Effect of G + C content on distances

The G+C content of a sequence has an important influence on the frequency of oligonucleotides. For example, when the G+C content of a sequence is high, oligonucleotides with high G+C content are more abundant. **The aim** of this study was to determinate whether G+C content also influences the distances computed between DNA sequences.

Three different experiments were performed, but the basic procedure used in each was the same. Briefly, for each completely-sequenced genome, the oligonucleotide frequencies and oligonucleotide frequencies of randomly-selected subsequences were computed. The distances between genomes and subsequences were computed and graphically represented.

http://gscompare.ehu.eus

Experiment No. 1

GENOMES
1,234 completely-sequenced prokaryotic genomes were used in this study.

FREQUENCIES AND DISTANCES
The frequencies and distances used in this experiment are described in detail in a separate document.

The following types of frequencies were computed for **tetranucleotides** by searching **both DNA strands**: Tetranucleotide frequencies (Oligos4), tetranucleotide frequencies (Oligos4st), zero'th order Markov chain frequencies (ZOM), first order Markov chain frequencies (FOM), second order Markov chain frequencies (SOM) and Z-score values (Zscore).

To compute the distances between tetranucleotide frequencies, the following statistics were used: Standard Pearson's distance (Pd), Genomic Signature distance (GS), Euclidean distance (E).

METHOD
- From each completely sequenced prokaryotic genome, 250, 500, 1,000, 5,000 and 20,000 bp subsequences were randomly selected. For each length, 1,000 subsequences were searched.
- The tetraonucleotide frequencies mentioned above were computed for each subsequence and for the complete genomic sequences.
- The frequencies of each subsequence, as well as the frequencies of the complete genome, were compared using the distances above.
- **The median** length of each genome and subsequence was recorded. As the subsequences selected from particular prokaryotic genomes sometimes show oligonucleotide compositions that differ significantly from that of the genome (for example, when the complete subsequence, or part of it, was acquired by horizontal transfer), we believe the median is a statistic that better describes the distance than average value, which may be skewed by the presence, in the genome, of subsequences with the afore-mentioned non-characteristic frequencies. In fact, the median is a robust estimate of average value.
- The median values were graphically represented. One of these graphs is amplified and described in figure 1, to show how these values are represented.

Figure 1: The median of the Euclidean distances between the standardized tetranucleotide frequencies of 1,234 prokaryotic genomes and 1,000 subsequences of each of the genomes are represented. The subsequences were 250 to 20,000 bp long, and each length was coloured differently. Prokaryotic genomes are represented in the graph according to their G+C content.

