Computation of oligonucleotide frequencies and distances

Contents:

http://gscompare.ehu.eus

Computation of oligonucleotide frequencies

In this document it is described how the scripts at http://gscompare.ehu.eus compute oligonucleotide occurrences and frequencies from a given sequence. Although there may be other ways to describe the oligonucleotide composition of a sequence, we have selected the frequencies shown in Figure 1 to create the script in the belief that they are the most used ones in the literature.

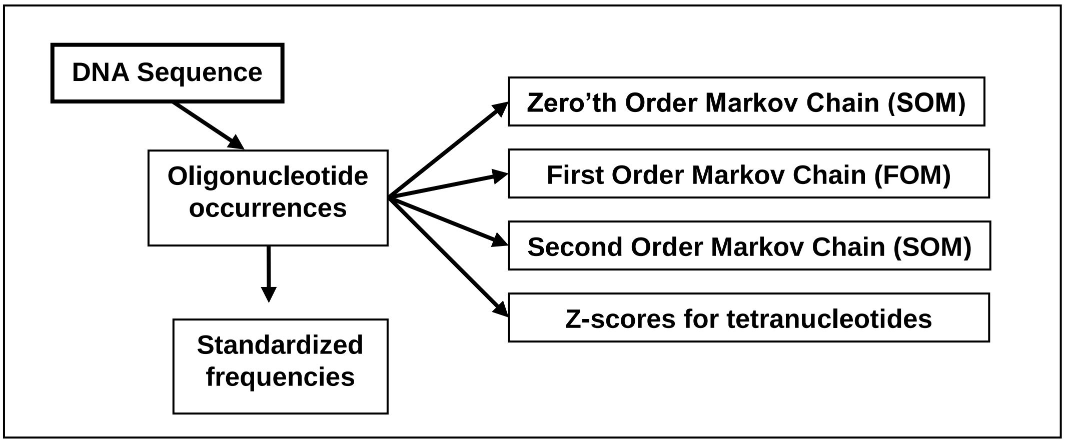
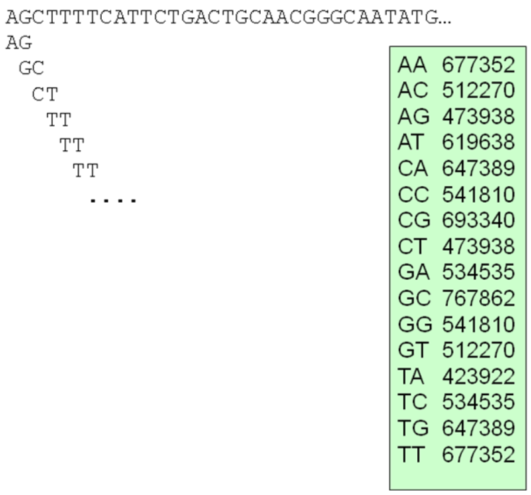

DNA Sequence
Oligonucleotide occurrences
Standardized frequencies
Zero'th Order Markov Chain (SOM)
First Order Markov Chain (FOM)
Second Order Markov Chain (SOM)
Z-scores for tetranucleotides

Figure 1: Oligonucleotide occurrences are extracted from sequence, and those data are later used to compute standardized frequencies. Tetranucleotide occurrences are used to compute SOM, FOM and SOM frequencies, and also z-scores for tetranucleotides (which are based on SOM frequencies).

Computing oligonucleotide occurrences

Many authors had used Chaos Game Representation images (CGR) or its derivative, Chaos Game Representation of [oligonucleotide] frequencies (FCGR), to describe the oligonucleotide content of sequences or genomes. Consequently, the description of the images, and also the description of the methods to compare and compute distance among CGR and FCGR images, became complicated due to the need to describe the images as two-dimensional objects. Although they are interesting representations, for practical purposes, in this document the oligonucleotide frequencies will be described as one-dimensional matrixes.

