UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Reconstructing phylogenies based on distances between transforms

Auteurs : DOURT Nicolas, MOREL Guillaume
Promoteur : CATANZARO Daniele
Année académique 2020-2021
Travail de fin d'études (TFE) en vue d'obtenir le titre de
Master (60) en Sciences de Gestion
Horaire de jour

Reconstructing phylogenies based on distances between transforms

Nicolas Dourt and Guillaume Morel
Advisor : Prof. Daniele Catanzaro

Louvain School of Management, Université Catholique de Louvain, Belgium

Abstract -

For the last decades, phylogenetics have become more and more important in genetics researches and of prime interest for DNA sequences analysis. Alignment-free methods have overtaken multiple sequence alignment methods, quite complex and memory consuming. But several of these alignment-free methods are still complex and do not use all the information available (loss of information content). To overcome this issue, new methods have been developed using content of the signals in another domain, using transforms preserving the energy levels, and so, the information). Transforms such as Fourier or Ramanujan are now used for that purpose. DNA sequences (composed of the four types of nucleotides) are often represented using binary code to be usable. The goal of this paper is to introduce new transforms and distances (applied on transforms), not used in this context before, and more convenient for binary signals or to compute distances between spectra. The accuracy of the new methods is assessed using phylogenetic trees to evaluate the quality of the new distance computation. They are built based on artificial and real datasets. The results show that some combinations of new transforms and metrics produce accurate distance computation and phylogenetic trees.

Keywords -

Phylogenetic Trees, Transforms, Distances.

1 Introduction

Researches and technologies development during the last decades have led to the analysis of large volume of sequence data, especially in genetic sequencing. So, new methods and algorithms are needed to study such large data sequences. Recent discoveries in genomics (genes, DNA, codon sequences and so on) are based on new methods of classification and comparison of different genetic sequences. This is the research domain of phylogenetics. Phylogenetics studies hierarchical evolution and relationships among genes, genomes and more broadly species and taxa using molecular genetic data as explained before. So, its purpose is to find the most effective and accurate similarity measures[1][2].

The main tool used is called a phylogeny, or more generally, a phylogenetic tree. It represents quite simply and clearly the evolution of species as a tree, with parents and forebears at the origin of the tree and children at the leaves. This tool and phylogenetics have a lot of fields of interest and practical application such as classification of species, medical and pharmaceutical applications (origins and evolution of pathogens and diseases), epidemiology (understanding and preventing evolution of pandemics), bioinformatics (developing software based on biological behavior), and so on[1][3]. Building a phylogenetic tree such as we do in this paper is a two-step process. The first one is to calculate the distance (based on a metric) two by two between sequences of a dataset and compute the distance matrix between these sequences. The main goal of this phase is to implement the most effective metric(or distances) to have the best similarity : the shortest distance must have the best link with reality, the best real meaning. Once this matrix is obtained, the tree can be built using techniques as UPGMA or neighbor-joining that are agglomerative (bottom-up) methods that build the tree itteratively from the matrix[2].

The main challenge of the computation of such phylogenetic trees is to find the best distance method to have the most realistic results. Litterature is full of papers proposing a lot of methods. The main problem is that is difficult to assess the best metric or validate it empirically due to the large amount of data and the several interpretations of the results. Each method has its hypothesis leading to different results.

Methods for distance computation have evolved over the years. The first kind of methods used is the multiple sequence alignment (MSA)[4][5]. In MSA methods, a number of sequences are aligned by adding gaps in sequences such that they all have the same length (greater than the maximal length of the sequences). Then, sequences are compared for each character using metrics as the Euclidian distance for example. The goal of these methods is to find the best way of aligning sequences to have the shortest distance. But these methods are quite complex and require a lot of memory. Furthemore, trying to minimize the number of gaps can lead to solutions that create more misaligned sequences and do not implement the best distance.

More recently, a new group of methods have been developed to overcome issues of MSA methods. These are alignment-free methods[6]. They solve the problem of misalignment. Some of them also greatly reduce computational complexity. The first ones were based on frequencies of words (nucleotides sequences) in large sequences. One of the first ones, still used today, is the method based on k-mer frequency words. These methods also have drawbacks. Some of them do not improve performances and also need large computation time and memory space. Another issue is that these methods often lose information in DNA sequences and do not use the whole information contained in the signal.

In order to overcome the last drawback, another type of methods have been used : transforms of signal in other domains that do not lose information. The first transform used is the Fourier transform that transforms the signal from time series domain to the frequency domain. The spectral content can bring other information about the signal, periodicities of the distribution of the nucleotides for example. According to Parseval's theorem, the transform does not affect the energy levels and so the information is conserved. But the quantity of data can be reduced (and so complexity) when the spectral content is not distributed over all frequencies (mainly the case for signals with a small number of frequencies). Classical metrics can then be used to compute distance in the new domain. Finally, not so many transforms have been used in the litterature about phylogenetics. Two of them can be mentioned because of the good results : Fourier transform[2] and Ramanujan transform[7] (quite close to Fourier).

Based on this overview of the different distance computations methods,the goal of this paper is to propose new methods developed in other contexts and observe if they can be useful for phylogenetics purpose. As a reference for comparison, a distance based on the mutation model named General Time-Reversible model (well-known in phylogenetics) is used[1][8]. The basic method for distance computation is based on the Fourier transform and the Euclidian metric, so they are briefly recalled in this paper. After that, new methods or combination are studied with two new transforms in this field (Walsh-Hadamard[9] and Wavelet[10]) and two distances (Euclidian and Likelihood ratio[11]).

After this first section of introduction, the structure of the paper is composed of two sections and a conclusion. The second section is dedicated to the theoretical aspects. Transforms and distances are defined. Binary signals are mainly used in this article so the way DNA sequences are represented as binary sequences is also introduced. The algorithms used to build phylogenetic trees and implement mutations to create artificial datasets are also explained. The third section is dedicated to results. Practical methods to find and use the datasets are explained and results are analyzed.

2 Methods

2.1 Numerical DNA sequences representations

Molecules of DNA are made out of chains of nucleotides of four types : Adenine(A) , Cytosine(C), Guanine (G) and Thymine(T). These molecules can be seen as permutations of these four types of nucleotides at different lengths. In order to compute transforms of these signals, a mapping from the four nucleotides to a numerical sequence is needed. A solution commonly found in literature is to map them to binary sequences.

The principle is to decompose a DNA sequence

$s(n)$ into four binary indicator sequences : $m_A(n)$, $m_C(n)$, $m_G(n)$ and $m_T(n)$. For each index of an indicator sequence $m_X(n)$, the value of the sequence is 1 if the nucleotide at this index of the DNA sequence $s(n)$ is X-type, 0 otherwise. This indicator mapping of DNA sequences can be defined as follows :

$$m_X(n) = \begin{cases} 1 & \text{if } s(n) = X \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

where $X \in$ A,C,G,T. The four indicator sequences therefore indicate the appearance of each of the four types of nucleotides at each position n. For example, if $s(n)$ = AAGC, then $m_A(n)$ = 1100, $m_C(n)$ = 0001, $m_G(n)$ = 0010 and $m_T(n)$ = 0000.