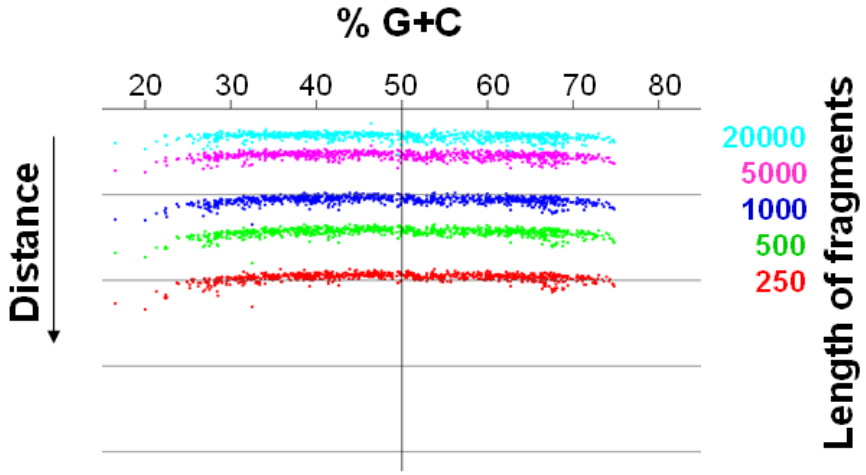
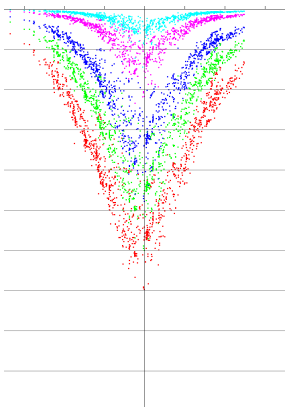
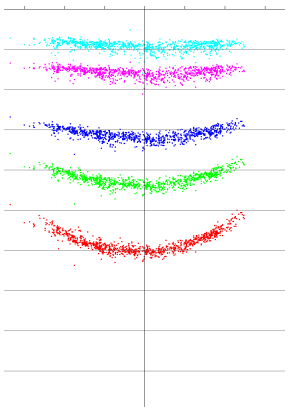
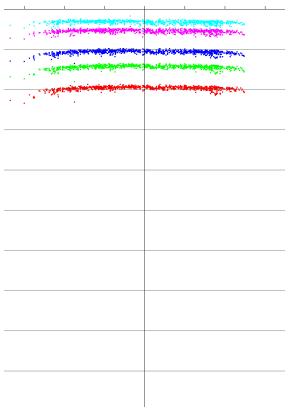
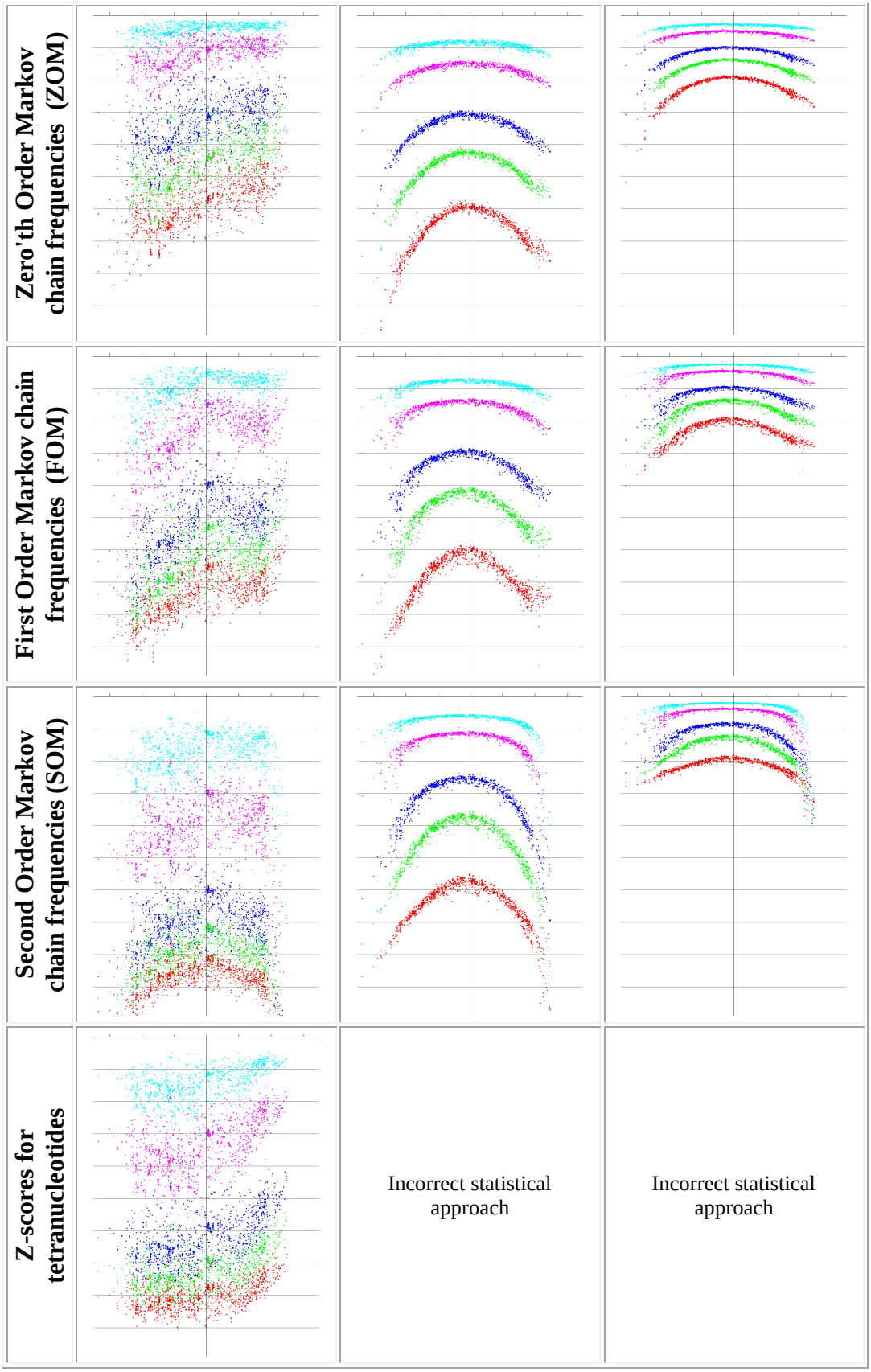


