Detection of latent sequence periodicities

Elisabetta Pizzi, Sabino Liuni[1] and Clara Frontali*
Laboratorio di Biologia Cellulare, Istituto Superiore di Sanità, Rome and [1]Centro Studi Mitocondri e Metabolismo Energetico, University of Bari, Italy

Received April 16, 1990; Accepted June 8, 1990

ABSTRACT

A method is proposed for the automatic detection of serial periodicities in a linear sequence. Its application to DNA subtelomeric sequences from two lower eukaryotes, *P.falciparum* and *S.cerevisiae*, reveals ordered patterns organised in hierarchical periodicities, not easily recognizable by other methods. The possible implications concerning the evolution of tandemly repetitive arrays are discussed in light of a model which involves, as successive steps, random repeat modification, the fusion of differently modified repeat versions into longer units, and the amplification of (and/or homogenization to) the more recent repeat units.

INTRODUCTION

DNA sequences which are the result of the tandem reiteration of basic repeat units are frequent in the genomes of both lower and higher eukaryotes (e.g. in satellite DNAs). The individual units which make up the array are not always perfectly identical. This is not surprising, since it is reasonable to expect that an accumulation of random events modifies the individual repeat units in irregular fashion.

In several cases, however, periodic correlations are observed in the occurrence of repeat variants. Their effect is such that suprarepeat units, comprised of a set of variously modified basic repeats, exhibit a higher degree of homology than the shorter repeats they contain. A typical example is given by the array of 36 bp units present in the Y′ subtelomeric sequence of *S.cerevisiae* (1). Other examples take the form of shifted, conserved patterns of base substitutions in the regions coding for repeated epitopes in plasmodial antigen genes (2–9).

A hierarchy of at least four related periodicities was described in the mouse satellite DNA as early as 1975 by E. Southern in a what was truly a pioneering work (10). According to his model, which involves successive rounds of repeat modification and multiplication, the observed periodicities reflect successive stages in the evolution of the sequence. G.Dover (11–14), when dealing with this kind of concerted evolution, postulates a role for active mechanisms of repeat homogenization, such as biased gene conversion, unequal crossing-over or slippage replication.

The widespread procedure which consists in extracting from a series of imperfect repeats just a consensus sequence, often masks the periodic recurrence of identical modifications of a basic repeat and hinders the detection of regular supra-periodic patterns. The computer programs which are generally employed to search for internal (perfect or imperfect) homologies in a given sequence do not allow for the easy recognition of such latent periodicities. The repetitive structure is in many cases not easily detectable even to the expert eye, because of the extensive diversification of the individual repeats merged in longer units.

The purpose of the present paper is to test a very simple algorithm designed to detect latent periodicities by a reinforcing procedure, which is somewhat similar to the methods employed to enhance regular patterns in the analysis of electron microscope images. The regions which were chosen for this test are the above-mentioned subtelomeric sequence of *S.cerevisiae* (1) and a recently published subtelomeric sequence from *P.falciparum* (15). In both cases our analysis allowed the identification of a repetitive pattern containing hierarchical periodicities. The analysis of short-range periodicities allowed us in both cases, furthermore, to suggest plausible models of pattern evolution from an original array of short repeat units, through steps involving random repeat modification, the fusion of modified units in longer suprarepeats, and the amplification of, and/or homogenization to, the new repeat units.

THE ‘ENHANCE’ ALGORITHM

Let us consider a string of n nucleotides and its distribution along a random sequence of N nucleotides ($N >> n$) of known base composition. Let p be the ‘a priori’ probability to find the oligonucleotide at some position along the sequence. p^2 will represent the probability to find the same oligo at two different positions. If d is the distance between any two positions of the n-mer, the frequency distribution in d will be represented by the straight line

$$f(d) = (N-n-d+1)p^2 \quad (1)$$

Equation (1) should adequately represent the distance frequency distribution of a given oligo, randomly occurring in a known sequence of sufficient length, when in the place of p we use the actual frequency of the oligomer, i.e. the number of times the given n-mer is represented in the sequence divided by the total number, $N-n+1$, of strings of length n which can be accommodated in the same sequence.

* To whom correspondence should be addressed

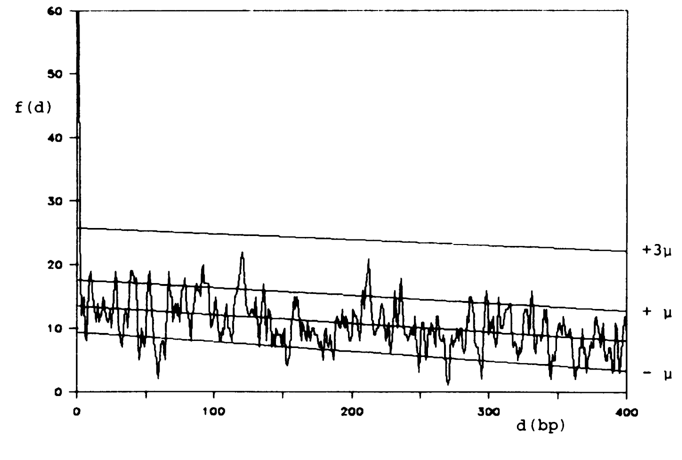

Figure 1. Frequency distribution of distances (d, expressed in base pairs) between AAA trinucleotides in a 1191 bp long, 74% AT rich sequence cloned from *P.berghei* chromosome 5 (Pace et al., in preparation). No regular spacing between groups of Adenines appears in the distribution. This latter is indistinguishable from a random distribution, being completely contained within three mean square errors (μ), with the exception of very high population numbers in classes d = 1 – 3 bp, indicating that AAA trinucleotides are very often part of longer (5 – 6 bp) Adenine runs.