2.2 Transforms

2.2.1 Fourier Transform

The Discrete Fourier Transform (DFT) is a conversion of a finite number of observations from the *time domain* to the same finite number of new observations in the *frequency domain*. It is a very powerful tool used in many domains and many applications : from regular spectral analysis of time series, to data compression, or even the resolution of partial differential equations. The DFT of a sequence $m(n)$ is given by :

$$Y(k) = \sum_{n=0}^{N-1} m(n) \exp \frac{-nk2\pi i}{N}$$

The different Y_k being the DFT coefficients over the frequency domain of the function. An important notion linked to the different transforms used here is the Power Spectrum of a transform. It is defined as :

$$PS(k) = |Y(k)|^2$$

This notion is important in this context as it has already been used in phylogenetics to show the periodicity of coding genes through the *3-Base Periodicity Property* [12]. The principle is the following : for a coding gene, using the first decomposition in binary sequences (equation 2.1), and computing the Power Spectra of the four sequences then adding them together, we get a signal :

$$P(k) = |Y_A(k)|^2 + |Y_C(k)|^2 + |Y_G(k)|^2 + |Y_T(k)|^2$$

This signal shows a peak in the frequency $\frac{N}{3}$, which is a result of the way that nucleotides function in triplets (codons), the bases that make up amino acids (and thus proteins, per the role of the coding genes), and not in another way. However, this does not mean that the DFT is useless for non-coding genes. For any gene, the Fourier Transform contains information on the distribution of the different nucleotides along the DNA sequences. This information is useful as it allows us to make comparisons between the sequences in another domain.

2.2.2 Hadamard Transform

The Hadamard Transform, also called Walsh-Hadamard Transform, is another transform that can be useful to treat signals [9]. For signals with a quite small spectral content (a combination of a few sinusoids with different frequencies), it can be useful to correlate the signal with sines and cosines to obtain The Fourier Transform in the spectral domain. For signals taking discrete values (a lot of different frequencies), some other transforms, as the Hadamard transform, are more useful. This transform also consist of a projection but this time in a sequency domain (based on the matrix of the Hadamard transform). The Walsh-Hadamard transform has been used for several purposes such as word recognition, image processing, transmission or filtering. It has also been used in phylogenetics to build phylogenetic trees from molecular data. It could also be used in a reverse mode to compute a maximum likelihood estimation of a tree but is less powerful because other methods are more efficient.

The Hadamard transform is defined, based on Hadamard functions (or matrices), as the product of the Hadamard matrix and the signal. It is also possible to write it for each index H_p :

$$Y(k) = \sum_{n=0}^{N-1} (H_p)_{k,n} \ \ m(n)$$

with N the length of the signal $m(n)$. The Hadamard matrices are defined, starting numbering the rows and columns from 0, as [13] :

$$(H_p)_{i,j} = \frac{1}{2^{\frac{p}{2}}} (-1)^{i \cdot j}$$

The normalization factor $\frac{1}{2^{\frac{p}{2}}}$ can be omitted. $i \cdot j$ is the bitwise dot product of the binary representations of i and j. It can be used to define Hadamard matrices only for $p \geq 2$. For example, $(H_3)_{3,4} = (-1)^{3 \cdot 4} = (-1)^{(0,1,1)\cdot(1,0,0)} = (-1)^0 = 1$. For p=0 and p= 1, there are no formulas. So for p=1..2,matrices without normalization factor are :

$$H_0 = +(1) \qquad H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$H_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

The Hadamard matrices of different sizes can also be build recursively from the first matrix H_1 and based on the Kronecker product. This product places at each entry of the first matrix the product of the entry and the second matrix. It can be written :

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \dots & a_{nn}B \end{pmatrix}$$

The p^{th} Hadamard matrix can be computed using this product between the first Hadamard matrix and the p^{th}-1 matrix : $H_p = H_1 \otimes H_{p-1}$. The matrix form is the following one.

$$H_p = \begin{pmatrix} H_{p-1} & H_{p-1} \\ H_{p-1} & -H_{p-1} \end{pmatrix}$$

It can be observed that there is a p index depending on the size of the Hadamard matrix. This index can be linked to the size of the signal studied N = 2^p. So, for a signal of size N=2^p, the p^{th} Hadamard matrix H_p is used. So, this matrix has a size NxN (it is obvious from the product between the matrix and the signal). This forces the studied signals to have a length of size N=2^p.

Formulas used to compute the Hadamard matrices and transform are finally quite simple but are time and ressources consuming. Using the matrix multiplication results in a complexity of N^2 or 2^{2p}. Many Fast Walsh-Hadamard transform (FWHT) algorithms exist and are much faster. They use symmetries coming from the recursive construction. Many of them have a complexity of $O(N \log N)$ or $O(p2^p)$. The one used in this paper is described here after [14] and illustrated on Figure 1. [15]

Algorithm 1: FWHT algorithm

Data: a signal m(n) of size N = 2^p
Result: Walsh-Hadamard transform of the signal
Build p+1 columns;
Fill the first column with values m(n) of the signal;
Start the loop in the first column;
Initialize a length constant M = N and a constant C=1 counting the number of column parts;
while *All columns are not filled up* **do**
  Divide all the actual column parts in two parts of same length M=M/2 (top half and bottom half);
  In the top half of the same part of the next column, compute the mutually exclusive sums of the respective values of the two parts;
  In the bottom half of this part, compute the mutually exclusive differences;
  There are now C = 2C parts of column
end
Group all the parts together to obtain the transform of the signal

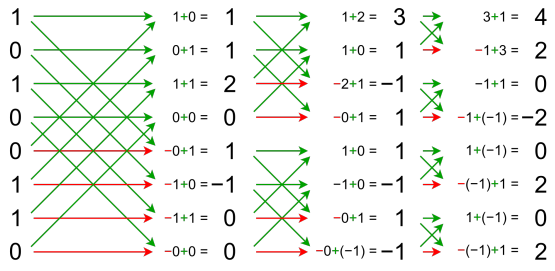

Figure 1: Fast Walsh-Hadamard transform algorithm

There exist a lot of different forms of Hadamard matrices. The more rigorous is the one mentioned before : H_p. But some others exist. The best known alternative is the sequency ordered [16], or Walsh ordered, Walsh-Hadamard transform W_p. From the classical H_p, W_p is build by rearranging the rows of

the matrix such that the number of the row (starting from 0) is the number of sign changes (from +1 to -1 or conversely) when the row is browsed from end to the other. For example, here is W_2 :

$$W_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}$$

2.2.3 Wavelet Transforms : Haar Transform

There exist a lot of other transforms that are more useful than the Fourier transform for some kinds of signals. One of them is the Wavelet transform that overcomes some disadvantages of the Fourier transform [10] [17] [18]. One problem with Fourier is that it captures the information about frequency on the whole signal but it is not suited for signals that have non-zero values on a short time interval (ECG for example). So, Fourier transform is well suited for long signals without sharp transitions. The Wavelet transform is more convenient for such signals because it captures information about frequency and localization of signals (in keeping with the uncertainty principle). Another advantage of the Wavelets is the large library of functions that can be used to compute the transform (see after for mathematical expression) according to the application.

