume **684** of *Lecture Notes in Computer Science*, Springer-Verlag, New York.

[Lewin 1997] Lewin, B. 1997, *Genes VI*, Oxford University Press, New York.

[Li 1998] Li, W., Stolovitzky, G., Bernaola-Galvan Pedro and Oliver, J.L. 1998. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes, *Genome Research*, **8**, 916-928.

[Milosavljevic 1993] Milosavljevic A., and Jurka, J. 1993, Discovering Simple DNA sequences by the Algorithmic Significance Method, *Comput. Appl. Biosci.*, **9**, 407-411.

[Mott 1999] Mott, R., and Tribe, R., 1999. Approximate statistics of gapped alignments, *J. Comp. Bio.* **6**, 91-112.

[Primrose 1995] Primrose, S.B. 1995, *Principles of Genome Analysis*, Blackwell Science, Cambridge, Massachussets.

[Rigoutsos 1998a] Rigoutsos, I., and Floratos, A., 1998. *Bioinformatics*, **14**, 55-67.

[Rigoutsos 1998b] Rigoutsos, I., and Floratos A., 1998. Motif discovery without alignment or enumeration, 221-227. *In* Istrail, S., Pevzner, P., and Waterman, M., eds., *Proc. 2nd Annual ACM Intl Conference on Comp. Mol. Biol, RECOMB 98.*

[Rivals 1995] Rivals, E., and Delgrange, O., 1995. A first step towards chromosome analysis by compression algorithms, 233-239. *In* Bourbakis, N.G. ed., *First International IEEE Symposium on intelligence in Neural and Biological Systems*, IEE Computer Society Press.

[Sagot 1998] Sagot, M.-F., and Myers, E., 1998. Identifying Satellites in Nucleic Acid Sequences, 234-242. *In* Istrail, S., Pevzner, P., and Waterman, M., eds.,*Proc. 2nd Annual Intl Conference on Comp. Mol. Biol, RECOMB 98.*

[Thompson 1994] Thompson, J.D., Higgins, D.G., and Gibson, T.J., 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acid Res.*, **22**, 4673-4680.

[Waterman 1995] Waterman, M.S., 1995. *Introduction to Computational Biology*, Chapman & Hall, London.

[Yeast Genome] Data obtained from *ftp.ebi.ac.uk*, directory *pub/databases/yeast.*

A Tandem Repeats in yeast Chromosome 1

nest	start	stop	k	T	coding	% score	% indels	% global matches	consensus length	clustalw score
<	0	76	2	38	noncoding	79	5	79	39	158
<>	1	43	2	21	noncoding	81	9	81	22	87
>	15	75	2	30	noncoding	70	6	70	31	104
	254	276	2	11	noncoding	100	0	100	11	51
	727	751	2	12	noncoding	83	0	83	12	42
	1354	1382	2	14	noncoding	71	0	71	14	79
<	6079	6107	2	14	noncoding	71	0	71	14	74
>	6082	6106	2	12	noncoding	83	0	83	12	63
	9273	9307	2	17	noncoding	82	0	82	17	81
	*11873	11927	2	27	YAL065C	88	0	88	27	173
	*12257	12320	3	21	noncoding	82	0	71	21	390
<	*12465	12849	8	48	noncoding	88	0	58	48	8795
<>	12466	12850	4	96	noncoding	89	0	75	96	3862
>	12543	12768	5	45	noncoding	84	10	50	57	2994
	12864	12900	2	18	noncoding	72	0	72	18	100
<	*12996	13131	3	45	noncoding	81	0	68	45	737
>	13024	13120	2	48	noncoding	72	11	72	51	280
	*14790	14816	2	13	noncoding	92	0	92	13	74
	14892	14928	2	18	noncoding	83	0	83	18	97
	17153	17183	2	15	noncoding	73	0	73	15	72
	18041	18079	2	19	noncoding	73	0	73	19	78
	23737	23763	2	13	noncoding	84	0	84	13	53
	*24305	24359	2	27	YAL063C	88	0	88	27	173
	*24687	24771	4	21	YAL063C	80	0	61	21	806
	*25161	25296	3	45	YAL063C	81	0	68	45	737
<	*25300	27190	14	135	YAL063C	87	0	23	137	74690
>	25397	27287	7	270	YAL063C	85	0	43	272	34493
	30990	31016	2	13	noncoding	76	0	76	13	71
<	33392	33428	2	18	noncoding	72	0	72	18	72
>	33394	33424	2	15	noncoding	80	0	80	15	62
	39118	39148	2	15	noncoding	80	0	80	15	73
	40247	40283	2	18	YAL056W	73	10	73	19	96
	42750	42782	2	16	noncoding	75	0	75	16	65
	43265	43291	2	13	YAL054C	76	0	76	13	58
	48205	48235	2	15	YAL053W	73	0	73	15	61
	49227	49263	2	18	YAL051W	72	0	72	18	74

nest	start	stop	k	T	coding	% score	% indels	% global matches	consensus length	clustalw score
	54825	54851	2	13	noncoding	76	0	76	13	62
	55585	55627	2	21	YAL047C	71	0	71	21	90
	57510	57540	2	15	YAL045C	73	0	73	15	72
	57900	57920	2	10	noncoding	100	0	100	10	52
	63749	63777	2	14	YAL041W	71	0	71	14	68
	65144	65168	2	12	YAL041W	83	0	83	12	54
	67717	67753	2	18	noncoding	72	0	72	18	74
	71671	71703	2	16	noncoding	75	0	75	16	83
	72264	72294	2	15	YAL038W	80	0	80	15	82
<	73314	73344	2	15	noncoding	73	0	73	15	59
<>	73316	73342	2	13	noncoding	84	0	84	13	61
>	73321	73349	2	14	noncoding	85	0	85	14	70
	73612	73636	2	12	noncoding	100	0	100	12	60
	76645	76681	2	18	YAL035W	72	0	72	18	90
<	76699	76723	2	12	YAL035W	83	0	83	12	70
>	76705	76735	2	15	YAL035W	86	0	86	15	86
	76788	76812	2	12	YAL035W	83	0	83	12	69
	76823	76853	2	15	YAL035W	86	0	86	15	90
	76977	77049	3	24	YAL035W	75	0	58	24	359
	77066	77124	2	29	YAL035W	70	6	70	30	156
<	77491	77527	2	18	YAL035W	94	0	94	18	121
<>	77498	77558	2	30	YAL035W	73	0	73	30	148
>	77503	77551	2	24	YAL035W	75	0	75	24	112
	77571	77607	3	12	YAL035W	79	0	58	12	110
	77677	77703	2	13	YAL035W	76	0	76	13	63
	78880	78904	2	12	YAL035W	83	0	83	12	63
	88083	88123	2	20	YAL029C	75	0	75	20	96
	89987	90015	2	14	YAL029C	71	0	71	14	77
	91451	91475	2	12	YAL029C	91	0	91	12	68
<	93836	93881	3	15	YAL028W	73	0	60	15	202
>	93852	93876	2	12	YAL028W	83	0	83	12	46
	94037	94067	2	15	YAL028W	73	0	73	15	85
	94581	94605	2	12	noncoding	83	0	83	12	65
	95568	95602	2	17	noncoding	76	0	76	17	84
	98184	98208	2	12	YAL026C	83	0	83	12	51
	*99939	99967	2	14	noncoding	100	0	100	14	88

nest	start	stop	k	T	coding	% score	% indels	% global matches	consensus length	clustalw score
<	100356	100404	2	24	YAL025C	75	0	75	24	111
<>	*100359	100413	3	18	YAL025C	74	0	55	18	214
>	100371	100395	2	12	YAL025C	83	0	83	12	61
	100456	100510	2	27	YAL025C	85	0	85	27	111
	*101465	101510	3	15	noncoding	79	0	60	15	194
	105997	106023	2	13	noncoding	76	0	76	13	72
	110711	110739	2	14	noncoding	85	0	85	14	79
	112739	112775	3	12	YAL021C	83	0	75	12	228
	113287	113317	2	15	YAL021C	93	0	93	15	114
	116169	116199	2	15	YAL019W	73	0	73	15	66
	116306	116338	2	16	YAL019W	81	0	81	16	67
	116478	116514	2	18	YAL019W	88	0	88	18	121
<	118304	118358	2	27	YAL019W	70	0	70	27	87
>	118310	118348	2	19	YAL019W	73	0	73	19	68
	118474	118510	2	18	noncoding	77	0	77	18	72
	119027	119057	2	15	YAL018C	80	0	80	15	70
	120115	120147	2	16	noncoding	75	0	75	16	91
	124188	124228	2	20	YAL017W	70	0	70	20	87
	124723	124755	2	16	noncoding	81	0	81	16	92
	124949	124979	2	15	YAL016W	80	0	80	15	67
	125626	125648	2	11	YAL016W	90	0	90	11	56
	128579	128603	2	12	YAL014C	83	0	83	12	74
	129700	129730	2	15	YAL013W	80	0	80	15	98
	132295	132325	2	15	YAL011W	86	0	86	15	100
<	132934	132976	2	21	YAL011W	76	0	76	21	78
>	132952	132978	2	13	YAL011W	76	0	76	13	57
	133377	133419	2	21	YAL011W	76	0	76	21	100
	134140	134168	2	14	noncoding	78	0	78	14	58
	134468	134496	2	14	YAL010C	78	0	78	14	73
	135256	135292	2	18	YAL010C	73	10	73	19	105
	137500	137526	2	13	YAL008W	76	0	76	13	54
	137537	137563	2	13	noncoding	76	0	76	13	66
	137910	137952	2	21	YAL007C	71	0	71	21	94
	138124	138148	2	12	YAL007C	83	0	83	12	54
	139546	139606	5	12	YAL005C	72	0	41	12	510
	141248	141278	2	15	YAL005C	86	0	86	15	80
	142793	142823	2	15	YAL003W	73	0	73	15	79
	149047	149075	2	14	YAL001C	71	0	71	14	61