Table 1: Influence of G+C content on the distance between subsequences of prokaryotic genomes and the complete genome. Each graph in the table has the same characteristics described in figure 1. Tetranucleotided frequencies and zero'th order Markov chain frequencies were standardized prior to the computation of Genomic Signature distance and Euclidean distance.

	Pearson's distance	Genomic Signature distance	Euclidean distance
Tetranucleotide frequencies			



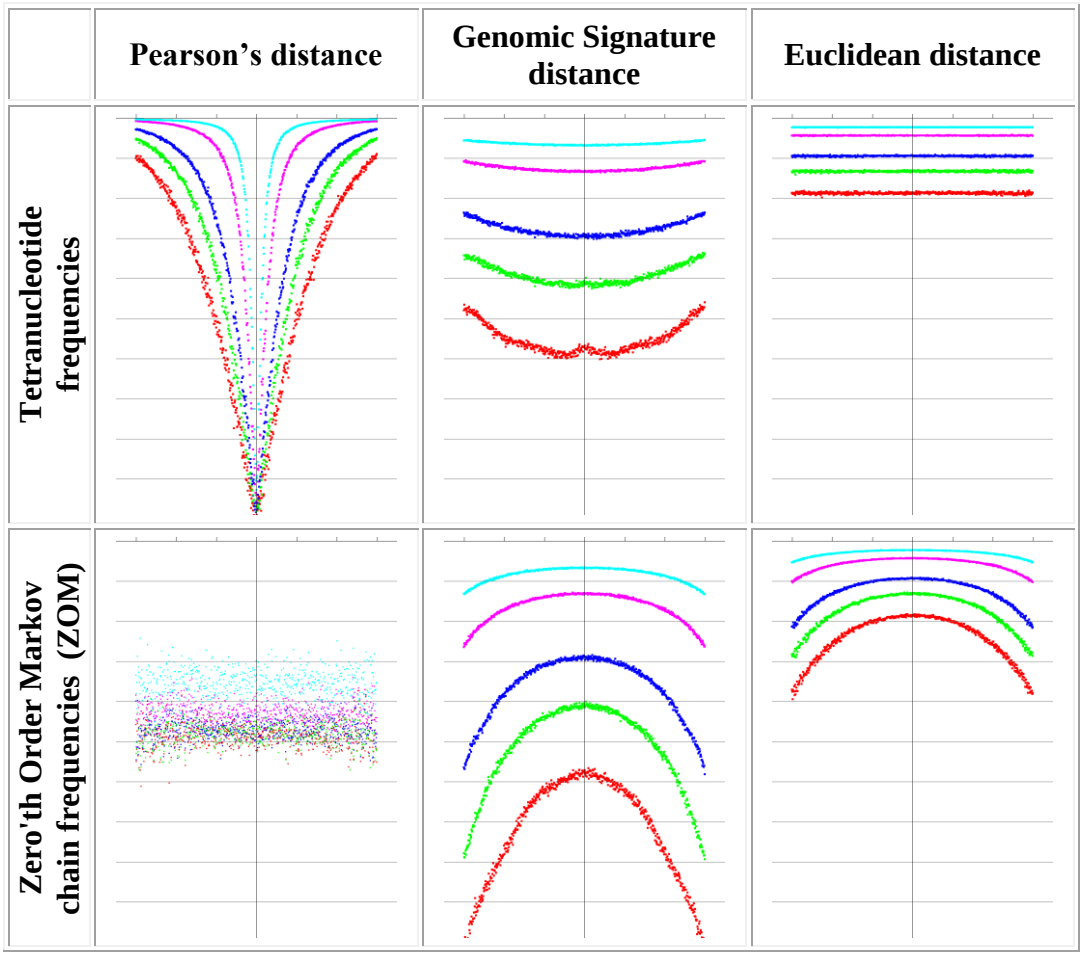
Additional types of frequencies were also searched, but the results of those searches are not included in this document: Standardized ZOM, FOM and SOM values were