Wavelets are wave-like functions that are localized in frequency but also in time. They have two major properties : the scale that captures frequency information and the location that captures time information. One example of such a function is the derivative of the Gaussian function that is represented on Figure 2 and has the following expression (ignoring a multiplication factor). The parameter a is the scaling factor and b is the location factor[10].

$$f(t) = -(t-b)e^{\frac{-(t-b)^2/(2a^2)}{\sqrt{2}\pi a^3}}$$

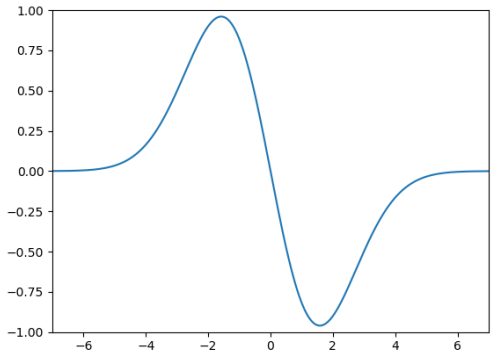

Figure 2: Wavelet function : derivative of Gaussian function

Changing values of a and b leads to respectively stretched and shifted functions that are shown on Figure 3 (orange for the stretched function and green for the shifted function).

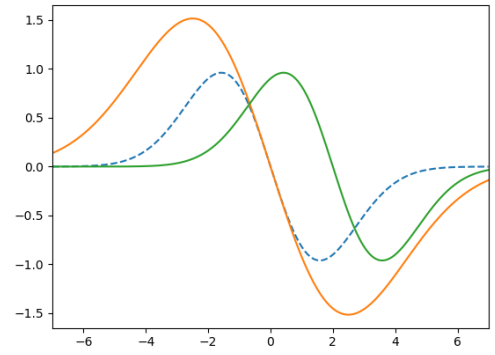

Figure 3: Stretched and shifted Wavelet functions

Using such functions (called wavelets) allows to compute the Wavelet transform (continuous and discrete transforms) for different scaling and location factors (a and b for the continuous form, and m and n for the discrete form)[10][17].

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t) \Psi \frac{t-b}{a} dt$$

$$T_{m,n} = \int_{-\infty}^{+\infty} x(t) \Psi_{m,n}(t) dt$$

This transform can be compared to a convolution product : for given scaling and locations factors, the signal is multiplied by the Wavelet function which

is shifted over the entire time domain to compute "how much a certain wavelet is in the signal for each location". So, many functions can be used as Wavelet functions and these functions can be computed for several scaling and location factors. One of the family of functions often used (this is the one implemented in this paper) is the Haar functions[17][19]. These functions are quite simple and are defined with the following expression and parameters. This is the continuous form but an analogous form can be used for discrete time series.

$$H_0(x) = H_{0,0}(x) = \frac{1}{\sqrt{N}} \qquad \text{for } x \in [0,1]$$

$$H_k(x) = H_{p,q}(x) = \begin{cases} \frac{1}{\sqrt{N}} 2^{\frac{p}{2}} & \text{for } \frac{q-1}{2^p} \leq x \leq \frac{q-\frac{1}{2}}{2^p} \\ -\frac{1}{\sqrt{N}} 2^{\frac{p}{2}} & \text{for } \frac{q-\frac{1}{2}}{2^p} \leq x \leq \frac{q}{2^p} \\ 0 & \text{otherwise} \end{cases}$$

Depending on the desired number of Haar functions, parameters are defined as :

$$0 \leq p \leq n-1$$

$$\begin{cases} q = 0, 1 \text{ for } p = 0 \\ 1 \leq q \leq 2^p \text{ for } p > 0 \end{cases}$$

$$k = 2^p + q - 1$$

The last expression gives the following (p,q) combination : $k = 0 \rightarrow (0,0); k = 1 \rightarrow (0,1); k = 2 \rightarrow (1,1)$ and so on. Neglecting the factor $\frac{1}{\sqrt{N}}$, the first two Haar functions (k=0 and k=1) are represented on Figure 4.

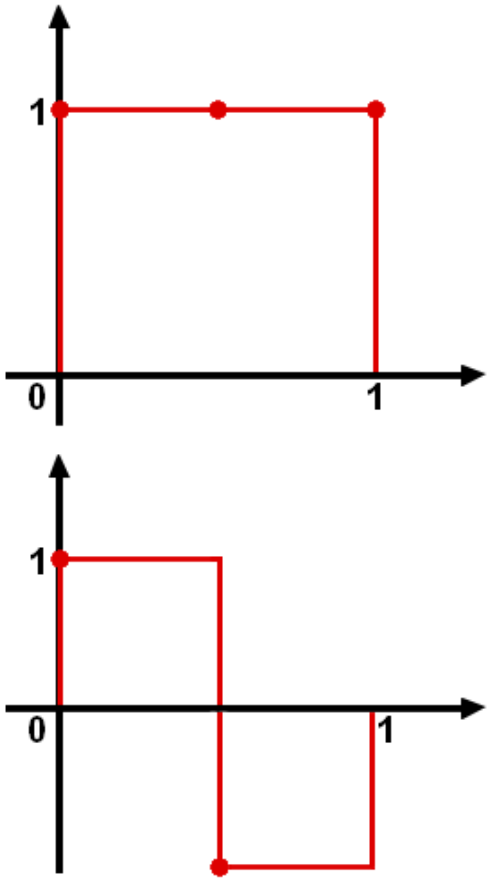

Figure 4: First two Haar functions

These functions can be used in a discrete way to compute Haar transform over discrete series[20][18]. For that purpose, points on Figure 4 have to be considered (at positions x=0 and x=1/2). These functions are applied to pairs of data (so entries of the series are grouped by two before the transform is applied). These transforms compute the sum and the difference of these pairs of data. These simple functions are the ones implemented in this paper to compute the Haar Wavelet transform according to the algorithm explained just before (group data by pairs and compute sums and differences). So the vector of data has the same length but is presented in another way. The advantages of this form are multiple. First, changes over the set of data can be observed in the difference function. Data is easier to quantize and there are fewer digits. The information (or energy of the signal) is concentrated in fewer values (mainly the sum function).

As for Hadamard transform, a matrix form of the transform can be developed. It is more complex mathematically so it is not computed here but it rises the same condition on the length of the used signals(N=2^p).

2.3 Computing distances

In order to compare the different signals obtained, we need to have distances. The first two distances will be used in 2.4 to reconstruct the phylogenetic trees that we want in the end. The distances we need here do not necessarily need to be metrics, in the sense that they do not need the axioms of *identity of indiscernibles*, *symmetry* and *triangular inequality* to be respected. It is better to do so, but in some motivated cases, it might be interesting to use a distance such as the *likelihood ratio distance* presented further below. The last distance is a distance specifically made for phylogenetics, and which will be used as a comparison for our methods.