nest	start	stop	k	T	coding	% score	% indels	% global matches	consensus length	clustalw score
<	151477	151513	2	18	noncoding	72	0	72	18	67
>	151479	151517	2	19	noncoding	73	0	73	19	77
	155851	155881	2	15	YAR003W	73	0	73	15	62
	157135	157159	2	12	YAR007C	83	0	83	12	57
	162448	162478	2	15	YARCTy1-1	80	0	80	15	91
	168114	168156	2	21	YAR014C	71	0	71	21	93
<	169202	169242	2	20	noncoding	76	9	76	21	100
>	169206	169238	2	16	noncoding	75	0	75	16	67
	173527	173567	2	20	YAR019C	70	0	70	20	88
	176545	176573	2	14	noncoding	85	0	85	14	76
	176957	176999	2	21	YAR020C	80	0	80	21	119
	181906	181936	2	15	noncoding	73	0	73	15	57
	183111	183143	2	16	noncoding	75	0	75	16	80
	183171	183193	2	11	YARWdelta6	90	0	90	11	59
	183267	183293	2	13	YARWdelta6	76	0	76	13	72
	185489	185519	2	15	YAR028W	73	0	73	15	80
	189216	189242	2	13	noncoding	76	0	76	13	53
	*190123	190151	2	14	noncoding	92	0	92	14	64
	192077	192107	2	15	YAR035W	73	0	73	15	83
	192243	192265	2	11	YAR035W	90	0	90	11	74
<	192465	192501	2	18	noncoding	72	0	72	18	68
>	192466	192488	2	11	noncoding	90	0	90	11	51
<	193956	194016	2	30	YAR044W	73	0	73	30	136
<>	193958	194012	2	27	YAR044W	77	0	77	27	127
<>	193994	194030	2	18	YAR044W	72	0	72	18	101
<>	193996	194026	2	15	YAR044W	80	0	80	15	96
>	193999	194023	2	12	YAR044W	91	0	91	12	74
	194165	194189	2	12	YAR044W	83	0	83	12	63
	196002	196038	2	18	YAR044W	77	0	77	18	80
	*198829	198862	3	11	noncoding	86	0	72	11	206
	199938	199966	2	14	noncoding	71	0	71	14	58
	202636	202670	2	17	noncoding	76	0	76	17	79
	204072	204108	2	18	YAR050W	72	0	72	18	66
<	204209	206639	9	270	YAR050W	88	0	59	270	49726
>	*204223	206653	18	135	YAR050W	89	0	51	135	105008
<	206736	206826	6	15	YAR050W	81	0	53	15	1113
>	*206737	206827	2	45	YAR050W	77	0	77	45	232

nest	start	stop	k	T	coding	% score	% indels	% global matches	consensus length	clustalw score
	206933	206969	2	18	YAR050W	72	0	72	18	74
	*207219	207282	3	21	YAR050W	80	0	76	21	262
	*207612	207693	3	27	YAR050W	88	0	77	27	418
	212972	212994	2	11	noncoding	90	0	90	11	50
	215008	215044	2	18	noncoding	73	10	73	19	94
	216271	216295	2	12	noncoding	83	0	83	12	57
	216481	216513	2	16	noncoding	75	0	75	16	67
	217348	217388	2	20	YAR060C	80	0	80	20	75
<	217455	217481	2	13	YAR060C	84	0	84	13	69
>	217480	217502	2	11	YAR060C	90	0	90	11	62
	219173	219218	3	15	noncoding	76	0	66	15	218
	220806	220836	2	15	noncoding	86	0	86	15	81
<	222094	222118	2	12	noncoding	83	0	83	12	51
>	222114	222156	2	21	noncoding	71	0	71	21	88
	222177	222201	2	12	noncoding	83	0	83	12	46
<	222275	222329	3	18	noncoding	74	0	61	18	243
>	222283	222337	2	27	noncoding	77	0	77	27	119
	222535	222565	2	15	YAR068W	73	0	73	15	67
	224923	224947	2	12	noncoding	83	0	83	12	53
	225397	225439	2	21	noncoding	71	0	71	21	76
	*229745	229805	2	30	noncoding	80	0	80	30	153
	*229944	229977	3	11	noncoding	81	0	63	11	180
<	230081	230105	2	12	noncoding	83	0	83	12	65
>	230102	230138	2	18	noncoding	84	10	84	19	174
	230150	230204	3	18	noncoding	70	27	45	24	435

In this Appendix we have tabulated all the (p, μ) TR with $p = 0.05$ and $\mu = 70\%$ of yeast chromosome 1. The first column in that table shows the existence of a nesting of TR of different periods (see main text). The $<$ symbol indicates the first of the TR in a nest. The member of the nest which starts the latest (but does not necessarily end the latest) is indicated by a $>$ symbol. In between these extremes, the other members of the nest are indicated by the $<>$ symbol. The asterisks to the left of the second column in some of the entries correspond to the TR reported by Benson [Benson 1994]. In the fourth column k is the copy-number. In the fifth column, T is the period. The other columns are self explanatory.

Figure Captions

Figure 1. (a) and (b): The function $h_T(i)$ (defined in Eq. (5)) used to screen out regions that do not contain tandem repeats, for the first 22 Kbp of yeast chromosome 1, and (a) $T = 27$ and (b) $T = 135$. The insets of both (a) and (b) are enlargements of the biggest peak in the main figures. (c) and (d): The function $f_T(i)$ (defined in Eq. (6) used to compute the extremes of the tandem repeats, for the first 22 Kbp of yeast Chr. 1, and (c) $T = 27$ and (d) $T = 135$. The insets of both (c) and (d) are enlargements of the biggest peak in the main corresponding figures. The information contained in (c) and (d) is a low-pass filtering of the data of (a) and (b) respectively.

Figure 2. Mean value and standard deviation of the global alignment scores between two sequences of different lengths (abscissae) taken at random from chromosomes 1, 6 and 8 of S. Cerevisiae. The thick solid lines are the theoretical fits of Eqs. (9) and (10) to the data. The inset, in logarithmic scale, shows that the standard deviation decreases as the power $-1/2$ of the period.

Figure 3. Pseudocode of the algorithm presented in this paper.

Figure 4. Number of TR of copy number 2 (solid squares) and copy number greater than 2 (solid circles) detected by our algorithm as a function of the chromosome length in the 16 nuclear yeast chromosomes. The plus sign indicates the number of tandem repeats detected in a random sequence of same base composition as yeast, and approximately the same length as chromosome 4. For the random sequence, all but one of the TRs had copy number 2. The numbers on top of the solid circles indicate the chromosome number.

Figure 5. Percentage of the total chromosome length covered by TR of all copy numbers, as a function of the chromosome length, within ORFs (plus symbol) and outside ORFs (circles). In the inset we have plotted for each chromosome, the absolute length taken by TR within ORFs *vs* the absolute length taken by the tandem repeats outside ORFs. The dashed line in the inset is the identity. The fact that the overall trend of the data is to be steeper than the identity indicates that there are more TR within coding regions than outside of them.

Figure 6. Percentage of total chromosome length covered by TRs of copy numbers strictly greater than 2, as a function of the chromosome length, within ORFs (plus symbol) and outside ORFs (circles). In the inset we have plotted for each chromosome, the absolute length taken by the TRs within ORFs *vs* the absolute length taken by the TRs outside ORFs for TR of copy numbers greater than 2. The dashed line in the inset is the identity. The fact that the

overall trend of the data is to be steeper than the identity indicates that there are more TR within coding regions than outside of them.

Figure 7. Number of occurrences of TRs with a given copy number versus the copy number (for copy numbers greater or equal than 3) inside the coding regions (a) and outside of them (b). For better resolution, we plotted copy number equal to 3 in the insets. Each chromosome is indicated by a different shade, according to the coding of the inset of Fig. 7(a). The ordering of the chromosomes is according to their length.

Figure 8. Number of occurrences of TRs with a given period versus the period inside the coding regions (a) and outside of them (b), for all copy numbers. For better resolution, we plotted periods greater than 35 in the insets, and used a slightly non-linear scale in the abscissae. Each chromosome is indicated by a different shade, according to the coding of the inset of Fig. 7(a). The ordering of the chromosomes is according to their length.

