Fast analysis of highly repetitive sequences in human chromosomes using a novel search algorithm

VICENTE ARNAU* and IGNACIO MARÍN**
* Departamento de Informática; ** Departamento de Genética
Universidad de Valencia.
Campus de Burjassot. Avda. Vicent Andrés Estellés, s/n. 46100 Burjassot. Valencia.
SPAIN
* vicente.arnau@uv.es ** ignacio.marin@uv.es

Abstract: We have developed an algorithm that allows the exhaustive determination of words of up to 12 nucleotides in DNA sequences and it is fast enough as to be used at a genomic scale running on a standard personal computer. In this study, we apply our algorithm to compare the composition of Alu highly repetitive sequences in two human chromosomes, namely chromosomes 21 and 22, each of them about 33 million nucleotides-long. We show that our method allows the quantification and comparison of hundreds of thousands of Alu sequences in a few minutes. We determine that both human chromosomes have different numbers and types of Alu sequences, in agreement with previous data. We also observe that DNA methylation has a profound impact on Alu sequence conservation. Future applications of this method are discussed.

Key-Words: Human chromosomes, Alu, repetitive sequences, nucleotide words.

1 Introduction

In a previous work [1], we described an algorithm that allows the exhaustive determination of all 12-nucleotide-long words ("12-mers") present in a given DNA sequence. We also showed that it is so fast that it can be used at a genomic scale, to analyze whole chromosomes or even genomes. As an example, we used it to detect all sequences that were highly characteristic of a particular human chromosome, appearing 50 or more times, but were absent in another human chromosome [1]. Those analyses, performed on human chromosomes 21 and 22 [2, 3], each one about 33 Megabases (33 Mb or 33 millions of nucleotides) long, took just a few minutes on a standard personal computer. We have called the program that implements this algorithm as UVWORD.

In this study, we use a modification of the UVWORD original algorithm to perform comparisons of a given relatively short sequence of DNA, that we will call from now on "master copy" against different human chromosomes. The goal of this type of analysis is to establish the relative abundance in those chromosomes of all the 12-mers contained in the master copy. The name "master copy" refers to the fact that these analyses are especially indicated to establish the number of repetitive sequences in a given chromosome or genome. Thus, the short DNA sequence that is used as starting point for the analysis can be considered a particularly significant (i.e. "master") copy that serves as guide to recognize all the rest of copies of the repetitive sequence that are present. Those sequences will be detected if they have at least 12 nucleotides in common with the master copy. Most interestingly, because we expect to find by chance a particular 12-mer only once every 16.8 Mb, we can be confident that, even when whole chromosomes or genomes are analyzed, essentially all the sequences found actually correspond to the repetitive family. For example, the analysis of the whole human genome would generate about 200 false positives (i. e. sequences with a given word, but not part of any repetitive family), and the analysis of an average human chromosome less than 10 (and even less, about 2, would be found in the small chromosomes 21 and 22). If we compare those values with the expectations for several repetitive element families, that are present hundreds of thousands or even millions of times in the genome, it is easy to conclude that the fraction of false positives is negligible.

The generation of new tools for the analysis of repetitive sequences at a genomic scale is interesting for several reasons. First, to recognize the different types of repetitive sequences and their frequences in a given genome or to performe comparisons among different genomes. Second, to determine whether particular sequences tend to concentrate in particular chromosomes or regions within chromosomes. Finally, the analysis of specific repetitive families or subfamilies may give substantial insights about the dynamics of invasion of the genome by repetitive selfish elements, as transposons. In this work, we focus on the distribution of the most abundant of the repetitive families in our genome, the Alu family (reviewed in [4]) in two of the best characterized human chromosomes, 21 and 22. Those two chromosomes, in spite of being of very similar size, are known to be very different in gene composition, G+C content and distribution of repetitive elements, including Alu elements [2, 3]. Thus, their analysis is an excellent test to establish the performance of our program.

2 Problem formulation

The DNA sequence where the search is to be made (e. g. in this case, human chromosomes 21 or 22) are exhaustively analyzed using UVWORD, as explained in [1], to determine all 12-mers and their frequencies. Then, the program is asked to screen for each 12-mer present in the master copy sequence and thus to determine the frequency of those particular 12-mers in the chromosome.