2.3.1 Euclidean

The Euclidean distance between two series is the most commonly used notion of distance :

$$d((x_1, x_2, .., x_n), (y_1, y_2, .., y_n)) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2.3.2 Likelihood ratio

The Likelihood Ratio distance was introduced in [11], as an alternative to the use of the Euclidean distance when dealing with Fourier transforms. This distance, based on a log-likelihood ratio test between two series x and y is given by :

$$\Lambda = 2 \sum_{i=1}^{n-1} 2\log(a_i + b_i) - \log(a_i) - \log(b_i)$$

The terms a_i and b_i being the coefficients of the Power Spectra of the series x and y. Even though the distance between one series and itself does not give 0, it gives a constant value for series of the same length, and thus it will not impact our tree reconstruction as the analyzed series are the same length. The paper in which it was introduced concludes that this distance favours the detection of significant changes in the smaller coefficients of the Fourier Transform, whereas the Euclidean distance will not differentiate the changes between smaller and larger coefficients (the latter being high frequencies which are often synonymous with noise).

2.3.3 General Time-Reversible distance

The General Time-Reversible model is a model of evolution described by Lanave et al. in [21], and is one of the most used and popular model of substitution. From the basis of this model, it is possible to compute a distance between two DNA sequences. The reasoning comes from the following equation :

$$P(t) = e^{At}$$

with $P(t)$ being the transition probability matrix, and A being the rate matrix and t being the branch length between two sequences in a tree. The principle is that we can estimate $P(t)$ by $\hat{P}(t)$ using data, by making comparisons between sequences, and each column will be equal to the number of substitutions from one nucleotide to another divided by the total number of substitution for each nucleotide. It is then possible to compute

$$\hat{A}t = \log(\hat{P}(t))$$

in which the log is the matrix logarithm, and where we need to isolate $\hat{t}$. By fixing the total rate of changes equal to 1 per unit of time, we can say that the changes from one state i to a state j are then $\pi_j a_{ij}$ per unit of time, with π_j being the frequency of apparition of a the nucleotide j. From there, the result

$$-\text{trace}(\hat{A}\hat{\Pi}) = 1$$

appears, with Π being a diagonal matrix with its entries being the π_j. Then, it follows that :

$$\hat{t} = -\text{trace}(\log(\hat{P})\hat{\Pi})$$

and thus we have the value of the General Time-Reversible distance, or *evolutionary distance* [22].

2.4 Phylogenetic tree reconstruction

There are different ways to reconstruct phylogenetic trees. First, there are the distance or similarity based methods, such as *Neighbor-Joining* (NJ)

and *Unweigthed Pair Group Method with Arithmetic mean* (UPGMA). Using a distance (or similarity) matrix, NJ links the different DNA sequences together as an unrooted tree following a criterion of minimum value in a matrix Q, computed such that for two sequences x and y :

$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

with $d(i,j)$ being the distance between the two sequences. In the case where one uses a similarity matrix, then the same criterion can be used as long as the entries of the matrix are multiplied by a factor -1, so that the principles of distances and similarities are respected. Using a distance matrix, UPGMA links the different DNA sequences together as a rooted tree following a criterion of minimum value directly in the matrix :

Algorithm 2: UPGMA algorithm

Data: a distance matrix of the k DNA sequences $M(k,k)$
Result: A rooted phylogenetic tree
while *All sequences are not linked in the tree* **do**
  Find i, j such that $M(i,j) = min(M)$;
  Link i and j as a node (i,j) in the tree, and compute the distance between the new node and the sequences by averaging the distance between each sequence of the node and the other sequences;
  Update M by substituting i and j by the new node (i,j);
end
The last link makes up the root of the tree.

As this work is based on distances, we will use one of these two methods, the UPGMA, to reconstruct phylogenetic trees.
Other paradigms exist, and a few can be listed : *Maximum Parsimony*, first presented in [23], which is an optimality criterion under which the tree with the smallest number of evolutions (i.e. the tree with the lowest diameter) and that explains the data correctly will be considered as the best tree [22]; *Maximum Likelihood*, which uses the work of Fisher as a way to compute the most probable tree possible, given some DNA sequences and an evolutionary model which makes it possible to compute the probability of transition from one state to another ; *Hadamard Methods*, which use the very same transform presented in section 2.2.2 but in a different way : Steel et al. in [24] use the Hadamard Transform to compute a vector γ, called the *conjugate spectrum*, which can be compared to vectors (called *branch lengths*) computed from different tree topologies, and the vector that is the closest to the conjugate spectrum is the one associated with the best tree under this paradigm. Many other paradigms exist, and most of them were presented extensively in [22].

2.5 DNA sequences mutations

In order to try the methods presented hereabove, some tests were needed on artificial data to assess the accuracy of the phylogenetic tree generation. To do so, we built a so-called *mutator*, which is a function that takes a DNA sequence, and applies artificial mutations on it for a given amount of time. There are multiple models that exist to simulate mutations, such as *Jukes-Cantor's* model [25], *Kimura's* models [26, 27], *Felsenstein's* model [28], or *Tavaré's* General Time Reversible (GTR) model [29].
One should know that some principles of mutation are quite hard to implement, as it is a biological process with a stochastic behaviour. For example, some characteristics, such as the 3-base periodicity principle, make it so that in coding genes, mutations rarely occur in the first and second nucleotides of a codon, whereas in non-coding genes, mutations occur quite evenly between the first, second and third codons. Two types of point mutations (mutations that affect only one nucleotide at a time) exist : first, the substitutions, themselves divided in 2 categories : transitions and transversions. Substitutions are mutations where a nucleotide of the sequence is replaced by another one. In the case of a transition, nucleotides called purines (adenine and guanine) are replaced by other purines, and pyrimidines (cytosine and thymine) are replaced by other pyrimidines ; and in the case of a transversion, purines are replaced by pyrimidines, and pyrimidines are replaced by purines. The second type, insertions and deletions are events in which a nucleotide is inserted in the DNA sequence for the former, or in which a nucleotide is deleted from the DNA sequence for the latter. Whereas the first type

of point mutations are easily modeled, modeling the second type requires very extensive knowledge of the subject and uses very advanced mathematics [1], which is why we decided to concentrate on the first one, and thus only model substitutions in our *mutator*, whose use is presented in section 3.

3 Results

In this section, we present the results we obtained for the reconstruction of phylogenetic trees using the UPGMA algorithm, with a distance matrix based on both the *Walsh-Hadamard Transform* and the *Wavelet Transform*, and both the *Euclidean distance* and the *Likelihood Ratio Statistic*. But first, in order to justify the use of such methods for DNA sequences, we applied these methods to some "playground data", and therefore validated them.

3.1 Playground data

To justify the use of the aforementioned methods, we first tried them on some "playground data". The principle is that we used a DNA sequence of a Human from a dataset used in [22], which we will call X, and then inserted 10% (or $\lceil \frac{n}{10} \rceil$ for a sequence of length n) of random mutations (substitutions only, as mentioned before, without differentiating the two types) into it twice to generate two new sequences, A and B. From these two new sequences, we generated four new sequences $A1$, $A2$, $B1$ and $B2$ by following the same principle, inserting 10% of random mutations. From these four sequences again, we generated eight new sequences $A3$, $A4$, $A5$, $A6$, $B3$, $B4$, $B5$, $B6$. The results for the Walsh-Hadamard transform with the Euclidean distance can be found in figure 5. The results for the other combinations of transforms and distances can be found in the Appendix A.