Figure 9. The copy number *vs* the period for all the TRs detected in the 16 chromosomes of yeast. Each chromosome is indicated by a different symbol, according to the legends. The dashed lines are hyperbolae corresponding to TRs of fixed total length of 250 *bp* (dot-dashed curve) and $2,500$ *bp* (dashed curve).

Figure 10. The period (a) and copy number (b) of the TRs of chromosome 4, *vs* the position along chromosome 4 where the corresponding TR is located. The thick lines in both (a) and (b) corresponds to copy number strictly greater than 2. The fainter line corresponds to TR of copy number exactly 2.

Figure 11. For yeast chromosome 4: histogram of the periods for TRs of all copy numbers (a) and TRs with copy numbers strictly greater than 2. For better resolution, the period range was separated in periods smaller than 30 (main figures in (a) and (b)) and periods greater than 30 (insets in (a) and (b)). (c) and (d): histograms of copy numbers for TRs of all copy numbers, and copy numbers strictly greater than 2, respectively.

Figure 12. The period (a) and copy number (b) of the TRs detected in a random sequence (of length and base composition similar to those of yeast chromosome 4), *vs* the position along the sequence where the corresponding TR is located. The thick line in both (a) and (b) corresponds to copy numbers strictly greater than 2. The fainter line corresponds to TRs of copy number exactly 2.

Figure 13. For the random sequence described in the caption of the previous

figure: histogram of the periods for TRs of all copy numbers (a) and histograms of copy numbers (b).

Figure 14. Base composition in the TRs with copy number equal to 2 (a) and copy numbers greater than 2 (b), for all the yeast chromosomes. In (a) we indicated with open symbols the base composition in the TR of the random sequence described in the caption of Fig. 12, with abscissa equal to chromosome 4. The shade coding and legends for (b) are the same as the ones indicated in (a).

Figure 1 (a) and (b)

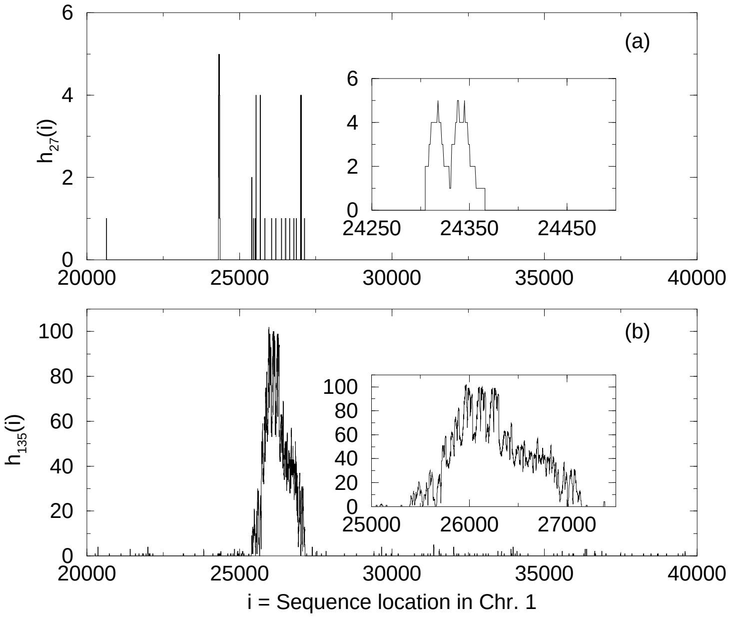


Figure 1 (c) and (d)

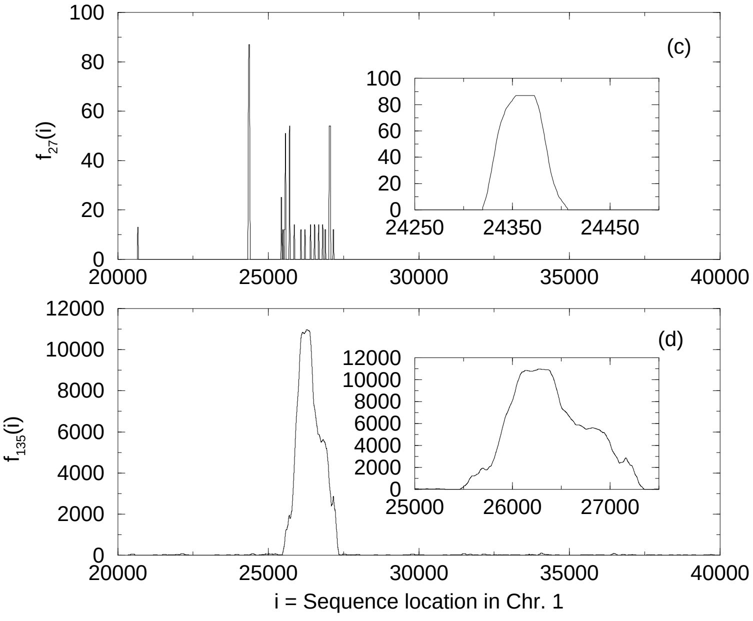
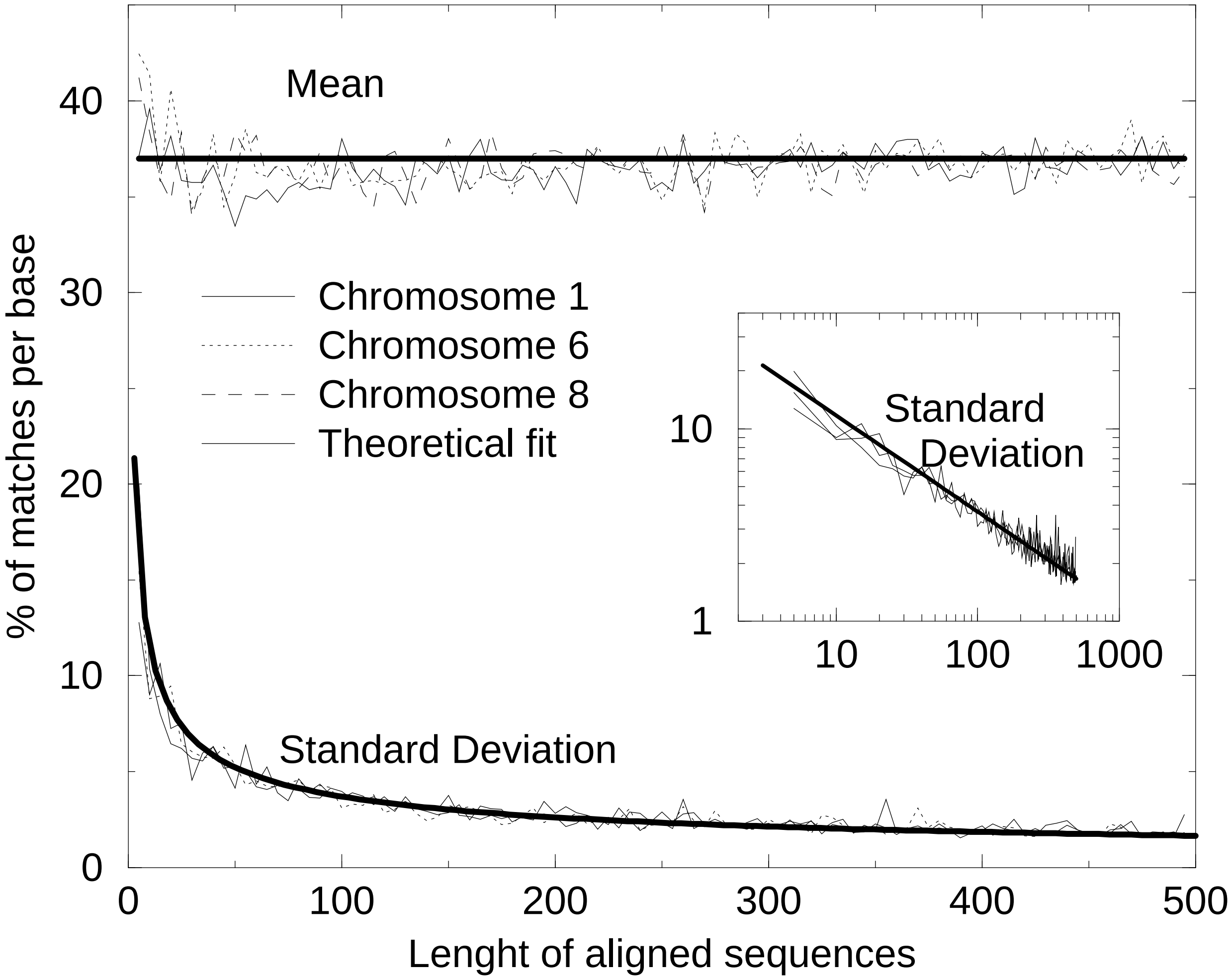
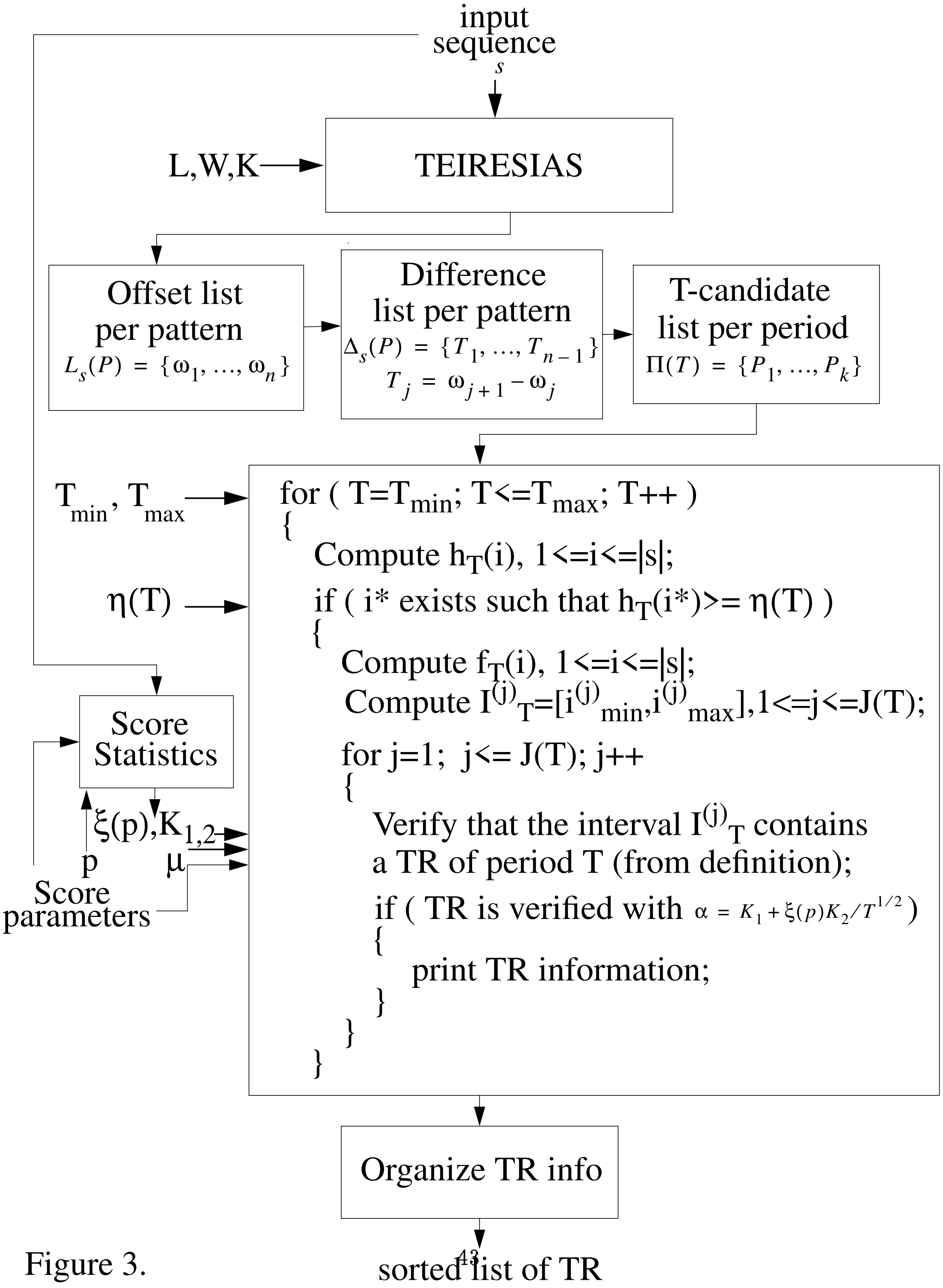