The first step is computing the number of oligonucleotide occurrences in the sequence. To compute oligonucleotide occurrences with length *k*, the sequence with length *m* will be scanned with a sliding window of length *k* along the sequence up to position m-k+1. An array/matrix *A* containing $4^k = n$ elements will be obtained, where each element f_i of the array is the number of occurrences of *i*-th oligonucleotide within the sequence.



```
AGCTTTTCATTCTGACTGCAACGGGCAATATG…
AG
 GC
  CT
   TT
    TT
     TT
      ....
```

AA	677352
AC	512270
AG	473938
AT	619638
CA	647389
CC	541810
CG	693340
CT	473938
GA	534535
GC	767862
GG	541810
GT	512270
TA	423922
TC	534535
TG	647389
TT	677352

Figure 2: Computing number of oligonucleotides for E. coli K-12. Check text above for details.

Array A mentioned in this document is equivalent to FCGR, the Frequency matrix extracted from a Chaos Game Representation (CGR) of the sequence, which was described by Almeida et al. (2001).

It is worth noting that the array containing occurrences of *k* long oligonucleotides may be used to obtain the occurrences of *k-1* long oligonucleotides within the sequence when the sliding window is run to the end of the sequence, so that the last tetranucleotides accounted will be of type NNN-, NN-- and N---, where N is a nucleotide, and “-“ denotes absence of nucleotide in that position. In that particular condition, occurrences for trinucleotide ACG are equal to the sum of occurrences of tetranucleotides ACGA, ACGC, ACGG, ACGT and ACG- (in case ACG were the last 3 nucleotides in the sequence). This property is very useful to the programmer, who may generate a script, which goes along the sequence only once to compute tetranucleotide composition (or longer oligonucleotides), and then may use that data to compute occurrences of trinucleotides and dinucleotides. At the end of the process, all oligonucleotides including gaps will be removed, so that all matrixes/arrays will be 4^k nucleotides long.

Standardized oligonucleotide frequencies

As mentioned above, oligonucleotide occurrences are computed and array A is obtained. Then, the frequencies are standardized by using the following formula:

$$f'_i = \frac{\mathrm{n}}{\sum_{i=1}^{n} f_i} f_i$$

where f'_i is the standardized frequency of *i*-th oligonucleotide, and f_i is its number of occurrences within the sequence. An array *A’* of size 4^k = n (the number of elements in the array) will be obtained where each element f'_i of the array will be the standardized frequency of a specific oligonucleotide. For standardized data, the sum of all values equals the number of elements.

ZOM, FOM and SOM frequencies

Frequencies based on Markov models are often used to describe oligonucleotide composition of sequences and for their ulterior comparison. Here we have used the notation by Bohlin & Skjerve (2009) to represent Zero'th Order Marchov Chain frequencies (ZOM), First Order Marchov Chain frequencies (FOM), and Second Order Marchov Chain frequencies (SOM).

$$\rho_{XYZW}(f) = \frac{f_{XYZW}}{f_X \, f_Y \, f_Z \, f_W} \quad \textbf{(ZOM)}$$

$$\xi_{XYZW}(f) = \frac{f_Y \, f_Z \, f_{XYZW}}{f_{XY} \, f_{YZ} \, f_{ZW}} \quad \textbf{(FOM)}$$

$$\eta_{XYZW}(f) = \frac{f_{XYZW} \, f_{YZ}}{f_{XYZ} \, f_{YZW}} \quad \textbf{(SOM)}$$

In the formulas above $\rho_{XYZW,}$ ξ_{XYZW} and η_{XYZW} are the ZOM, FOM and SOM-based frequencies for oligonucleotide XYZW. To compute them, the following must be previously computed: the number of occurrences in the sequence of tetranucleotide XYZW (f_{XYZW}), the number of occurrences of trinucleotides *(f_{XYZ} and f_{YZW})* and dinucleotides (f_{XY} , f_{YZ} and f_{ZW}) contained within the tetranucleotide, and the nucleotide composition of the sequence (f_X, f_Y , f_Z and f_W).

Arrays containing the frequencies of each of the 256 possible tetranucleotides were generated for ZOM, FOM and SOM values.

It must be pointed out that the arrays containing ZOM, FOM and SOM data will contain standardized data only for randomly generated sequences. F**or naturally occurring DNA sequences/genomes the ZOM, FOM and SOM values are not standardized values**, so that Genomic Signature Distance and Euclidean should not be applied to them. This is a big concern mostly for small DNA sequences, while long sequences are very close to standard.