This is shown in fig.1 for a non-repetitive subtelomeric sequence from *P.berghei*. The actual frequency distribution obtained by locating the AAA trinucleotide along the 1191 bp of the sequence, and by calculating distances for all possible position pairs, is shown as a histogram. Best-fit through the histogram columns yields the straight line also plotted on fig.1. Statistical uncertainty on straight line parameters, deriving from sample limitation, is indicated in fig.1 by the lines corresponding to one and three mean square errors. The expected linear decrease, calculated by means of equation (1) on the basis of the actual number of times the oligo is present in the given sequence, practically coincides with the best-fit line.

When applying the same analysis of distance frequency distribution to an oligonucleotide which is periodically repeated along the sequence (but which possibly occurs also elsewhere in it), a series of distinct peaks emerges from an irregular background. Peaks are considered as indicating significant periodicities whenever they exceed the line corresponding to three mean square errors. In the case of a perfectly regular spacing, the positions of the peaks will correspond to exact multiples of the basic periodicity, while their heights will be a linearly decreasing function of the mutual repeat distance. In the case which is more interesting for our purposes, i.e. the case in which some positions in the linear lattice are occupied by modified versions of the n-mer, these positions will not be recognised in the search for n-mer locations. Accordingly, there will be variations in the occupancy of distance classes, which will be reflected in deviations from the monotonous decrease of the peaks' heights. The statistical significance of these deviations is easily tested by interpolating a straight line through the series of relevant peaks, and by calculating the mean square errors on the straight line parameters. The presence of peaks exceeding three mean square errors suggests the existence of supraperiodicities.

The program functions as follows:

a – the user ascribes a value to n (e.g. n = 3 or n = 4)

b – for each of the n-mers progressively found along a given sequence:

b.1 – the n-mer positions along the sequence are identified

b.2 – the distances between all pairs of positions are calculated

b.3 – the distance frequency distribution is plotted

b.4 – the linear distribution (equation 1), to be expected if the same number of n-mer copies were randomly distributed along the sequence, is drawn on the same plot

b.5 – (only in relevant cases, as decided by the operator) the best straight line is interpolated through a given series of peaks corresponding to multiples of a basic periodicity; the best-fit line, with its confidence limits is plotted, and peaks which significantly exceed these limits are identified.

A rapid inspection of the frequency distribution plots usually allows one to identify those oligos whose distributions exhibit some regularity. Work is in progress on the development of an automatic identification system which will sort out only the most regular oligos. Empirical criteria may help in reducing the work only to those n-mers which appear particularly significant, or to the most abundant ones.

Observed periodicities may all be exact multiples of a basic distance, or may be related through some linear combination of different short-range periodicities. Case by case analysis of their mutual relationships allows the user to create a virtual repeat-unit, to which the whole repetitive region can be easily compared. While it is relatively easy to carry out such a comparison automatically, the computer printout indicating only differences with respect to the virtual unit, the fully automatic reconstruction of this unit would require more sophisticated systems of artificial intelligence.

At the present stage of elaboration of the program, the user is thus required to perform the following operations:

- choose the length (n) of the oligomers to be tested;
- decide whether all possible n-mers should be tested, or only those obeying some criterion derived from other knowledge;
- select series of peaks, corresponding to multiple distances, on which the statistical test for the significance of supraperiodicities is to be performed;
- introduce the reconstructed virtual unit, to which the whole array will be compared;

The program provides printed outputs for the plotting of the distance frequency distributions (see step b.3), for the supraperiodicity test (see step b.5), and for the final alignment to the virtual repeat. This third output allows the user to immediately locate supra-repeats and to verify their correspondence to the supraperiodicities identified on a statistical basis under step b.5.

The program was written in FORTRAN 77 and implemented on a VAX 11/780 computer, connected to a PGPLOT graphic interface.

THE LATENT PERIODIC PATTERN IN A SUBTELOMERIC SEQUENCE FROM *Plasmodium falciparum*

pPftel.1, the 3 Kb clone studied by Vernick and Mc Cutchan (15), was isolated in a search for hypervariable chromosomal regions. The presence of 7bp repeats typical of Plasmodium telomeres (16,17) showed unequivocally that it had been derived from a chromosomal extremity. Within the 3 Kb insert, the Authors identified a 864 bp region described as containing complex repeats. This 'complex repeat' region is adjacent on the distal side to the telomeric structure (approximately 1.3 Kb long) and on the proximal side to a 793 bp, apparently non-repeated region.

The 'complex repeat' region was screened by the Authors for 100% internal homologies. As a result of their computer analysis, they identified a block consisting of four copies of a 41 bp repeat (A-repeat) and a second block of four copies of a 27 bp repeat (B-repeat). Furthermore, multiple copies of a string of 18 bp bridging the head and the tail of two consecutive A repeats were found to be scattered without apparent order all over the 'complex repeat' region. The same was true of a 13 bp sequence spanning the junction between two consecutive B repeats.

The Authors suggest that these 18 and 13 bp sequences, bridging larger repeats but existing also independently of them, could play a role in sequence rearrangements, by acting as 'recombinogenic sequences', promoting their own amplification and dispersal.

A search for internal homologies, performed using the matrix method and allowing for 20% mismatch, enabled us to recognize periodicities of 14 and 28 bp in the 'complex repeat' region. To investigate the relationships between the various repeat units more deeply, and possibly to discover a hidden pattern, we employed the ENHANCE algorithm. The analysis was carried out separately for the 'complex repeat' region (from nucleotide 794 to 1657) and for the non-repeated region (from nucleotide 1 to 793) of pPftel.1.

Analysis of the 'complex repeat' region of pPftel.1

Tetranucleotides showing significant deviation from the random linear distribution are presented in fig.2.