Figure 2

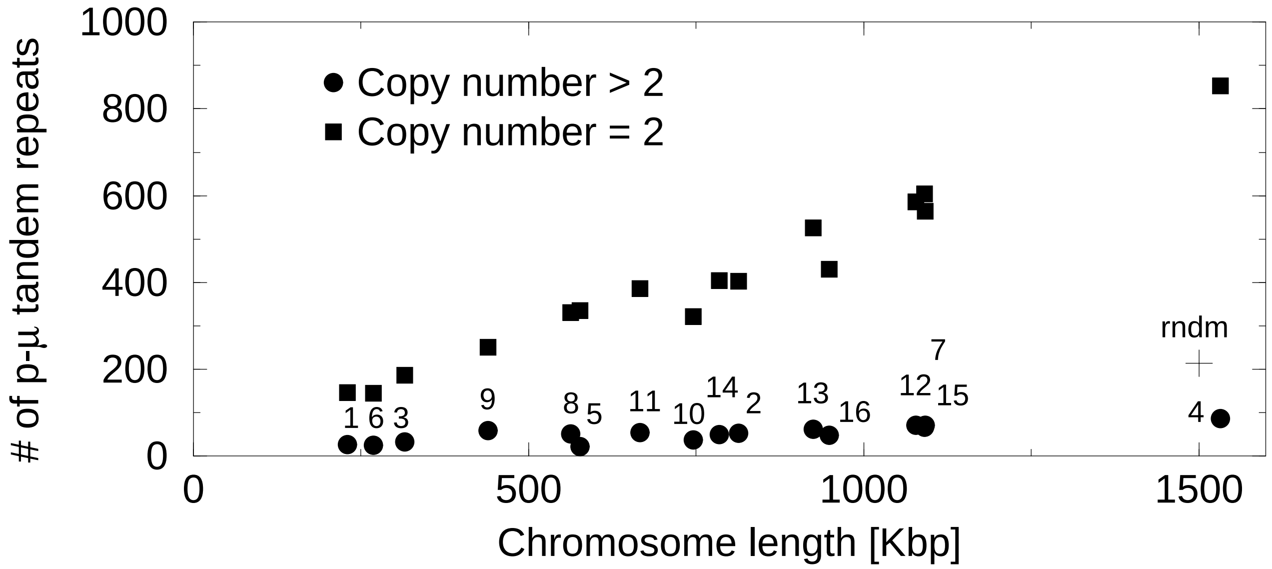

Mean
40
30
20
10
0
% of matches per base
Chromosome 1
Chromosome 6
Chromosome 8
Theoretical fit
Standard Deviation
Deviation
10
1
10
100
1000
Standard Deviation
0
100
200
300
400
500
Lenght of aligned sequences



```
for ( T=$T_{min}$; T<=$T_{max}$; T++ )
{
    Compute $h_T(i)$, 1<=i<=|s|;
    if ( i* exists such that $h_T(i^*)$>= $\eta(T)$ )
    {
        Compute $f_T(i)$, 1<=i<=|s|;
        Compute $I^{(j)}_T=[i^{(j)}_{min},i^{(j)}_{max}]$,1<=j<=J(T);
        for j=1;  j<= J(T); j++
        {
            Verify that the interval $I^{(j)}_T$ contains
            a TR of period T (from definition);
            if ( TR is verified with $\alpha = K_1 + \xi(p) K_2 / T^{1/2}$ )
            {
                print TR information;
            }
        }
    }
}
```

Figure 3.

Figure 4



# of p-μ tandem repeats
1000
800
600
400
200
0
0
500
1000
1500
Chromosome length [Kbp]
● Copy number > 2
■ Copy number = 2
rndm
7
1
6
3
9
8
5
11
10
14
2
13
16
12
15
4