In order to demonstrate the usefulness of our method to analyze the repetitive component of complex genomes, we applied this method to detect in human chromosomes 21 and 22 all the words of length 12 found in a master copy of 84 nucleotides, that corresponds to the most conserved region of the sequence of the Alu Y subfamily (positions 2 to 87 in [5]). The precise sequence analyzed is as follows: 5′ GCCGGGCGCGGTGGCTCACGCCT GTAATCCCAGCACTTTGGGAGGCCGAGGCG GGCGGATCACGAGGTCAGGAGATCGAGAC CA – 3′. Given the high sequence conservation in that region, an important fraction of sequences belonging to the other two main Alu subfamilies, Alu J and Alu S, should also be detected in many cases (i. e. only part of the 12-nucleotide long words analyzed are Alu Y-specific) [5]. In order to detect all the sequences presents in chromosomes 21 and 22, both direct and inverse/complemented sequences of those chromosomes, corresponding to the two strands of the double helix, were analyzed and the results for both strands were added. Thus, all results shown below correspond to the sum of results for both strands.

3 Problem solution

All analyses were performed in a few seconds (speed = 1.5 Mb/sec) on a standard PC computer (Intel Pentium ® 4 CPU 2.8 GHz.). A summary of the results is shown in Figure 1 to 4. Figure 1 shows a general overview of the results of the analyses for the two chromosomes. A first obvious observation is that the ability of the program to detect Alu sequences, that depends on the similarity of the different Alu sequences that are in a chromosome with the master copy, varies significantly along the analyzed sequence. This hetereogenity is due to two main causes. First, variation in the different families of Alu sequences, that are more similar in some regions and divergent in other regions [5]. Second, the presence of point mutations that alter particular sequences, making them different from the master copy. In particular, we have detected a significant effect of the presence of CG dinucleotides on the values observed, a fact that is related with the high level of mutational variation associated to the DNA methylation that is typical of many organisms, including humans, and that convert CG dinucleotides to TG or CA dinucleotides at a high rate [6]. Thus, many of the Alu sequences cannot be detected because they have accumulated one or several CG to TG or perhaps CG to CA changes in their sequences, thus being not identical to the master copy. Comparing our data with previous results derived from the human genome project [2, 3], we determined that as much as 30 % of the Alu sequences present in chromosomes 21 and 22 are different along their whole lengths from the master copy used for our analyses. It is also evident when inspecting Figure 1 that Alu sequences are about two times more frequent in chromosome 22 than in chromosome 21, in agreement with previous data [2, 3]. It is also obvious from that figure that the general profile of conservation (shape of the curve) is similar in both chromosomes. However, in order to determine whether there is some enrichment of particular sequences in one of the chromosomes, we calculated the ratio of frequencies between the two chromosomes for each 12 nucleotide-long word. Results are shown in Figure 2. As it can be seen in that figure, the relative frequency of the sequences changes when we move from the 5′ to the 3′ end of the molecule, with the highest ratios being observed in the 5′ end. This result indicates that the two human chromosomes have a different composition of Alu sequences, with those in chromosome 21 having significantly less similar 3´ends relative to the master copy than those of chromosome 22. Previous data obtained in the human genome sequencing projects fits with our data, in that composition of Alu sequences varies with the G+C content of chromosomes, and that content is qualitatively different from chromosomes 21 (41%) and 22 (48%) [7]. Finally, Figures 3 and 4 show the results when both orientations of the same chromosome are compared. We can conclude that the two human chromosomes do not display significant signs of polarity, i. e. both strands have extremely similar numbers of Alu sequences as detected with UVWORD. The slight, non-significant differences that can be observed, especially in Figure 3 for chromosome 21, must be just an effect of sampling, becoming more evident in the chromosome that has less Alu sequences.

4 Conclusions

The method used in this study allows the exhaustive determination of words of 12 nucleotides in very large sequences and in a very short time. Thus, its use may be generalized at the genomic level. Some general applications of our algorithm are the finding of singular sequences in chromosomes or genomes (as shown in [1]), the characterization of the number and types of repeated sequences in long DNA sequences (as shown in this study) or the determination of the number of times that some significant sequences are present in one or several DNA sequences, e.g. to obtain genomic signatures. More complex algorithms for determination of words in which ambiguities are allowed have been also developed by our group and will be detailed elsewhere.

Our analyses of the Alu sequences of two human chromosomes confirm previous findings regarding their high difference in Alu content, with chromosome 22 having a density of Alu sequences that is about twice that found in chromosome 21 (Figures 1 and 2). They also allow to establish two interesting facts: the absence of chromosome polarity (i.e. Alus are identically represented in both DNA strands; Figures 3 and 4) and the significant difference of conservation of Alu sequences, respect to the master AluY copy used, when both chromosomes are compared, a difference that has a clear 5´ to 3´ biass (Figure 2). In summary, our simple procedure allows the precise study of the repetitive component of chromosomes or full genomes, yielding high quality data in few minutes. We plan not only and to further explore different master Alu sequences to obtain a clearer picture of their diversity in the human genome, but also to expand this type of analyses in order to explore the degree of similarity in the repetitive components of multiple genomes.