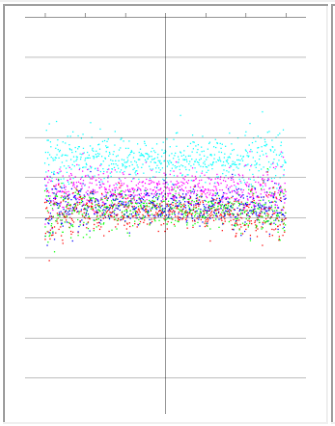
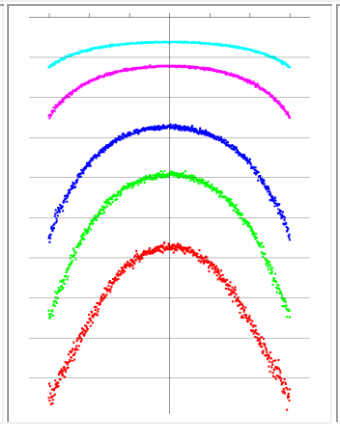
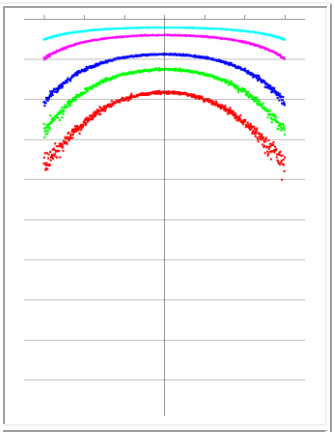
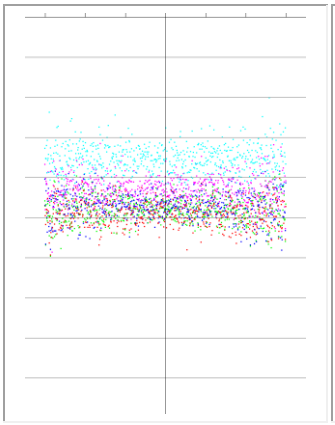
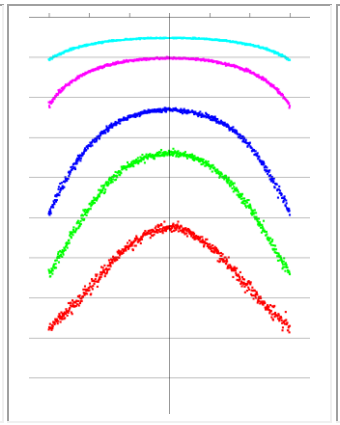
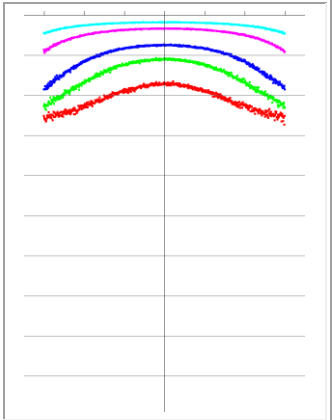
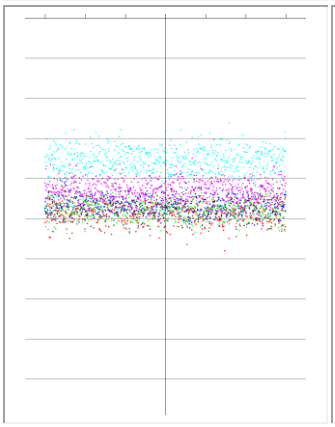
basically the same as the non-standardized values, but the figures seemed blurred, compared to the non-standardized figures.