Figure 5

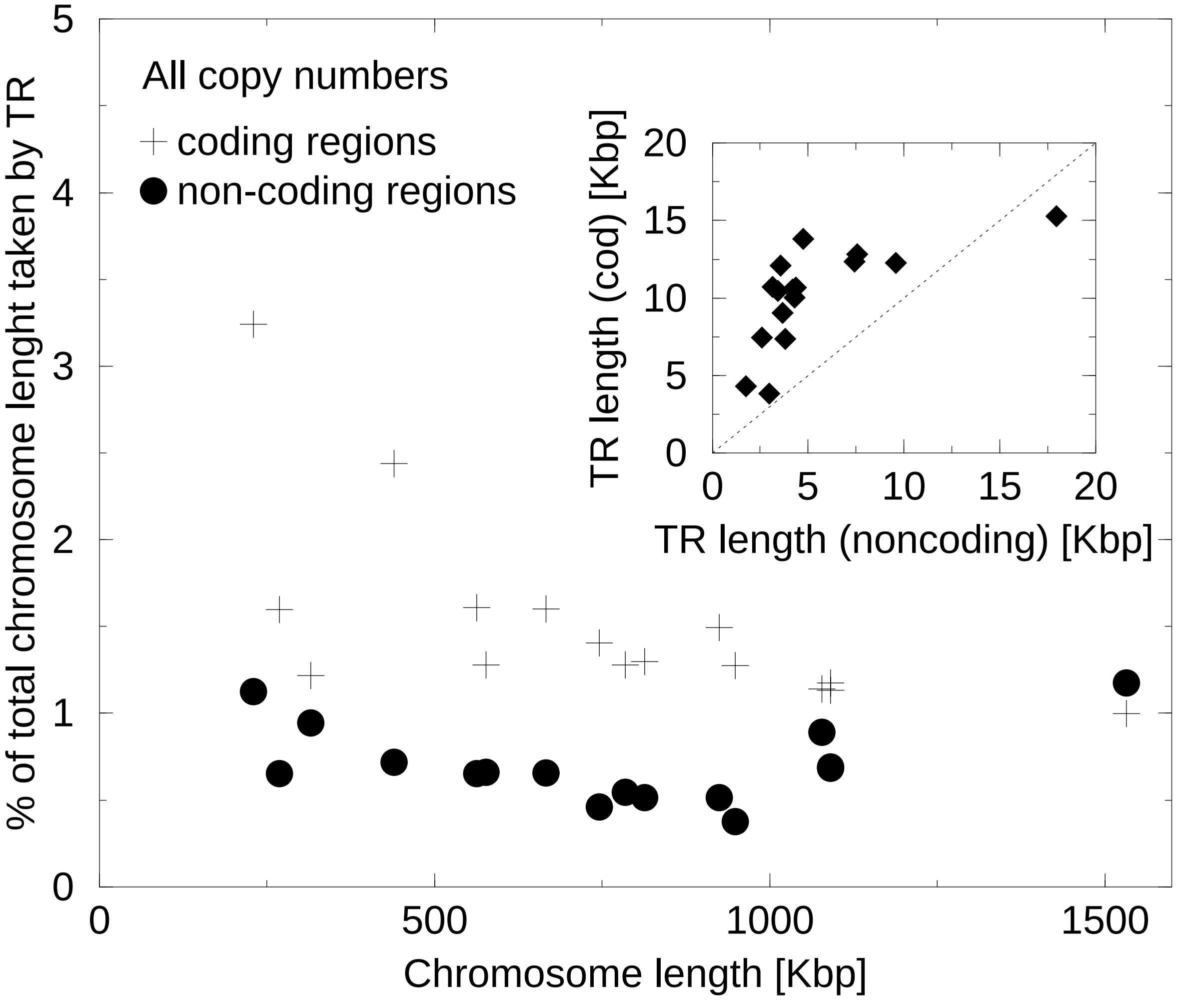

All copy numbers
+ coding regions
● non-coding regions
% of total chromosome lenght taken by TR
Chromosome length [Kbp]
0
1
2
3
4
5
0
500
1000
1500
TR length (cod) [Kbp]
TR length (noncoding) [Kbp]
0
5
10
15
20

Figure 6

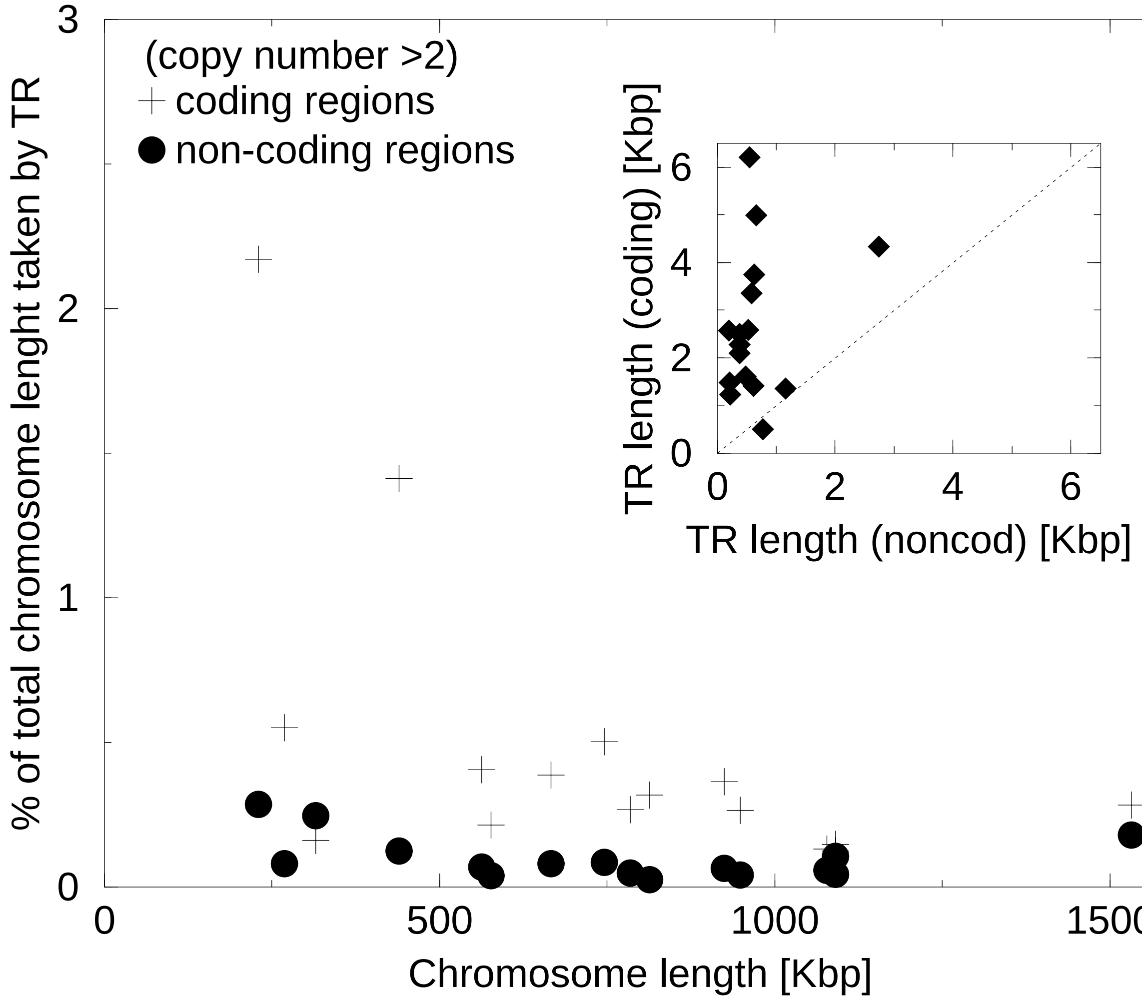

(copy number >2)
+ coding regions
● non-coding regions
% of total chromosome lenght taken by TR
Chromosome length [Kbp]
0
1
2
3
0
500
1000
1500
TR length (coding) [Kbp]
TR length (noncod) [Kbp]
0
2
4
6
0
2
4
6

Figure 7

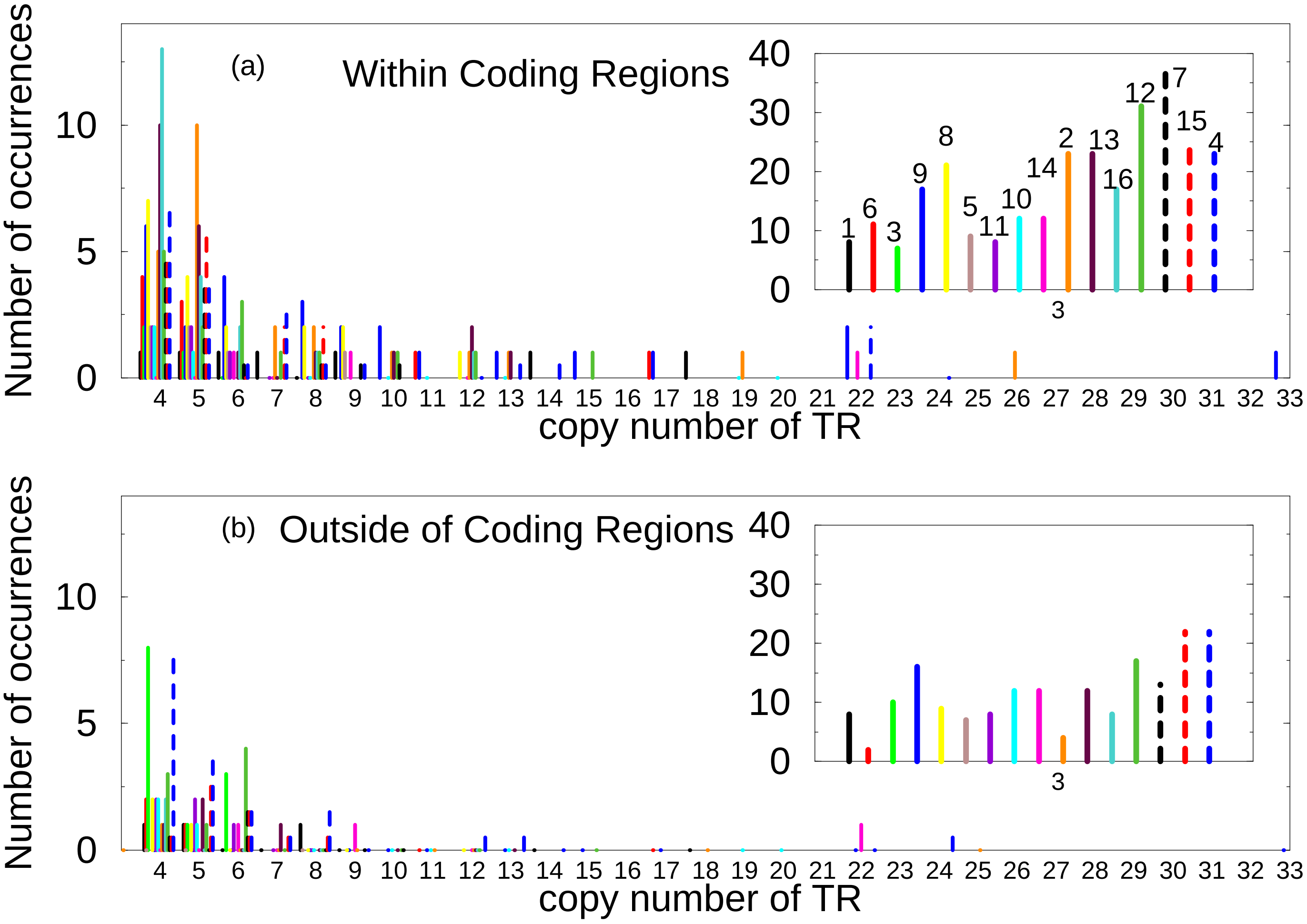

(a)
Within Coding Regions
Number of occurrences
0
5
10
copy number of TR
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
0
10
20
30
40
1
6
3
9
8
5
11
10
14
2
13
16
12
7
15
4
3
(b)
Outside of Coding Regions
Number of occurrences
0
5
10
copy number of TR
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
0
10
20
30
40
3