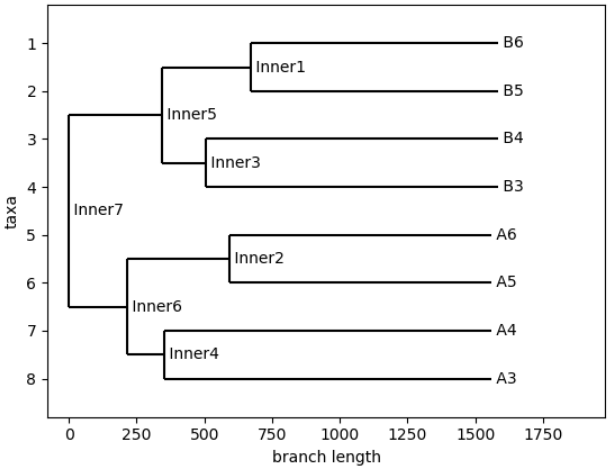

Figure 5: Phylogenetic Tree for the playground data using the Walsh-Hadamard transform and Euclidean distance

We can observe for all of these that the topologies of the trees are perfectly coherent with the expected output shown in figure 15. The only differences appear in the branch lengths, which represent a unit of time between evolutions, and thus vary because of the differences in transforms outputs and in distances. Because of these results, we thus validate the fact that the transforms and distances we are using in our methods are justified to construct phylogenetic trees, and should perform well on the real data as well.

3.2 Real Data

As our methods were validated on fake datasets, we performed tests on real datasets. The first dataset is composed of mammals DNA sequences. The comparison between the different trees was easier due to the evident knowledge about these animals. The second dataset is based on different flu strains, as it is a first step towards SARS-COV-2, and because the different strains are already well known and make the verification process easier. The last one is composed of SARS-COV-2 strains, as it is still a worldwide concern that made this work more stimulating and more in tune with the times. Results are compared to the case of the GTR distance to see the effectiveness of the different metrics and transforms. For each figure, the abscissae is the branch length, and it varies greatly between the different transforms and distance, but is only representative for the corresponding tree,

and they should not be compared from one tree to another. It represents a scale of the genetic divergence between the common ancestors and the leafs of the tree. As some branch lengths sometimes made the trees very difficult to read and interpret, we took the liberty of rescaling the abscissae at some points to make it clearer for the reader.

3.2.1 Mammals

The dataset used for these results comes from [22], and represents a part of a mitochondrial gene from 14 different mammals (12 primates, a bovine and a mouse). The sequences are 232 nucleotides long, and thus required the use of 0-padding in order to be of a length that is a power of 2, as needed for the Walsh-Hadamard and the Haar transforms.

Results for mammals are presented on Figures 6, 7 and 8. Other results are shown in Appendices. The different combinations of transforms and metrics can be compared to a more classical combination : Fourier transform and Euclidian distance. The 3 classifications seem quite good and logical. It is difficult to assess the good performance of a method because it is not based on a quantitative criterion. It is more based on the qualitative classification of the different breeds. But this dataset allows an easy idea of the performance of the classification.

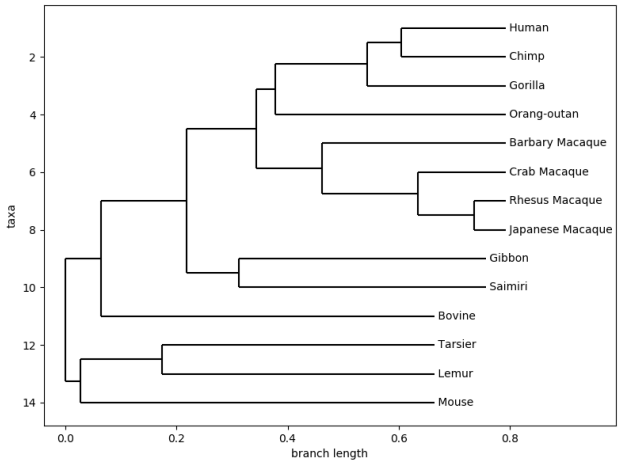

Figure 6: Phylogenetic Tree for the Mammals data using the GTR distance

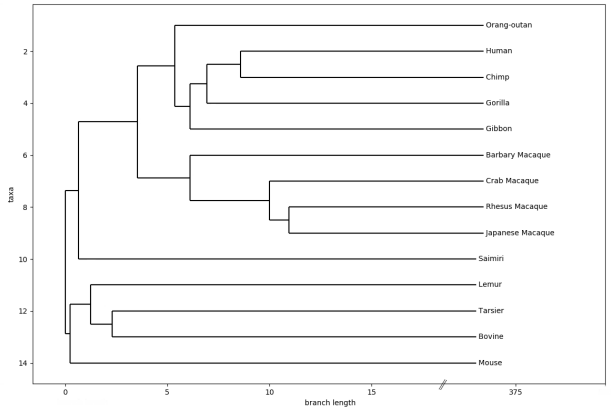

Figure 7: Phylogenetic Tree for the Mammals data using the Haar transform and Likelihood Ratio Statistic

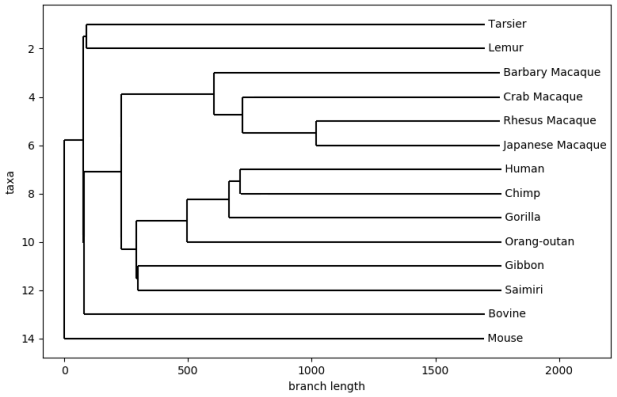

Figure 8: Phylogenetic Tree for the Mammals data using the Walsh-Hadamard transform and Euclidean distance

As we said, the classifications all seem to be logical, but we can observe a few differences : the Gibbon when using the GTR distance (figure 6) is further apart from other apes, which are closer to the macaques, and the place of the Mouse and the Bovine in all trees but the one of Walsh-Hadamard Transform with the Euclidean distance (figures 6,7,19,20) seems quite random, and it is only in figure 8 that we can see a true differentiation of the two non-primate species. However, we have to remember that this is the result of reconstructing trees based on a single gene, which cannot grasp the whole variation between the different genomes, thus often leading to non-perfect phylogenetic trees.

3.2.2 Influenza A

The dataset used for these results comes from [7], and represents the neuraminidase genes from 31 different Influenza A viruses (representing 5 strains : H1N1, H3N2, H7N3, H7N9 and H11N9). The sequences are between 1367 and 1424 nucleotides long, and thus required the use of 0-padding in order to be of a length that is a power of 2, as needed for the Walsh-Hadamard and the Haar transforms. To make it easier to grasp the clustering of the different strains, they are color-coded in the trees.