This paper is supported by Ministerio de Ciencia y Tecnología (Spain; projects GEN2001-4851-C06-02 and SAF2003-09506) and CICYT Grant No. TIC2000-1151-C07-04).

References:

[1] V. Arnau and I. Marín, A fast algorithm for the exhaustive analysis of 12-nucleotide-long DNA sequences: application to human genomics. *Proceedings of the 17th International Parallel and Distributed Processing Symposium,* IEEE Computer Society, 2003.

[2] I. Dunham, N. Shimizu, B. A. Roe, S. Chissoe, *et al.* The DNA sequence of human chromosome 22. *Nature,* Vol. 402, 1999, pp. 489-495.

[3] M. Hattori, A. Fujiyama, T. D. Taylor, H. Watanabe *et al.* The DNA sequence of human chromosome 21. *Nature*, Vol. 405, 2000, pp. 311-319.

[4] M. A. Batzer and P. L. Deininger, Alu repeats and human genomic diversity. *Nature Reviews Genetics* Vol.3; 2002, pp. 1-10.

[5] M. A. Batzer, P. L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C. M. Rubin, C. W. Schmid, E. Zietkiewicz, and E. Zuckerkandl, Standardized nomenclature for Alu repeats. *J. Mol. Evol.* Vol. 42, 1996, pp.3-6.

[6] M. W. Nachman and S. L. Crowell Estimate of the mutation rate per nucleotide in humans. *Genetics* Vol.156, 2000, pp. 297-304.

[7] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, Vol. 409, 2000, pp. 860-921.

Figure 1. Frequencies of appearance of words of size 12 derived from the AluY master copy in human chromosomes 21 and 22. "Final nucleotides", here and in the other figures, refer to the fact that the frequencies are estimated for words which last nucleotide is the one showed below. For that reason, the first eleven letters of the master sequence are not included in this figure.

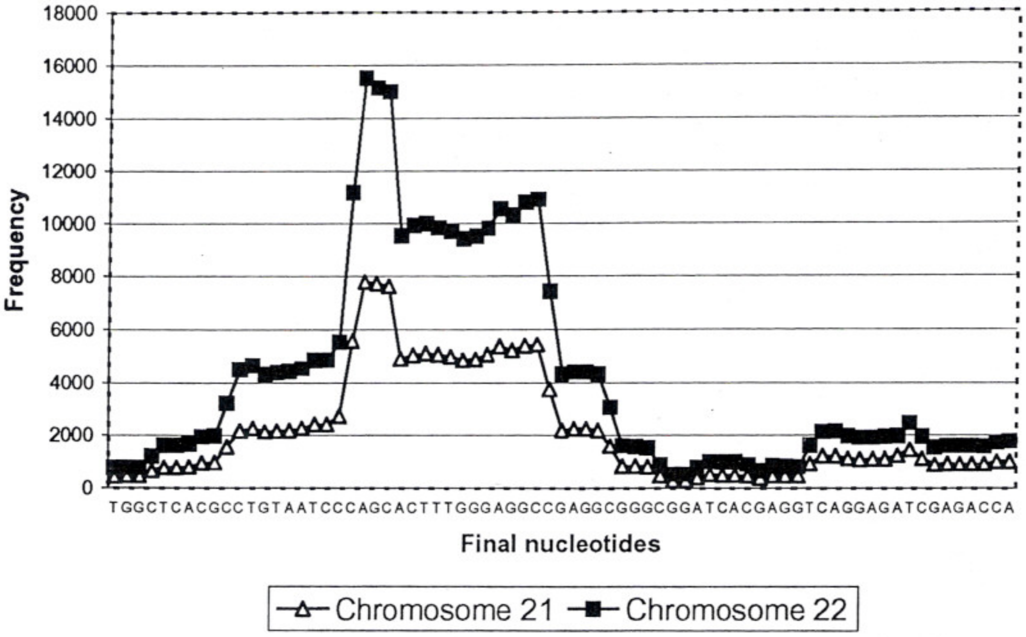


Figure 2. Ratio (Chrom. 22 / Chrom. 21) of the frequencies of 12-nucleotide-long words present in Alu sequences.