Figure 8

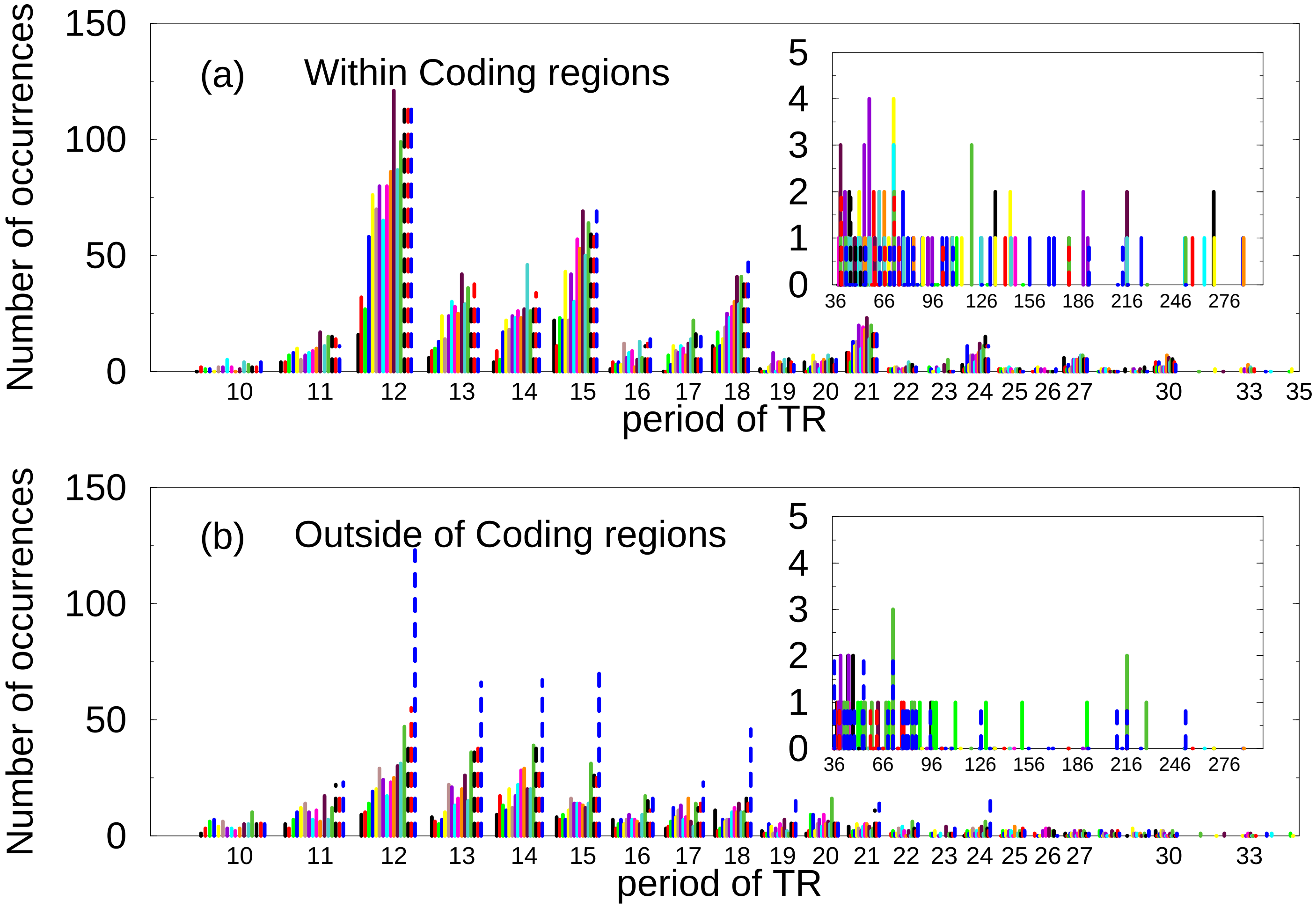

(a) Within Coding regions
(b) Outside of Coding regions
Number of occurrences
0
50
100
150
period of TR
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
30
33
35
0
1
2
3
4
5
36
66
96
126
156
186
216
246
276

Figure 9

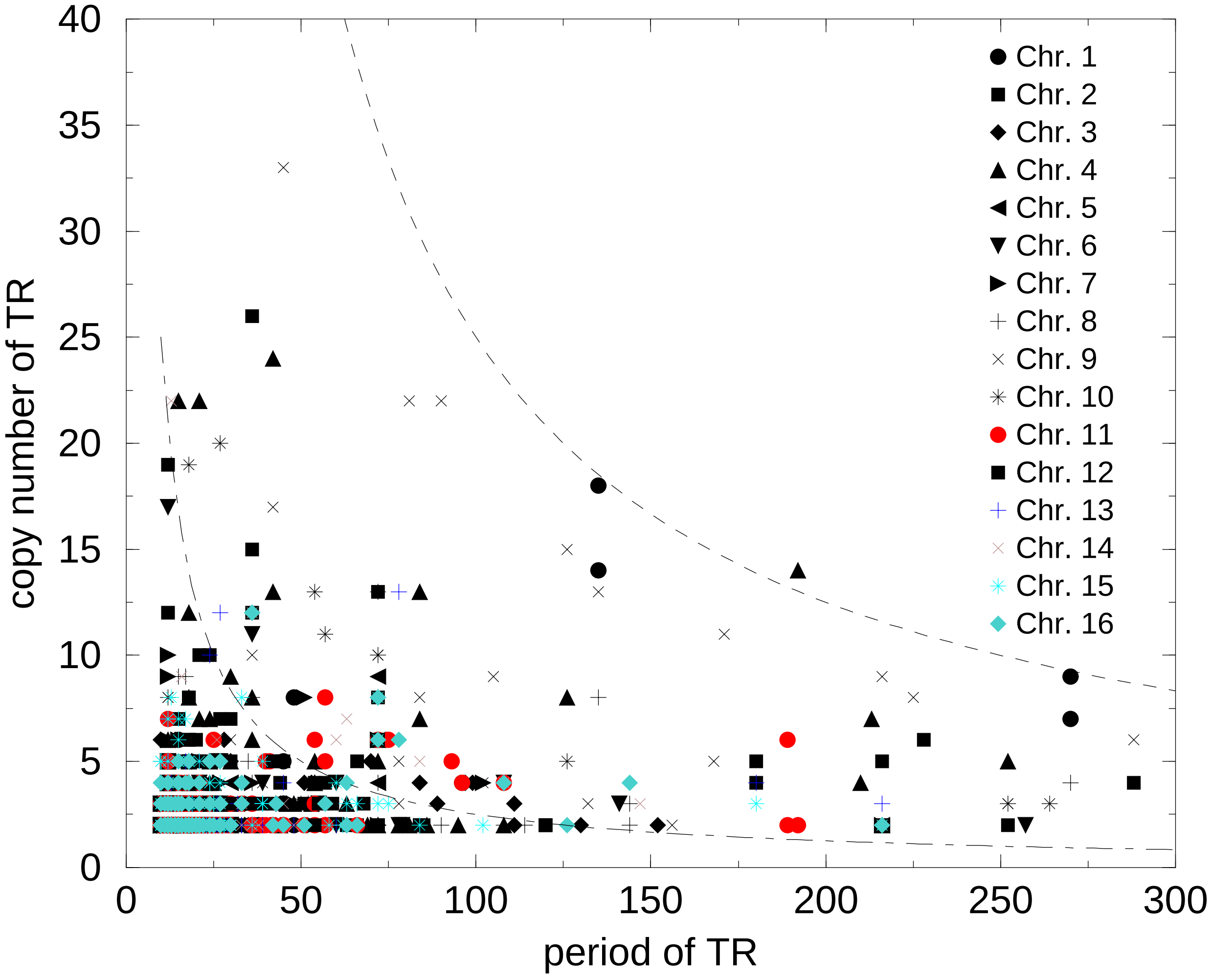

copy number of TR
period of TR
0
5
10
15
20
25
30
35
40
0
50
100
150
200
250
300
● Chr. 1
■ Chr. 2
◆ Chr. 3
▲ Chr. 4
◀ Chr. 5
▼ Chr. 6
▶ Chr. 7
+ Chr. 8
× Chr. 9
✳ Chr. 10
● Chr. 11
■ Chr. 12
+ Chr. 13
× Chr. 14
✳ Chr. 15
◆ Chr. 16