The absence of background noise is noteworthy. It suggests that there exists an extended order, which leaves no space for random occurrence of the same oligonucleotides over the whole (864 bp) region. Clear periodicities appear in the distributions: a pure periodicity of 14 bp appears in the distribution of the AGAG and GAGT strings, which also exhibit a significant 42 bp supraperiodicity (fig.3). The 14 bp periodicity is well marked also in the distributions of the AGAA, TAAG, AGTA and GTAA strings, all of which also contain identical sub periodicities (4,6,8 and 10 bp) manifested as, minor, but distinct peaks. The similarity of these distributions suggests that the various tetranucleotides are reiterated in phase, i.e. that they are linked in a longer repeated string, the slightly different profiles most probably reflecting local base substitutions. The obvious reconstruction of the ideal, reinforced pattern, satisfying all periodicities observed, is:

```
..... TAAG AG TAAG TAAG TAAG AG TAAG TAAG......
               4|_____|_____|4
   |__________|6
               |__________|8
   |_______________|10
   |________________________|14
```

resulting from the tandem reiteration of a 14 bp virtual unit. The 18 bp element identified by Vernick and Mc Cutchan (15) corresponds the addition of a fourth TAAG unit to the virtual repeat thus identified.

Fig.4 gives the alignment of the whole complex region to the virtual repeat. Perfect supra periodicities of 42 bp, whose existence was suggested by the statistical test in fig.3, are easily identified in this presentation. Supra-repeats of 28 bp also become evident, though their contribution to the frequency distributions of the particular set of tetranucleotides presented in fig.2 is not statistically detectable. (It can be detected, for example in the GACC distribution; data not shown). Supra-repeats indicated in fig.4 as A and B correspond to the two blocks of repeats already described (15) as containing units of 41 and 27 bp, but the units are actually 3 × 14 = 42 bp and 2 × 14 = 28 bp long, respectively. Note that the different supra-repeats tend to be contiguously arranged in different portions of the 'complex repeat' region.

The 14 bp, ideal repeat TAAGAGTAAGTAAG clearly results from the fusion of three simpler (TAAG) units, the first unit being followed by two base pairs (AG), most probably originated by a slipped replication event. The whole pattern thus seems to have evolved from an original simple array of tandem TAAG units in which the duplication of an AG pair, followed by amplification of (and/or homogenization to) the new repeat unit, created an array of 14 bp repeats:

```
....|TAAG |TAAG |TAAG |TAAG |TAAG |TAAG |TAAG .....
....|TAAG AG TAAG TAAG |TAAG AG TAAG TAAG |.....
```

Subsequent repeat diversification and homogenization processes may have further complicated the pattern, hiding its regularity. This reconstruction also shows that the 18 bp repeats which were suggested (15) to act as recombinogenic elements, are simply conserved parts of an earlier repetitive pattern. They can be easily detected in fig.3, whenever a perfect 14-mer is followed by a conserved TAAG unit.

Analysis of the non-repeated region of pPftel.1

Although current homology programs do not reveal any significant internal homology in this region, our algorithm is effective in detecting a significant periodicity of 14 bp in the frequency distribution of TAAG, the most abundant tetranucleotide in this as in the 'complex repeat' region. Peaks corresponding to 14 and 42 bp stand out significantly from a non-structured background as shown in fig.5.

A likely interpretation of this observation is that diversification prevailed over homogenization in this part of the sequence and blurred the regular array. A closer inspection reveals that imperfect 14 bp repeats extend into the non-repeated region for about one third of its length.

A HIERARCHY OF PERIODICITIES IN SUBTELOMERIC SEQUENCES FROM *S. cerevisiae*

Several chromosomes of *S.cerevisiae* bear, in subtelomeric position, a tandem array (up to 4 copies) of a highly conserved sequence (6.7 Kb in length) known as the Y'region (18–20).

Within the Y'region, Horowitz and Haber (1) identified a series of 12 imperfect copies of a 36 bp repeat, which are arranged in tandem. The 36 bp unit itself can be subdivided into three sub-repeats 12 bp long. The 12 bp units, note the Authors, are much more variable than the 36 bp unit, but their variation pattern shows clear correlations, the degree of homology among dodecamers that appear in I^{st} (or II^{nd} or III^{rd}) position being higher than that between dodecamers belonging to the same 36 bp unit. The latter thus appears to have originated from the fusion of three variants of a prototype dodecamer. In effect, Horowitz and Haber (1) are able to derive not only separate consensus sequences for dodecamers of type I,II and III, but also a general consensus sequence for a unified dodecamer (fig.6).