Z-scores of tetranucleotides

Z-scores of tetranucleotides were computed as reported by Teeling et al. (2004) to describe the oligonucleotide composition of a sequence.

$$Z_{XYZW} = \frac{N_{XYZW} - E_{XYZW}}{\sqrt{\mathrm{var}(N_{XYZW})}}$$

where N_{XYZW} is the number of occurrences for tetranucleotide XYZW, E_{XYZW} is the expected number of occurrences, and var(N_{XYZW}) is the variance of the tetranucleotide.

For each tetranucleotide XYZW, an expected frequency can be calculated by means of maximal-order Markov model:

$$E_{XYZW} = \frac{N_{XYZ} N_{YZW}}{N_{YZ}}$$

where N_{XYZ}, N_{YZW} and N_{YZ} are the frequencies of tetranucleotides and dinucleotides within the tetranucleotide XYZW.

The variance can be approximated as follows:

$$\mathrm{var}(N_{XYZW}) = E_{XYZW} \frac{(N_{YZ} - N_{XYZ})(N_{YZ} - N_{YZW})}{N_{YZ}{}^{2}}$$

Arrays containing the z-scores for all tetranucleotides were obtained for each searched sequence.

Computation of distances

Several methods may be used to compute distances for oligonucleotide frequencies. They all require previously obtaining an array of frequencies (or their derivatives) for all oligonucleotides of a given length. Without doubt, in the literature, the most searched oligonucleotides are tetranucleotides. Here we have described the methods to compare frequencies we believe have been used more often in the literature: two methods based on Pearson's correlation, and Genomic Signature Distance and Euclidean distance. Selection of the statistical method may depend on many factors, but there are a couple of ideas to be considered: comparison of sequences by using a correlation coefficient does not require standardization of data (the distances will be exactly the same ones), while computing Genomic Signature Distance and Euclidean distance does require the data to be standardized (check discussion above).

Pearson distance

To compute Pearson distance (d), first Pearson correlation coefficient (r_{xy}) must be computed. In the literature two different Pearson correlation coefficients have been used to compare oligonucleotide frequencies: the standard correlation and the weighted one.

Standard Pearson correlation coefficient ($r_{x,y}$) and Standard Pearson distance

The Pearson correlation r_{xy} for array $(X_i)_{(1\leq i\leq n)}$ and array $(Y_i)_{(1\leq i\leq n)}$ is defined by the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where x_i is the frequency i-th oligonucleotide, and $\overline{x}$ and $\overline{y}$ are the media of arrays $(X_i)_{(1\leq i\leq n)}$ and array $(Y_i)_{(1\leq i\leq n)}$.

To increase the computing speed, a single-pass alternative formula is available:

$$r_{xy} = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2}\sqrt{n\sum_{i=1}^{n}y_i^2 - (\sum_{i=1}^{n}y_i)^2}}$$

The formula above can be numerically unstable when the numbers of decimal values are limited during the computing. Consequently, both formulas must be checked in our particular computer to confirm they yield the same correlation values. As we detected no problems with the numerical instability of the formula, in all our tools this formula is used.

The standard Pearson distance (*d*) is defined as

$$d = 1 - r_{xy}$$

Weighted Pearson correlation coefficient ($rw_{x,y}$) and Weighted Pearson distance

The Weighted Pearson correlation rw_{xy} for array $(X_i)_{(1 \le i \le n)}$ and array $(Y_i)_{(1 \le i \le n)}$ is defined by the following formulas:

$$nw = \sum_{i=1}^{n} x_i y_i$$

$$\overline{xw} = \frac{\sum_{i=1}^{n} x_i^{\,2} y_i}{nw} \quad \overline{yw} = \frac{\sum_{i=1}^{n} x_i y_i^{\,2}}{nw}$$

$$sx = \frac{\sum_{i=1}^{n} (x_i - \overline{xw})^2 x_i y_i}{nw} \quad sy = \frac{\sum_{i=1}^{n} (y_i - \overline{yw})^2 x_i y_i}{nw}$$

$$rw_{x,y} = \frac{\sum_{i=1}^{n} \frac{x_i - \overline{xw}}{\sqrt{sx}} . \frac{y_i - \overline{yw}}{\sqrt{sy}} ..x_i y_i}{nw}$$

This modification of the standard correlation coefficient is explained in-depth by Almeida et al. (2001), and it is named Global distance in their article. Briefly, in this formula the oligonucleotides with low frequency are under-valorised to avoid small modification of their number (with may mean a big change on frequency) to influence the correlation coefficient.