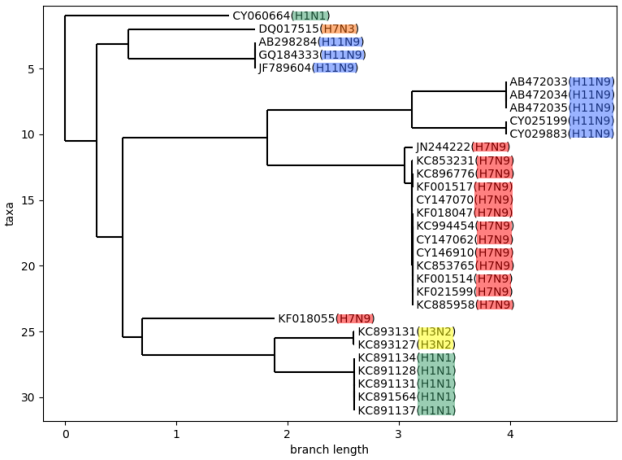

Figure 9: Phylogenetic Tree for the Influenza data using the GTR distance

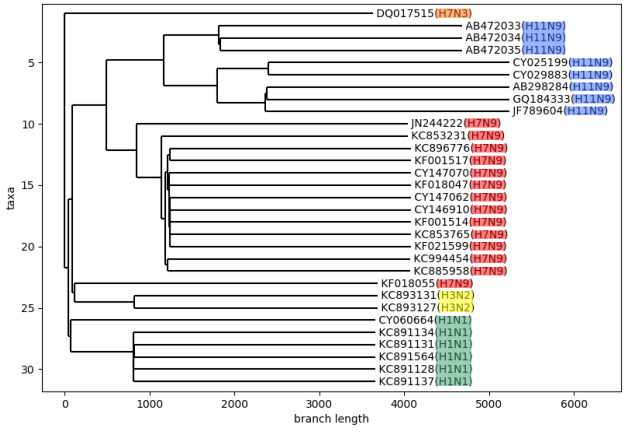

Figure 10: Phylogenetic Tree for the Influenza data using the Walsh-Hadamard transform and Likelihood Ratio Statistic

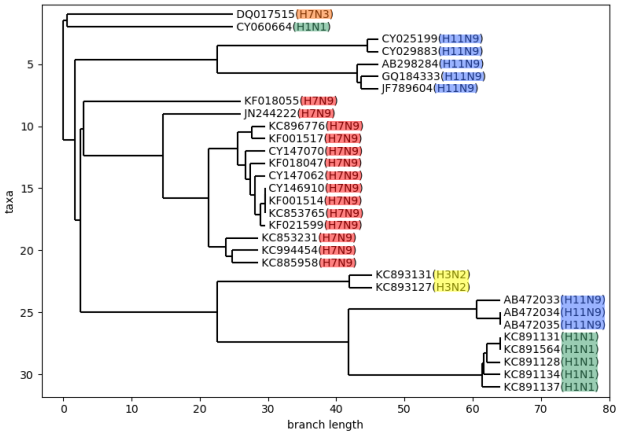

Figure 11: Phylogenetic Tree for the Influenza data using the Haar transform and Euclidean distance

Visually, we can observe that the clustering works quite well for all methods. There are some components that appear to be repetitive : the division of the H11N9 strain in two distinct branches in figures 9, 11 and 22, where the AB472033, AB472034 and AB472035 group often stands out ; the CY060664 sequence from the H1N1 group which stands out from the other sequences except in figure 10, which can possibly be explained by the fact that the other H1N1 sequences were very localized (all were gathered in Illinois in 2012, whereas CY060664 was gathered in Ontario in 2009) ; or KF018055 which also stands out, especially in figures 9, 10 and 21, and JN244222 to a lesser extent in all trees.

3.2.3 SARS-COV-2

The dataset used for these results was made from scratch, and represents the whole genome of 55 SARS-COV-2 viruses (representing 11 strains as identified by the project *Nextstrain* : 19A, 19B, 20A, 20B, 20C, 20E, 20G, 20H, 20I, 20J and 21A). The sequences are between 29695 and 29782 nucleotides long, and thus required the use of 0-padding in order to be of a length that is a power of 2, as needed for the Walsh-Hadamard and the Haar transforms. To make it easier to grasp the clustering of the different strains, they are color-coded in the trees.

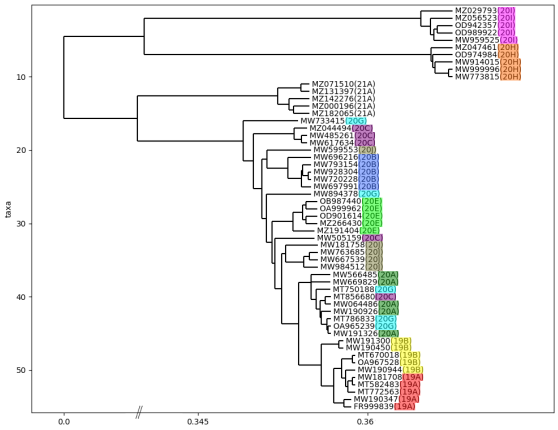

Figure 12: Phylogenetic Tree for the SARS-COV-2 data using the GTR distance

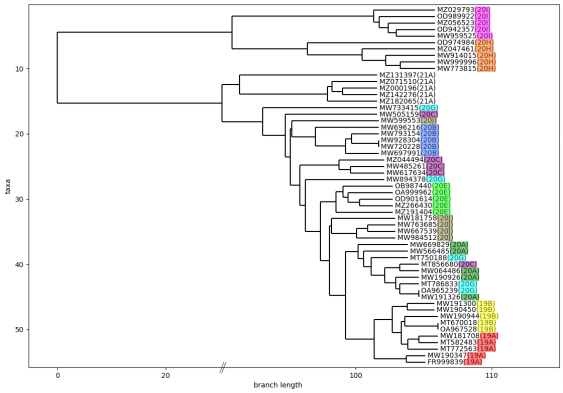

Figure 13: Phylogenetic Tree for the SARS-COV-2 data using the Haar transform and Euclidean distance

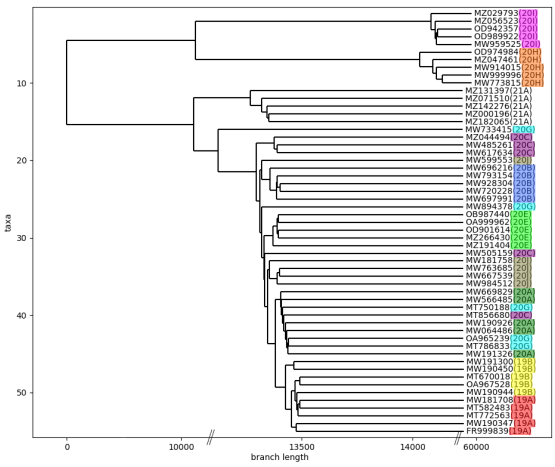

Figure 14: Phylogenetic Tree for the SARS-COV-2 data using the Walsh-Hadamard transform and Likelihood Ratio Statistic

Once again, on this data, we can see that the clustering performs quite well. We can see that the strains 20H and 20I are always well-defined and distant from the other strains ; and that the strains 19A, 19B, 20B, 20E and 21A are always well-defined as well. We observe that in all results, the strains 20A, 20C and 20G mix up and find themselves related. This can probably be explained by a different system of strain identification from the project *GISAID*, which groups some parts of the 20A strains and the 20C, 20G and 20H strains. However as we can see, there is no confusion between the 20H strain and the others. One sequence from the 20J (MW599553) is always far from the rest of the 20J strains, but without a clear explanation as to why.
Based on all these trees, we can conclude that these combinations of distances and transforms performs well to evaluate distances between DNA sequences.