Experiment No. 2

Random sequences with 20 to 80% G+C content were generated, and the procedure used in experiment no. 1 was applied to them. The aim of this experiment was to check whether the results shown in table 1 were similar for random sequences and for prokaryotic genomes. Results are shown in table 2.

Table 2: Influence of G+C content on the distance between subsequences of random sequences with different G+C content and the complete sequence. Each graph in the table has the same characteristics described in figure 1. Tetranucleotide frequencies and zero'th order Markov chain frequencies were standardized prior to the computation of Genomic Signature distance and Euclidean distance.



First Order Markov chain frequencies (FOM)			
Second Order Markov chain frequencies (SOM)			
Z-scores for tetranucleotides		Incorrect statistical approach	Incorrect statistical approach

Experiment No. 3

GENOMES
1,013 completely-sequenced prokaryotic genomes were used in this study.

FREQUENCIES AND DISTANCES
Both strands of sequences and subsequences were searched, and **hexanucleotide** frequencies, as well as standardized frequencies, were obtained. The Pearson's distance, Genomic Signature distance and Euclidean distance between each subsequence and the complete genome were calculated.

METHOD
- The prokaryotic genomes were sorted by G+C content (low to high).
- From each completely sequenced prokaryotic genome, 20,000 bp-long subsequences (100 subsequences) were randomly selected.
- Hexanucleotide frequencies were computed for each subsequence and for the complete genomic sequences. Standardized frequencies were also computed.
- For each genome, the distances between the genome and each of the 100 random subsequences were computed and recorded.
- One graph was generated for each type of distance.