Figure 10

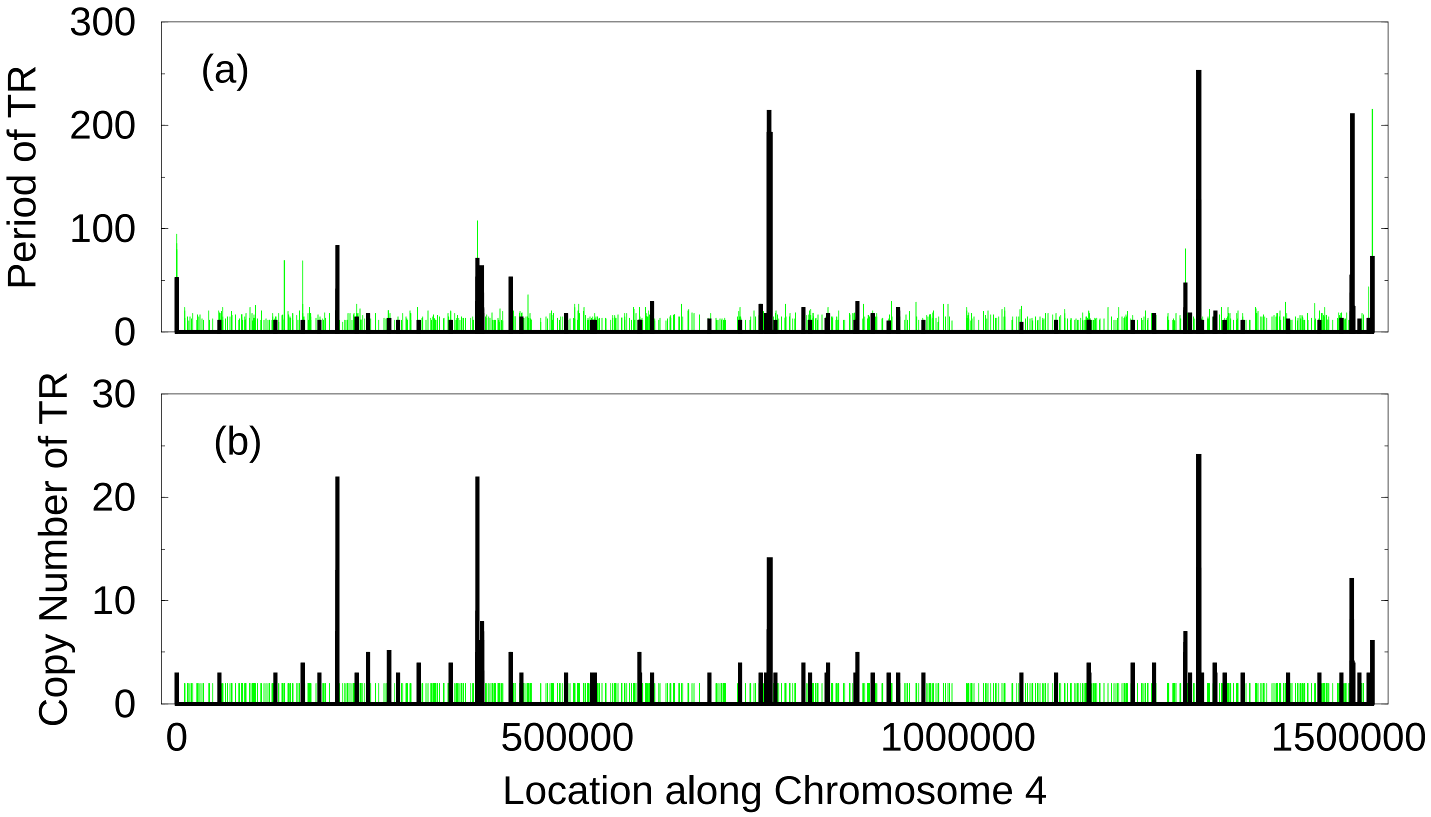

300
200
100
0
Period of TR
(a)
30
20
10
0
Copy Number of TR
(b)
0
500000
1000000
1500000
Location along Chromosome 4

Figure 11

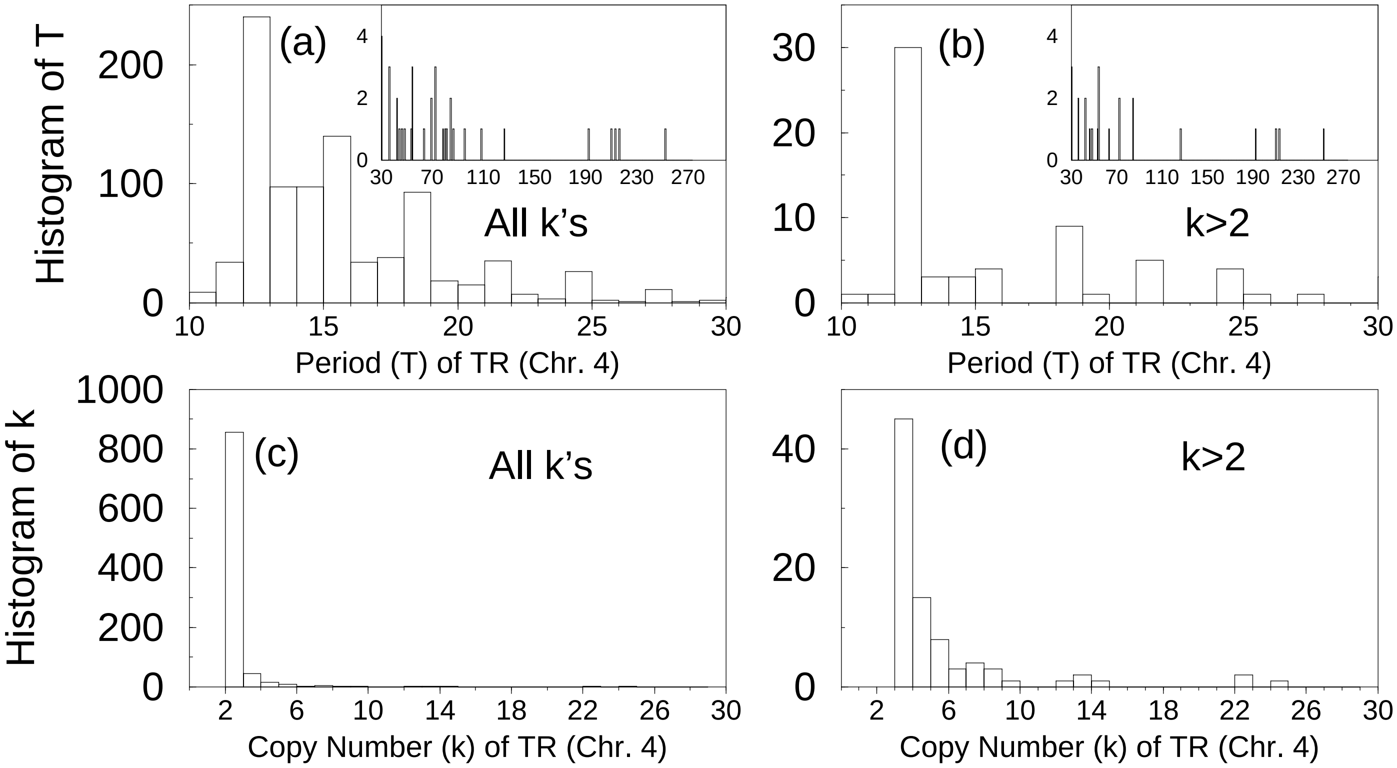
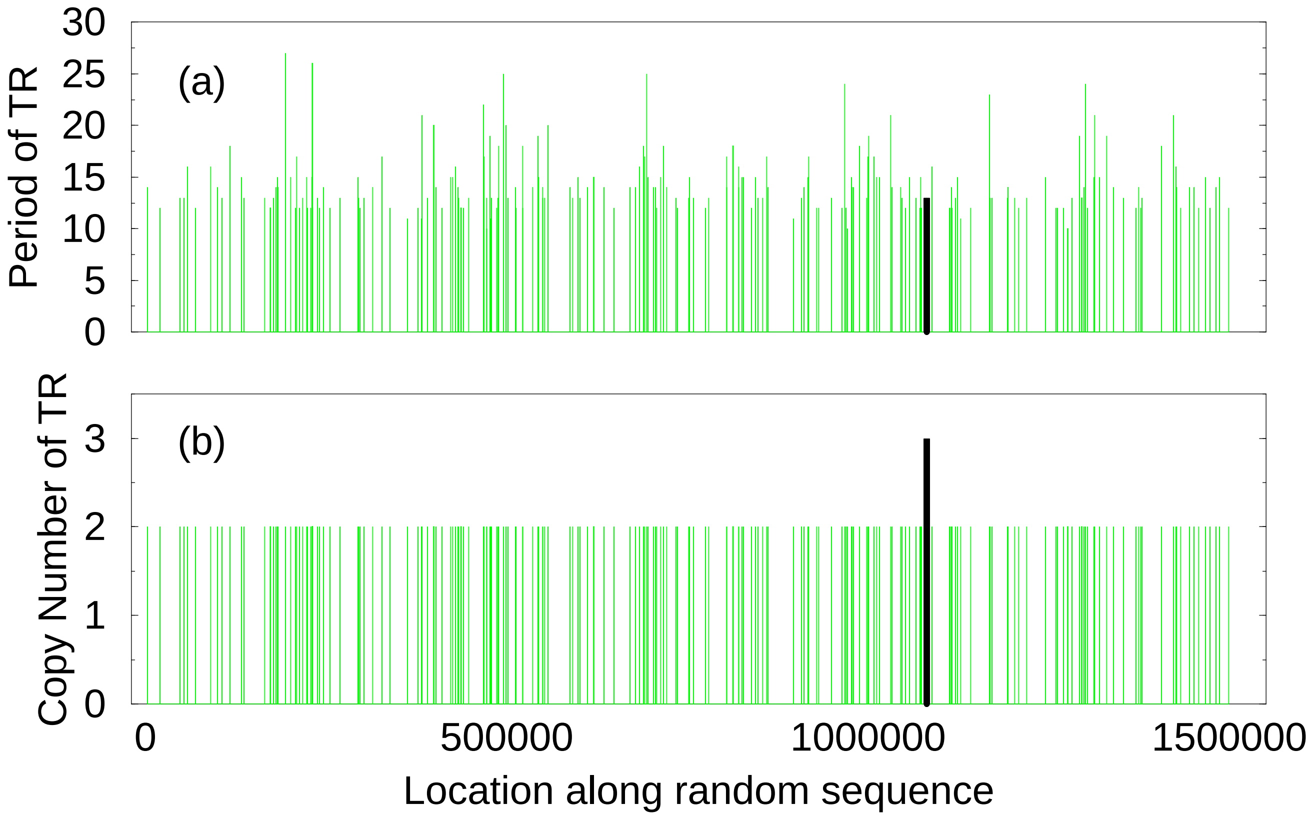