4 Conclusion

In this work, we have established new distances between DNA sequences based on a special ratio statistic between time series, and Fourier and Wavelet transforms. The method consists of transforming the DNA sequences into four numerical signals, and then applying the Walsh-Hadamard or Haar transforms on it. We then computed the distance, either the euclidean or the likelihood ratio statistic, between the power spectra of the transforms, in order to construct phylogenetic trees through the use of UPGMA. We first assessed the performance of our methods on synthetic data, and then we assessed their performances on real data, consisting of Mammals DNA, Influenza DNA, and finally SARS-COV-2 DNA. All methods provided accurate clustering between similar species. These results show that the use of such transforms and distance are effective on such data, and that multiple methods using new or different transforms should be encouraged for such work. It should also be noted that these kind of distances might as well work in the clustering of other data types than DNA sequences, and might open new roads in some disciplines where they could be used.

A Additional Figures

A.1 Playground Data

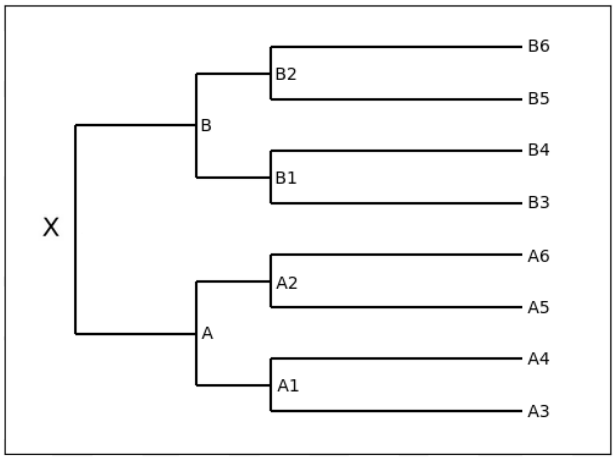

Figure 15: Real phylogenetic tree for the playground data

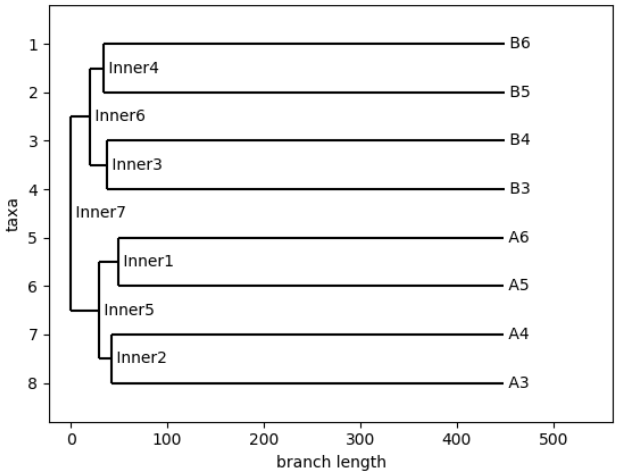

Figure 17: Phylogenetic Tree for the playground data using the Walsh-Hadamard transform and Likelihood Ratio Statistic

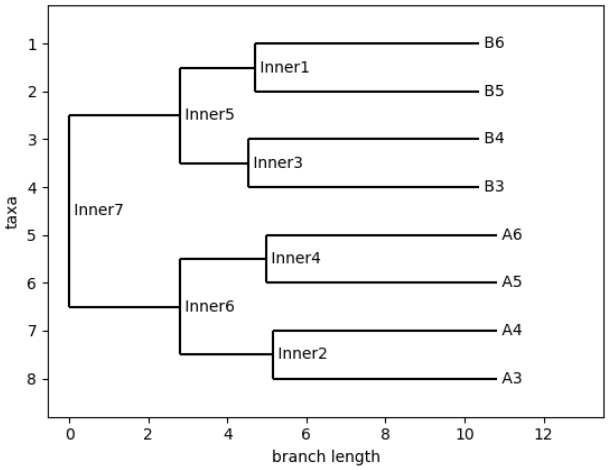

Figure 16: Phylogenetic Tree for the playground data using the Haar transform and Euclidean distance

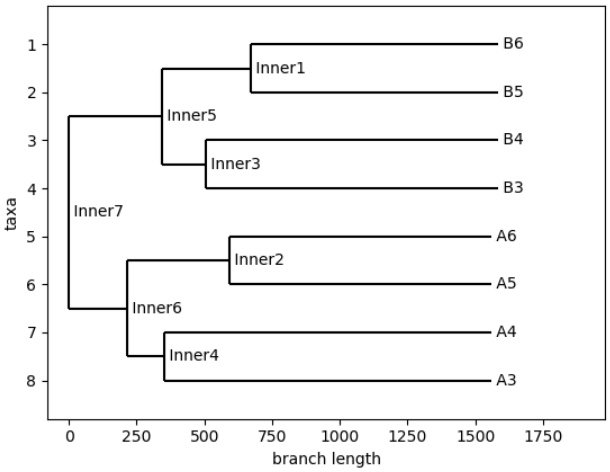

Figure 18: Phylogenetic Tree for the playground data using the Haar transform and Likelihood Ratio Statistic

A.2 Mammals Data

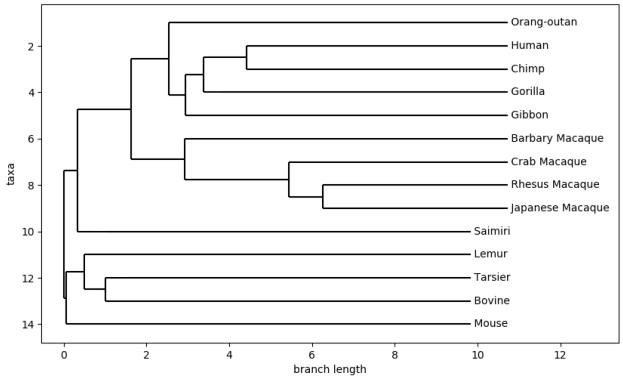

Figure 19: Phylogenetic Tree for the mammals data using the Haar transform and Euclidean distance

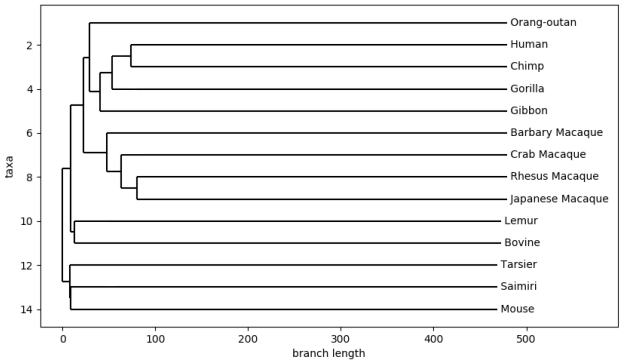

Figure 20: Phylogenetic Tree for the mammals data using the Walsh-Hadamard transform and Likelihood Ratio Statistic