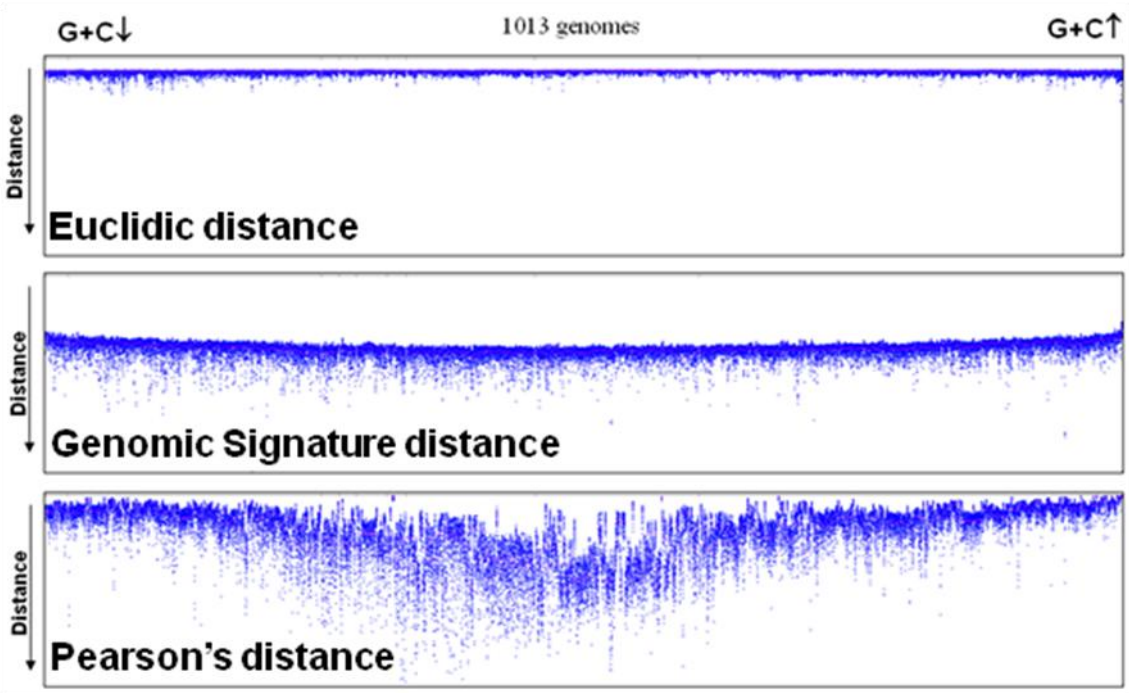


Figure 2: Influence of statistical procedure (Euclidean, Genomic Signature or Pearson's distance) and G+C content on the distances between hexanucleotide frequencies of prokaryotic genomes and 20,000 bp-long subsequences from the same genome. From left to right, 1,013 prokaryotic genomes are sorted according to G+C content. For each genome, the distances between the frequencies of the complete genome and the frequencies of 100 randomly-selected subsequences are represented with dots. All 100 dots for a given genome are located in the same vertical line, so that the relative variance of each statistical distance is also represented.

Discussion

Comparing oligonucleotide frequencies, it is apparent that G+C content influences distance. The effect depends on the type of frequencies and statistical distance used.

In our first experiment, we searched for the effects of G+C content on 13 different combinations of tetranucleotide frequencies and distances (additional combinations were removed from our results). All combinations, except for those obtained with Euclidean distance, were influenced by G+C content.

When applying Pearson's distance to oligonucleotide frequencies, we observed an important increase in distance when G+C content got closer to 50%. We repeated this experiment with random sequences in experiment no. 2, and similar results were obtained. This increase in distance is due to a very simple fact: In random sequences with a G+C content of 50%, the probabilities of all the 256 tetranucleotides are identical (all data will be located at the same point in a bidimentional space), and consequently, no correlation exists.

To our knowledge, the afore-mentioned increase in distance has not been examined by other authors, so we designed experiment no. 3 to graphically show whether the increase in Pearson's distance was related to an increase in the variance of this distance. We searched the hexanucleotides of 1,013 prokaryotic genomes and compared them to 100 subsequences from each genome using Pearson's, Genomic Signature and Euclidean distances. Though the ZOM, FOM and SOM frequencies, as well as Z-scores, of hexanucleotides and other oligonucleotides can be computed, we have not found such computations in the literature, and in fact, we believe the effort required to obtain such computations is worthless. In the graph, genomes were sorted from low to high G+C content, and the 100 distances obtained from each genome were plotted. This experiment graphically showed the increase in Pearson's distances when G+C content was near 50%. It must be pointed out that a significant percentage of the completely-sequenced genomes used in the experiment was close to 50%. With regards to Genomic Signature distance, a similar effect was detected, but it was much less evident. When it comes to Euclidean distance, G+C content did not seem to influence distance. Concerning variations in the data, the graph makes evident that variance is low for Euclidean distances, higher for Genomic Signature distances and highest for Pearson's distances.