(a)
All k's
Histogram of T
0
100
200
10
15
20
25
30
Period (T) of TR (Chr. 4)
0
2
4
30
70
110
150
190
230
270
(b)
k>2
0
10
20
30
10
15
20
25
30
Period (T) of TR (Chr. 4)
0
2
4
30
70
110
150
190
230
270
(c)
All k's
Histogram of k
0
200
400
600
800
1000
2
6
10
14
18
22
26
30
Copy Number (k) of TR (Chr. 4)
(d)
k>2
0
20
40
2
6
10
14
18
22
26
30
Copy Number (k) of TR (Chr. 4)

Figure 12



(a)
Period of TR
30
25
20
15
10
5
0
(b)
Copy Number of TR
3
2
1
0
0
500000
1000000
1500000
Location along random sequence

Figure 13

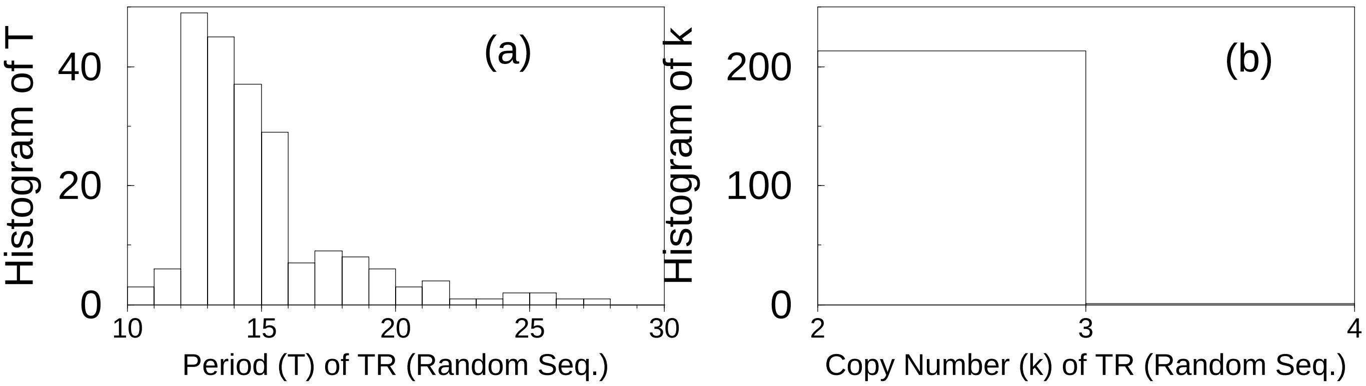
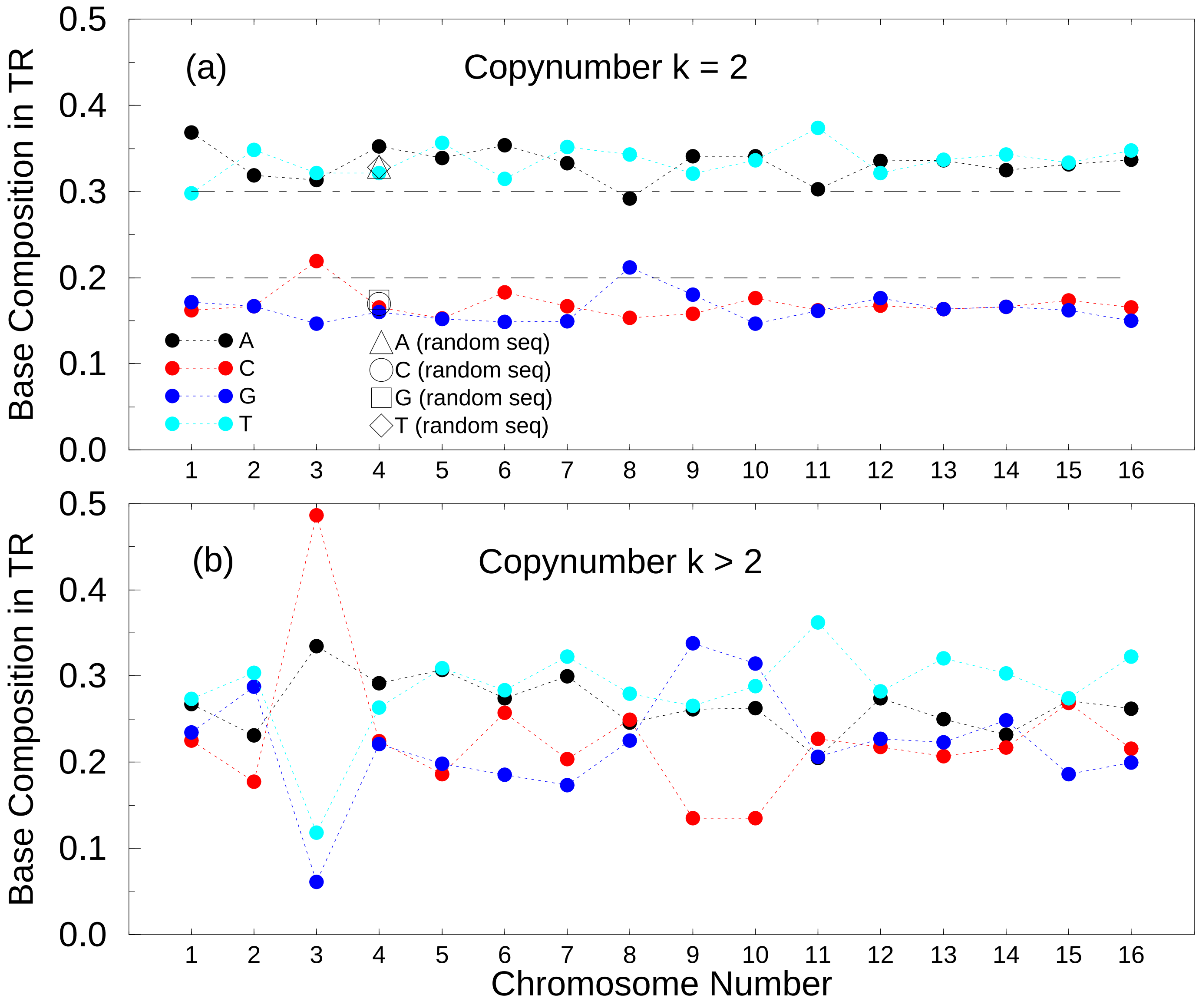

Histogram of T
40
20
0
10
15
20
25
30
Period (T) of TR (Random Seq.)
(a)
Histogram of k
200
100
0
2
3
4
Copy Number (k) of TR (Random Seq.)
(b)

Figure 14


(a)
Copynumber k = 2
Base Composition in TR
0.0
0.1
0.2
0.3
0.4
0.5
A
C
G
T
A (random seq)
C (random seq)
G (random seq)
T (random seq)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
(b)
Copynumber k > 2
Base Composition in TR
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Chromosome Number