The periodic structure is certainly much more evident here than

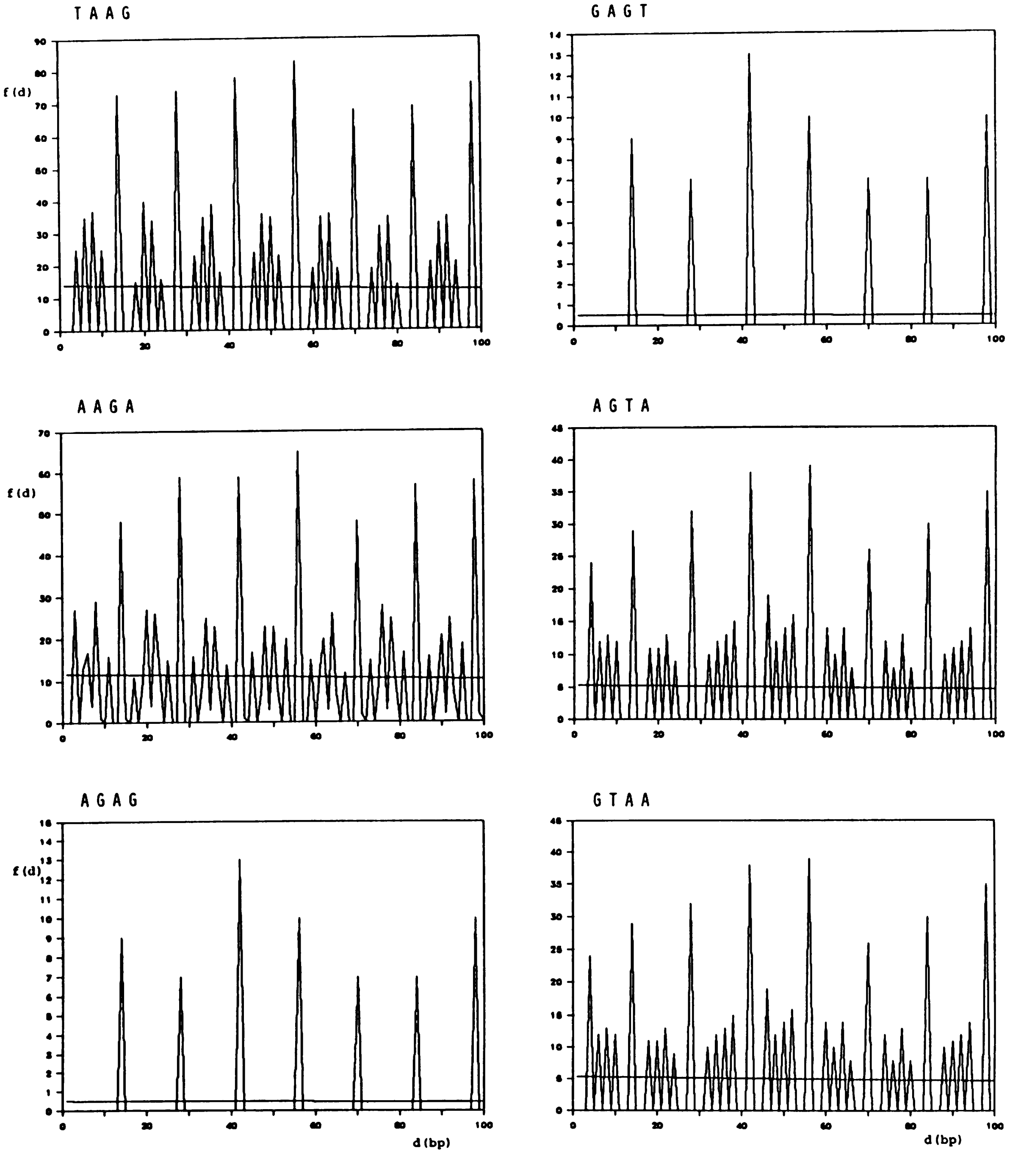

Figure 2. Distance frequency distributions for a set of tetranucleotides (TAAG, AAGA, AGAG, GAGT, AGTA, GTAA) exhibiting clear regularity in their occurrence along the 'complex repeat' region of pPftel.1 (16). The line drawn in each distribution plot is calculated by equation (1) using the actual frequency of each tetramer in the given sequence (0.13, 0.12, 0.02, 0.02, 0.08, 0.08 respectively).

in the case discussed in the previous section, the 12 bp and 36 bp repeats being discernible by simple inspection. Nevertheless, application of our algorithm yields further indications concerning latent shorter periodicities and their evolution into a complex pattern.

A first suggestion of the existence of shorter repeat units comes from the observation of internal redundancies within dodecamer units. From the consensus sequences reported in fig.6 it can be seen that the CTA triplet appears three times in dodecamer I, while dodecamer II and III can often be subdivided into two hexamers beginning with CCA.

We decided to use our ENHANCE algorithm to detect possible latent periodicities in the 12×36=432 bp of this imperfectly repetitive sequence. We started with the analysis of distance distributions for triplets such as CCA or CTA, which are reiterated within dodecamers, or such as CTG, which appears only once in the 36 bp unit (being present, and conserved, only in type II dodecamers), or such as TGC, which can be present in two consecutive dodecamers (being always present in dodecamers of type I, and 7 out of 12 times in dodecamers of type II).

Fig.7 shows the results, compared with expectations based on

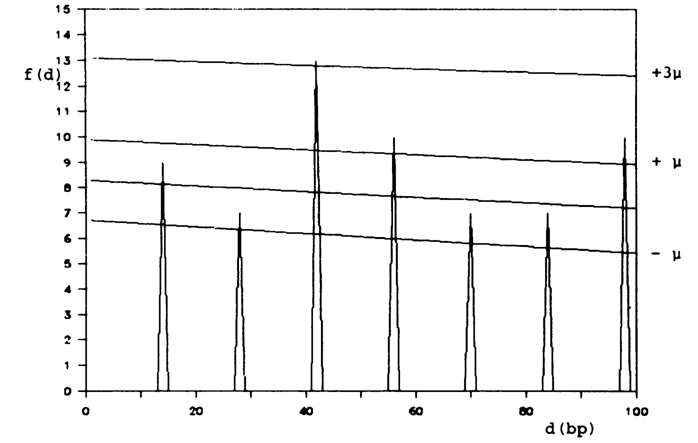

Figure 3. The statistical significance of the deviation from monotonous decrease of the peak heights is tested, as described in the text, by interpolating the 14n peak series. The peak corresponding to d=42 bp, exceeding three mean square errors (µ) on the best fit line, indicates the existence of a 42 bp supra-periodicity in the distribution of AGAG tetranucleotides in the pPftel.1 'complex repeat' region (15).