The weighted Pearson distance (*d*) is defined as

$$d = 1 - rw_{xy}$$

Euclidean distance

The Euclidean distance *d* for array $(X_i)_{(1\leq i\leq n)}$ and array $(Y_i)_{(1\leq i\leq n)}$ is defined by the following formula:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Where x_i and y_i are the frequencies of *i*-th oligonucleotide in each array.

In order to adapt the standard Euclid distance so that the distance values for different k values will be in the same range, a constant is added to the formula above, where k is the length of the searched oligonucleotides (Wang et al., 2005).

$$d = \frac{\sqrt{2^k}}{4^k}\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

In all our tools this formula is used.

Genomic Signature Distance

Campbell et al. (1999) defined the **Genomic Signature Distance** *d* between array $(X_i)_{(1\leq i\leq n)}$ and array $(Y_i)_{(1\leq i\leq n)}$ for dinucleotides with the following formula:

$$d = \frac{1}{n}\sum_{i=1}^{n}\left| f'_{xi} - f'_{yi}\right|$$

where f'_{xi} and f'_{yi} are the standardized frequencies of the *i*-th oligonucleotide in arrays X and Y and n is equal to 4^k (where k is the number of bases of the oligonucleotides). Campbell et al. (1999) applied the formula specifically to compare dinucleotide relative abundance profiles and named it **average absolute dinucleotide relative abundance difference (δ*)**.

Wang et al. (2005) applied the same formula for longer oligonucleotides, calling it the Hamming distance, but as the Hamming distance is usually defined for strings or vectors and only accounts for the number of positions at which they differ, the present study will call this the Genomic Signature Distance.

As shown in a separated document, the formula for Genomic Signature Distance may be described this way:

$$d = \sum_{i=1}^{n}\left|\frac{fx_i}{sumX} - \frac{fy_i}{SumY}\right|$$

where f_{xi} and f_{yi} are the number of occurrences of the *i*-th and *j*-th oligonucleotides within the arrays X_i and Y_i and sumX and sumY are precomputed constant values for Σf_{xi} and Σf_{yi}.

Other statistics

Almeida et al. (2001) used the formulas mentioned in this document as Weighted Pearson distance to compare tetranucleotide frequencies. In their article this distance is named **Global distance**.

Pride et al. (2006) computed **Tetranucleotide Usage Deviations (TUD)** based in Zero'th Order Markov chain approximation. The value for each tetranucleotide is described as the ratio of observed (f_{XYZW}) to the expected (N*[$f_X\, f_Y\, f_Z\, f$]; N is the length of the sequence), so that, an array containing all oligonucleotides of a specific size (the TUD) will be equivalent to ZOM values divided by a constant (the length of the sequence). As Pride et al. (2006) used linear regression analysis to compare TUD, correlation coefficients obtained by ZOM and TUD must yield the same results. Computing TUD may require more computational resources than computing ZOM.

Additionally, Pride et al. (2001) computed **tetranucleotide differences** for frequencies derived from a Second Order Markov Chain approximation. Those frequencies correspond, according to the description of the formula, to SOM frequencies above. Anyway, tetranucleotide differences do not seem to be useful to compare different sequences. Finally, in the same text, it is mentioned that Euclidean distance is used to compare TUDs even though the formula for Hamming distance is shown.

Wang et al. (2005) defined **Image distance**, which is indeed a Hamming distance, to compare density matrixes derived from frequency of oligonucleotides in FCGR images. To our knowledge, no other groups had used Image Distance, and it has not been included in this document. Additionally, calculation of the density matrixes is computationally quite expensive, and to determine the neighbourhoods, an R radio must be defined, which may require a lot of searching, or even setting up the value empirically.

REFERENCES

Almeida, J.S., Carriço, J.A., Maretzek, A., Noble, P.A., Fletcher, M. 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics 2001; 17: 429–437

Campbell, A., Mrazek, J., Karlin, S., 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 96, 9184– 9189. (PubMed)

Pride, D.T., Wassenaar, T.M., Ghose, C., Blaser, M.J.. 2006. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 2006, 7:8. (BMC)

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glöckner, F.O. 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics. 2004; 5: 163. (PubMed)

Wang, Y., Hill, K., Singh, S., Kari, L. 2005. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene. 2005;346:173-85. (PubMed)