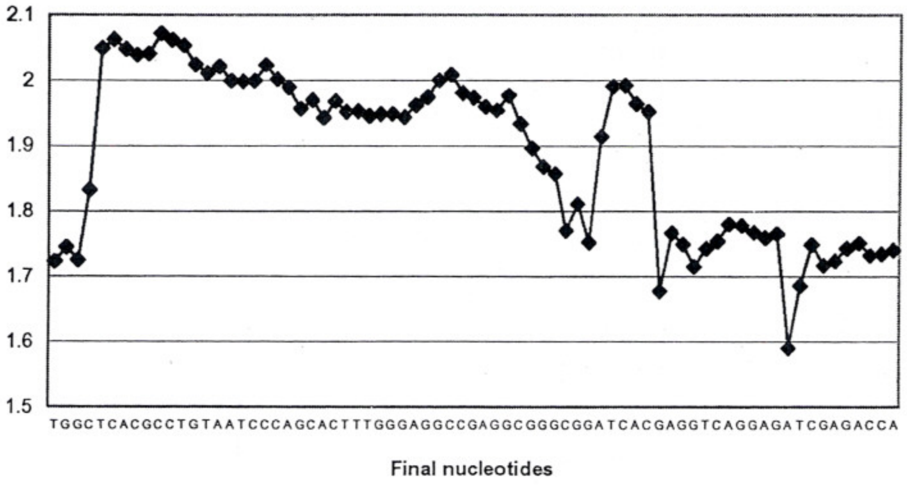


Figure 3. Frequencies of Alu sequences detected in both strands (i. e. the direct and the inverted/complemented strands) of the double helix of chromosome 21. Because the curves obtained for the two strands are so similar that they often completely overlap, white circles are generally barely visible in this representation.

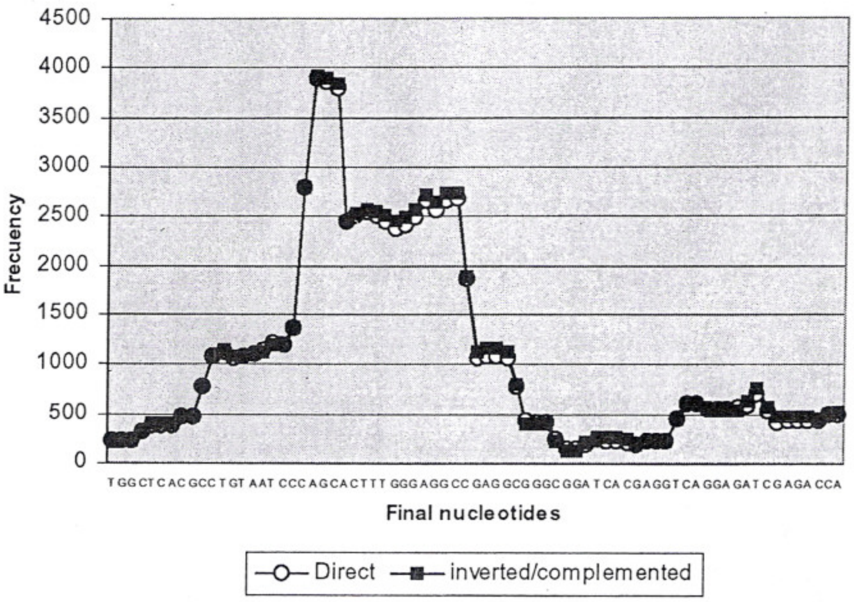
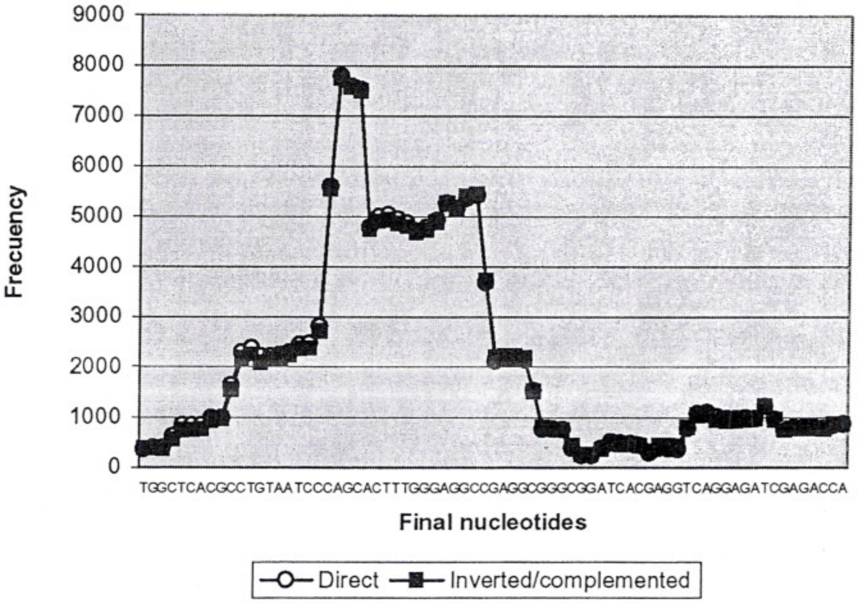


Figure 4. Frequencies of Alu sequences in both strands of chromosome 22.


Frecuency
TGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCA
Final nucleotides
Direct
Inverted/complemented