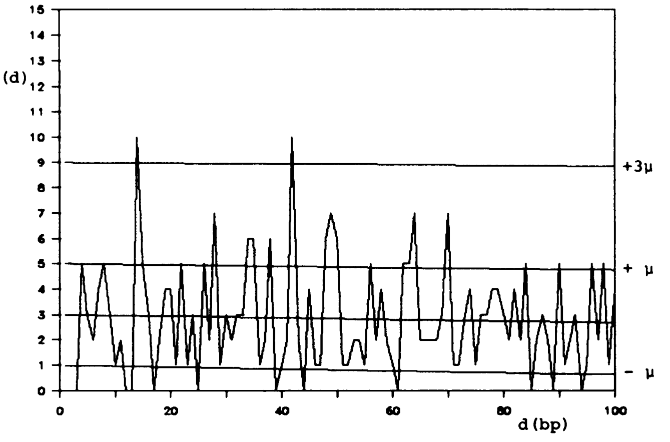

Figure 5. Statistical test for the significance of the 14 bp periodicity in the TAAG distribution along the 'non-repeated region' of pPftel.1 (16). The best fit line interpolated through all d classes is plotted with +/− 1 and + 3 mean square errors (µ).

```
       TAAG AG TAAG TAAG

794    .... .A G... A...
     [ .... .. .... G..T
A    [ G..A .. .... A.GA
     [ .... .. .... ....
     [ .... .. .... G..T
A    [ G..A .. .... A.GA
     [ .... .. .... ....
     [ .... .. .... G..T
A    [ G..A .. .... A.GA
     [ .... .. .... ....
       .... .. .... G..T
     [ .... .. .... G..T
A    [ G..A .. .... A.GA
     [ .... .. .... ....
       .... .. .... A.GA
       .... .A G... A...
       .... .. .... ....
B    [ .... .C .... A..T
     [ G..A .. .... A.GA
B    [ .... .A G... A..T
     [ G..A .. .... A.GA
B    [ .... .A G... A..T
     [ G..A .. .... A.GA
B    [ .... .A G... A..T
     [ G..A .. .... A.GA
       .... .. .... ....
       .... .. .... G..T
       G..A .. .... A.GA
       .... .A G... A...
       .... .. .... ....
       .... .C .... A..T 1227
```

```
       TAAG AG TAAG TAAG

1228   G..A .. .... A.GA
       .... .. .... ....
       .... .C .... A..T
       G..A .. .... A.GA
       .... .A G... A...
       .... .. .... ....
       .... .. .... G..T
       G..A .. .... A.CA
       .... .. .... ....
       .... .. .... ....
       .... .C .... A..T
       G..A .. .... A.GA
       .... .. .... ....
       .... .. .... ....
       .... .C .... G.CC ]
       .... .A C... A... ] c
       .... .C .... G.CC ]
       .... .A C... A... ] c
       .... .A .... G.CC ]
       .... .A .... A..T ] d
       .... .C .... G.CC ]
       .... .A .... A..T ] d
       .... .C .... G.CC ]
       .... .A C... A... ] c
       .... .C ...A G.CC ]
       .... .A .... G... ] e
       .... .C .... G.CC ]
       .... .A C... G..T ] e
       .... .C ...A G.CC ]
       .... .A .... G... ] e
       .... .C .... 1657
```

Figure 4. Alignment of the whole (864 bp) 'complex repeat' region of pPftel.1 (15) to the virtual repeat unit (boxed). Capital letters (A and B) identify supra-repeats corresponding to those described in the original paper (15). Low-case letters (c,d,e) identify a group of inter-related 28 bp supra-repeats not noticed before. Mismatches are underlined. Imperfect supra-repeats containing more than 3 mismatches are not indicated.

I	C T A $\genfrac{}{}{0pt}{}{A}{G}$ T G C T A C T A
II	C C A C T G $\genfrac{}{}{0pt}{}{A}{C}$ $\genfrac{}{}{0pt}{}{A}{C}$ A G $\genfrac{}{}{0pt}{}{T}{C}$ A
III	C C A A C $\genfrac{}{}{0pt}{}{T}{G}$ $\genfrac{}{}{0pt}{}{C}{T}$ C A G $\genfrac{}{}{0pt}{}{C}{G}$ A
General	C $\genfrac{}{}{0pt}{}{T}{C}$ A $\genfrac{}{}{0pt}{}{A}{C}$ $\genfrac{}{}{0pt}{}{T}{C}$ G C $\genfrac{}{}{0pt}{}{T}{C}$ A $\genfrac{}{}{0pt}{}{G}{C}$ $\genfrac{}{}{0pt}{}{T}{C}$ A

Figure 6. Consensus sequences reported (1) for dodecameres in first (I), second (II) and, third (III) position in the 36 bp repeat region contained in the Y' sequence of *S.cerevisiae*. As noted by the Authors (1) they all conform to a general consensus (last line).

a random distribution of the same triplets. Sub-periodicities of 3, 6 and 9 bp clearly appear in the frequency distribution of CTA, along with their combinations 18, 21, 27, 30, and 33, a much higher peak marking the 36 bp periodicity. By comparing the distributions of CCA and TGC, one notices the absence in the latter of periodicities 6, 18, 30, 42, etc (i.e. 6+12n with n=1,2,3...) which are instead present in the CCA distribution, while the 12n (n=1,2,3,...) periodicity is common to both. The CCA distribution also suggests a supra-periodicity of 108 bp, whose statistical significance is demonstrated in fig.8a. As expected, the CTG distribution exhibits a pure 36 bp periodicity. This distance (as well as its multiples) is marked in all the other examined distributions by peaks significantly exceeding the linear interpolation through the other peaks of the 12n (n=1,2,3,...) series (see for example the TGC distribution in fig.8b).

In the reconstruction of a possible pathway leading to the formation of the complex pattern, special attention has to be paid to short range periodicities, which may be remnants of an original array of simple sequences. The 3, 6 and 9 bp periods in the CTA distribution strongly suggest an original simple repetition of CTA triplets. The schemes in fig.9 illustrate possible alternatives for

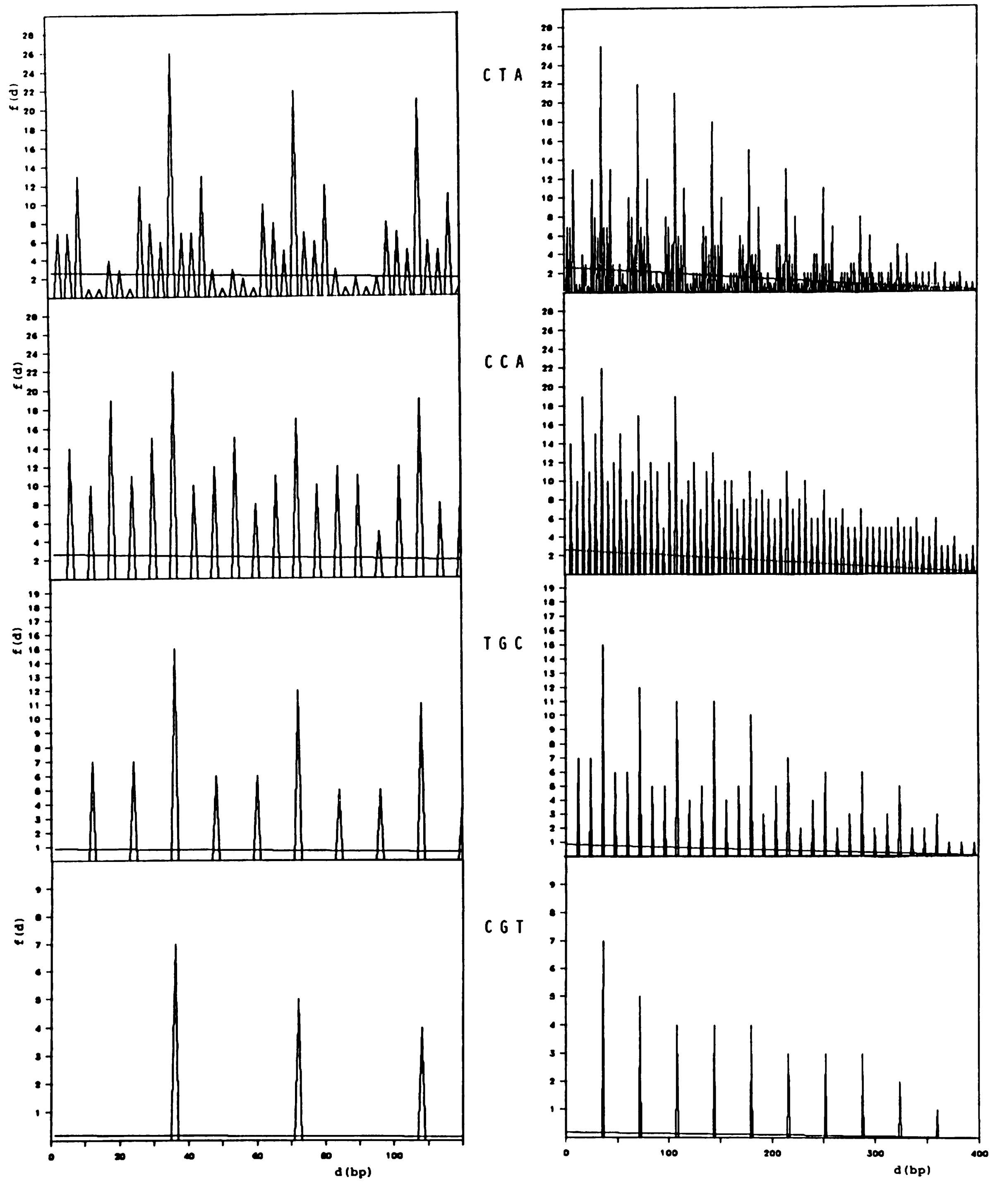

Figure 7. Distance frequency distribution for a selected set of trinucleotides (see text) in the 36bp-repeat region of the *S.cerevisiae* Y' unit (1). Each distribution is also plotted on an extended range of distances (right-hand panels). Straight lines calculated by equation (1) on the basis of actual frequencies (0.08 for CTA, 0.08 for CCA, 0.04 for TGC, 0.02 for CTG) are also plotted.

the series of modification / fusion / modificat-ion steps which may have led to the observed consensus sequences for dodecamers of type I,II and III. Note the appearance of the 6 bp periodicity for CCA before diversification between type II and III dodecamers.

Having recognised the possibility of a common origin of the various dodecamers, anyone of them may be taken as a virtual unit for the alignment of the whole (432 bp) sequence and for the identification of actual supra-repeats. Fig.10 shows the alignment to the dodecamer CCA CTG CCA GTA, chosen as a possible ancestor of type II and III dodecamers.

Inspection of fig.10 readily reveals the presence of supra-repeats: these involve, as expected, groups of three dodecamers, but larger units as well. In effect the two groups indicated by brackets in fig.8 and consisting of nine dodecamers, differ by just 1 out of 108 bp and are clearly responsible for the 108 bp

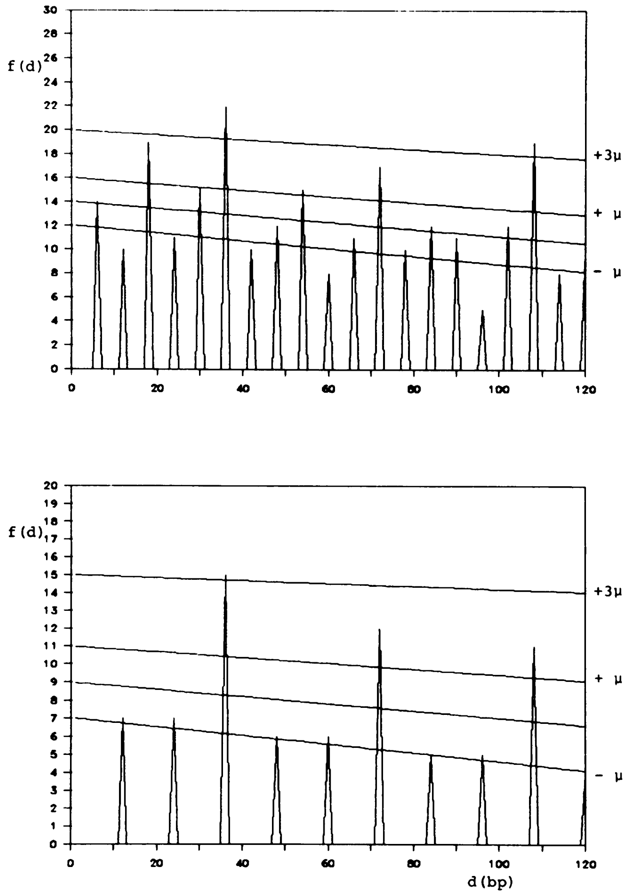

Figure 8. Statistical significance test of: a) 36 bp and 108 bp supraperiodicities in the CCA distribution of fig.7 and b) 36 bp supraperiodicity in the TGC distributions of fig.7. Similar tests for the other distributions of fig.7 indicate that, in all of them, the 36 bp peak exceeds 3 mean square errors (μ) on the line interpolated through the 12n peak series.

```
        |CTA|CTA|CTA|CTA|CTA|CTA|CTA|CTA|
                        ↓
        |CTA CTG CTA CTA|CTA CTG CTA CTA|
               *                *
                        ↓
              A                A
 [ I ]  |CTA  TG CTA CTA|CTA  TG CTA CTA|
              G                G
              *                *


        |CTA|CTA|CTA|CTA|CTA|CTA|CTA|CTA|
                        ↓
        |CCA CTA|CCA CTA|CCA CTA|CCA CTA|
          *       *     ↓ *       *
                        ↓
 [ II ] |CCA CTG CCA GTA|CCA CTG CCA GTA|
               *     *  ↓      *     *
                        ↓
                        ↓
 [ III ]|CCA ACG CCA GCA|CCA ACG CCA GCA|
             **      *       **      *
```

Figure 9. Alternative evolution pathways possibly leading from a primitive simple array of tandem CTA units to the dodecamer units I, II and III present in the 36 bp repeat region of the *S.cerevisiae* Y′ unit (1). Asterisks mark single substitution events occasionally occurring in subsequent steps (indicated by arrows). Boxed latin numbers refer to the consensus sequences given in fig.6.

```
    | CCA CTG   CCA GTA |

106.T. A..   ... ...
... A..      .G. C..
... ACT      ... .C.
.T. A..      .T. C..  ┐
... ...      ... .C.  │
... AC.      T.. .G.  │
.T. G..      .T. C..  │
... ...      ... .C.  │ 108 bp
T.. AC.      T.. .G.  │
.T. G..      .G. C..  │
... ...      AA. ...  │
... ACT      ... .C.  ┘
.T. A..      .T. C..  ┐
... ...      ... .C.  │
... AC.      T.. .G.  │
.T. G..      .T. C..  │
... ...      ... .C.  │ 108 bp
T.. AC.      T.. .G.  │
.T. G..      .G. C..  │
... ...      AA. ...  │
... ACT      ... AC.  ┘
.T. G..      .T. C..  ┐
... .C.      AA. ...  │ 36 bp
..G ACT      ... AC.  ┘
.T. G..      .T. C..  ┐
... .C.      AA. ...  │ 36 bp
..G ACT      ... AC.  ┘
.T. G..      .T. C..  )
... ...      .T. .C.  )  36 bp
... ACT      ... .C.  )
.T. A..      ... C..  )
... ...      .T. .C.  )  36 bp
... ACT      ... .C.  )
.T. A..      ... C..
... ...      AA. ...
... AC.      .T. ..G 537
```

Figure 10. Alignment of the whole 36 bp repeat region (432 bp in total) to the virtual 12 bp repeat (boxed). Mismatches are underlined. Brackets identify nearly perfect supra-repeats, carrying not more than 2 mismatches.

periodicity apparent in the CCA frequency distribution. Below the two groups of 108 bp there are two tandem copies of identical 36 bp units, followed by two more copies different from the preceding ones, and almost identical to each other (2 mismatches out of 36 bp).

Only the first and the last group of three dodecamers in the array differ from the others, though both are still 70−80% homologous to the virtual unit.

Thus the hierarchy of repeats, as revealed by our approach, extends from 3 to 108 bp, in several steps (3,6,12,36,108).

DISCUSSION

When closely inspecting DNA sequences formed by the tandem reiteration of imperfect repeats, it is not infrequent to find regular patterns in the distribution of repeat variants. Such regularity is surprising because one would expect, rather, that random base substitutions would, if tolerated, inevitably lead to the diversification of individual repeats and to a progressive blurring of what had originally been a repetitive pattern. Instead, correlations in repeat variation often lead to the formation of longer, perfect suprarepeats, as if an elaboration of the repetitive pattern were taking place. Sometimes the final complexity of the pattern is such as to hide its periodic features.

The work described in the present paper aims at enhancing such regular patterns, which are not readily detectable using the methods currently available. Latent periodicities are reinforced by means of an algorithm based on the frequency distribution of distances between oligonucleotides which are present many times in a given sequence.

The application of this simple algorithm to subtelomeric sequences from *Plasmodium* (15) and yeast (1) is instructive. In the first case, a clear pattern with a fundamental periodicity of 14 bp emerges, which had gone completely unnoticed by the Authors (15). Both in this and in the second case (where periodicities of 12 and 36 bp were already evident) our computer algorithm revealed an extended range of periodicities, starting from very short ones (4 and 3 bp, respectively) and including multiple or combined periodicities.

By taking into consideration the various elements of information it was possible to reconstruct how these short-range periodicities were connected in a virtual repeat, whose reiteration, as such or in slightly modified versions, could represent the whole repeated region. In both cases, alignment of the repetitive region with this virtual repeat showed that groups of modified repeat units (or supra-repeats) tend to be repeated in tandem, creating the supraperiodicities which had been identified through the statistical analysis. The virtual repeat itself can originate as a suprarepeat, as a result of the fusion of shorter, basic units.

The general picture thus appears to support a model in which a string, formed by the merging of different versions of the same repeat (randomly modified or unmodified with respect to a prototype), is propagated in tandem by some repeat amplification and/or homogenization mechanism. This model would explain several features often found in repetitive regions, namely the fact that long repeat units are often internally imperfectly repetitious, and the fact that, when repeats of different length coexist, they are generally not interspersed but occupy separate portions of the repetitive region. According to this model, short-range periodicities present in complex patterns would represent remnants of an original, simpler array, which evolved by modification and fusion steps.

One possible justification for the existence of mechanisms leading to the multiplication of the more recent fusion units is that such mechanisms would counteract the normal process of repeat diversification. These mechanisms should be active at least in those cases where the conservation of a repetitive structure is more important than the conservation of a definite repeat sequence. This could be true in particular for subtelomeric regions, which are known to be prone to rearrangements and polymorphisms (21−25), or for intragenic repeats coding for repetitive epitopes in plasmodial antigen genes. In the latter case, in effect, the rapid inter- and intra-specific divergence, together with faithful conservation within a given array , most probably responds to the need to escape the host's immune response through rapid change of repeated immunogenic determinants (2,4,6,7). In a notable case (the S-antigen gene of *P.falciparum* strain K1) two coexisting repeat sets (12 and 15 bp respectively, both internally repetitious, as if they shared a common fusion origin) are related to each other by a frameshift event which completely alters the codified epitope (26). This situation is clearly possible only in the absence of phenotypic selection rules based on a specific sequence in the translation product. In such conditions the nucleotide sequence is free to drift under the opposing actions of random base substitutions or losses (leading to repeat diversification) and of a tendency to maintain a repeated structure through some repeat homogenization mechanism.

This interplay may lead to the appearance of very complex patterns, in which the original simplicity (27) is barely discernible. The method we propose is particularly useful in such cases.

Another application of the proposed algorithm is the detection of phased groups of nucleotides related to variations of the helical parameters, such as Adenine runs spaced 10−11 bp, involved in the bending of the helix axis.

The method might prove useful also for multiple comparisons among sequences from different genome locations (e.g. short interspersed repeats) or even from different sources (e.g. eukaryotic promoters). By merging them into a single composite sequence (care being taken to combine sequences of the same length) our algorithm might be applied to identify phased motifs and possibly to extract the virtual consensus sequence. Such a procedure transforms multiple parallel alignments and comparisons into a linear search for series periodicities.

As a final, general comment, extreme caution should be recommended when attempting to draw from those features which are well conserved in an otherwise imperfect, tandemly repetitive pattern, conclusions concerning particular constraints or functions. If repeat amplification and/or homogenization mechanisms are active in repetitive pattern evolution, conservation of a 'motif', as in kaleidoscopic images, can be as random as its modification, and should not necessarily be considered as implying a biological function or significance.

ACKNOWLEDGEMENTS

Thanks are due to Prof. C.Saccone for stimulating discussion and for critical revision of the manuscript. The work received support from EEC in the framework of the Programme 'Science and Technology for Development' (contract no. TS2-0082).

REFERENCES

1. Horowitz, H. and Haber, J.E. (1984) Nucleic Acids Res., **12**, 7105−7121.
2. Favaloro, J.M., Coppel, R.L., Corcoran, L.M., Foote, S.J., Brown, G.V., Anders, R.F. and Kemp, D.J. (1986) Nucleic Acids Res., **14**, 8265−8277.
3. De la Cruz, V.F., Lal, A.A. and Mc Cutchan, T.F. (1987) J.Biol.Chem, **262**, 11935−11939.
4. Arnot, D.E., Barnwell, J.W., Tam, J.P., Nussenzweig, V., Nussenzweig, R.S. and Enea, V. (1985) Science, **230**, 815−818.
5. De la Cruz, V.F., Lal, A.A, Welsh, J.A. and Mc Cutchan, T.F. (1987) J.Biol.Chem., **262**, 6464−6467.
6. Godson, G.N., Ellis, J., Svec, P., Svhlesinger, D.H. and Nussenzweig, V. (1983) Nature, **305**, 29−33.
7. Galinski, M.R., Arnot, D.E., Cochrane, A.H., Barnwell, J.M., Nussenzweig, R.S. and Enea, V. (1987) Cell, **48**, 311−319.
8. Enea, V., Galinski, M.R., Schmidt, E., Gwadz, R. and Nussenzweig, R.S. (1986) J.Mol.Biol., **188**, 721−726.
9. De la Cruz, V.F., Lal, A.A. and Mc Cutchan, T.F. (1988) Mol.Biochem.Parasitol., **28**, 31−38.
10. Southern, E. (1975) J.Mol.Biol., **94**, 51−69.
11. Dover, G.A. (1982) Nature, **299**, 111−117.
12. Dover, G.A. and Flavell, R.B. (1984) Cell, **38**, 622−623.
13. Dover, G.A. (1987) J.Mol.Evol., **26**, 47−58.
14. Dover, G.A. (1988) Nature, **331**, 121.
15. Vernick, K.D. and Mc Cutchan, T.F. (1988) Mol.Biochem.Parasitol., **28**, 85−94.
16. Ponzi, M., Pace, T., Dore, E. and Frontali, C. (1985) EMBO J., **4**, 2991−2995.
17. Dore, E., Pace, T., Ponzi, M., Scotti, R. and Frontali C. (1986) Mol.Biochem.Parasitol., **21**, 121−127.
18. Chan, C.S. and Tye, B.K. (1983) Cell, **33**, 563−573.
19. Walmsley, R.M. (1987) Yeast, **3**, 139−148.
20. Zakian, V.A. and Blanton, H.M. (1988) Mol.Cell.Biol., **8** 2257−2260.
21. Horowitz, H., Thornburn, P. and Haber, J.E. (1984) Mol.Cell.Biol., **4**, 2509−2517.
22. Borst, P. and Greaves, D.R. (1987) Science, **235**, 658−667.
23. Corcoran, L.M., Thompson, J.K., Walliker, D. and Kemp, J. (1988) Cell, **53**, 807−813.
24. Vernick, K.D., Walliker, D. and Mc Cutchan, T.F. (1988) Nucleic Acids Res., **16**, 6973−6985.
25. Dore, E., Pace, T., Ponzi, M., Picci, L. and Frontali, C. (1990) Mol.Cell.Biol. in press.
26. Saint, R.B., Coppel, R.L., Cowman, A.F., Brown, G.V., Shi, P.T., Barzaga, N., Kemp, D.J. and Anders, R.F. (1987) Mol.Cell.Biol., **7**, 2968−2973.
27. Tautz, D., Trick, M. and Dover, G.A. (1986) Nature, **322**, 652−656.