A.3 Flu Data

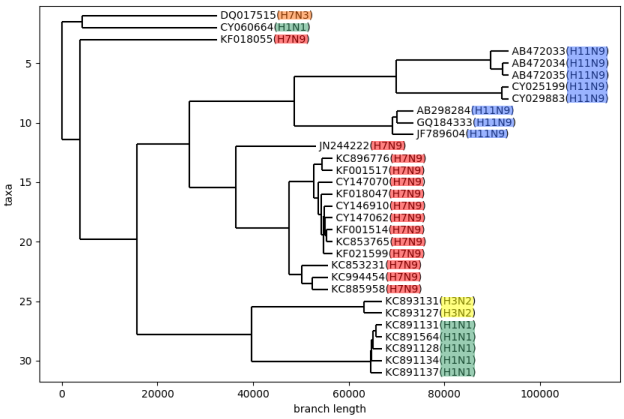

Figure 21: Phylogenetic Tree for the Influenza data using the Walsh-Hadamard transform and Euclidean distance

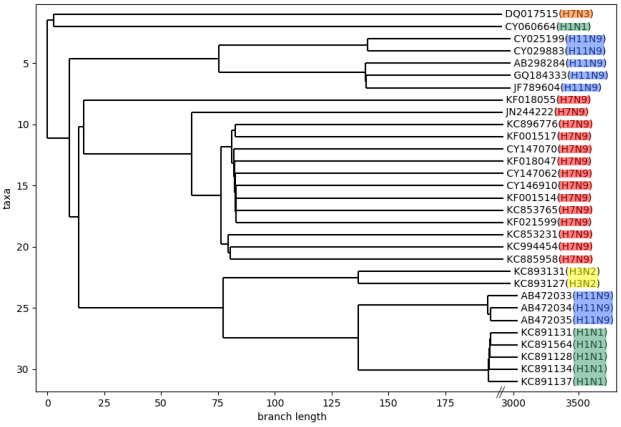

Figure 22: Phylogenetic Tree for the Influenza data using the Haar transform and Likelihood Ratio Statistic

A.4 SARS-COV-2 Data

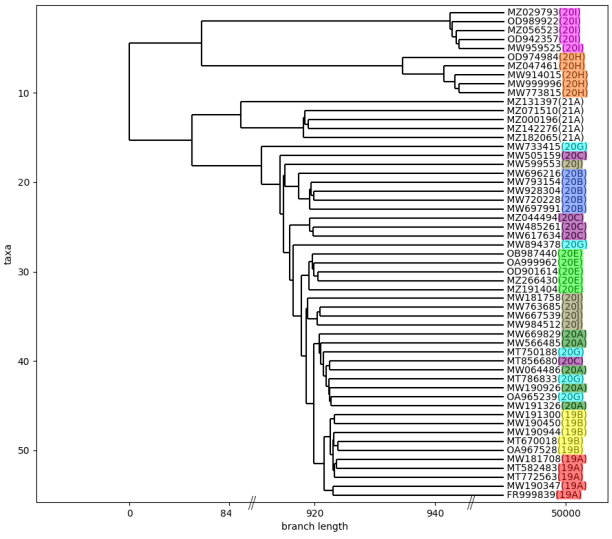

Figure 23: Phylogenetic Tree for the SARS-COV-2 data using the Haar transform and Likelihood Ratio Statistic

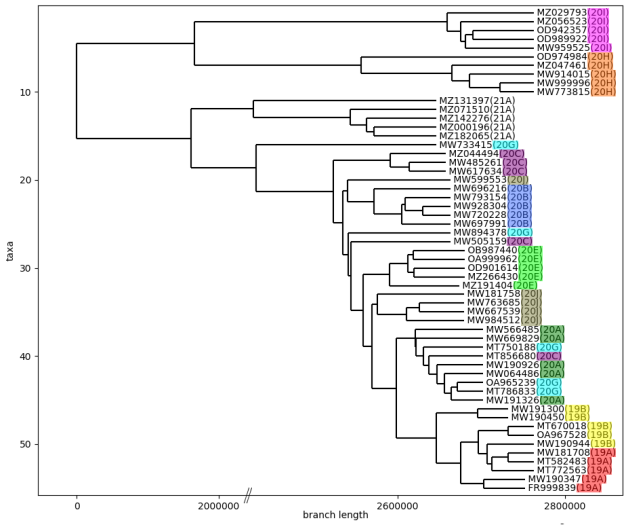

Figure 24: Phylogenetic Tree for the SARS-COV-2 data using the Walsh-Hadamard transform and Euclidean Distance

B GenBank references

The following genomes were accessed and used in our simulations, and come from the NCBI databases accessible on their webpage [30].

GenBank ID	Strain
FR999839	19A
MW190347	19A
MT772563	19A
MT582483	19A
MW181708	19A
MW190450	19B
MW191300	19B
OA967528	19B
MW190944	19B
MT670018	19B
MW191326	20A
MW064486	20A
MW566485	20A
MW669829	20A
MW190926	20A
MW697991	20B
MW696216	20B
MW720228	20B
MW928304	20B
MW793154	20B
MT856680	20C
MW617634	20C
MW505159	20C
MW485261	20C
MZ044494	20C
MZ191404	20E
MZ266430	20E
OD901614	20E
OA999962	20E
OB987440	20E
MW733415	20G
MT786833	20G
MW894378	20G
MT750188	20G
OA965239	20G
MW773815	20H
OD974984	20H
MZ047461	20H
MW914015	20H
MW999996	20H
MW959525	20I
OD989922	20I
MZ056523	20I
OD942357	20I
MZ029793	20I
MW984512	20J
MW667539	20J
MW599553	20J
MW181758	20J
MW763685	20J
MZ182065	21A
MZ131397	21A
MZ142276	21A
MZ071510	21A
MZ000196	21A

Table 1: GenBank accession number and Strain of each sequence for the SARS-COV-2 data

References

[1] D. Catanzaro. Estimating phylogenies from molecular data. In R. Bruni, editor, *Mathemat-*

ical approaches to polymer sequence analysis and related problems, chapter 8, pages 149–176. Springer, New York, 2010.

[2] Stephen S.-T. Yau Changchuan Yin, Ying Chen. A measure of dna sequence similarity by fourier transform with applications on hierarchical clustering. *Journal of Theoretical Biology*, 2014.

[3] EMBL-EBI. Why is phylogenetics important? URL https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/why-is-phylogenetics-important/.

[4] Multiple sequence alignment (msa) methods, . URL https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/multiple-sequence-alignment.html.

[5] Multiple sequence alignment, . URL https://en.wikipedia.org/wiki/Multiple_sequence_alignment.

[6] Alignment-free sequence analysis. URL https://en.wikipedia.org/wiki/Alignment-free_sequence_analysis.

[7] Changchuan Yin and Jiasong Wang. A novel method for comparative analysis of dna sequences by ramanujan-fourier transform. *CoRR*, abs/1403.1523, 2014.

[8] Daniele Catanzaro, Raffaele Pesenti, and Michel C. Milinkovitch. A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the gtr model. *Bioinform.*, 22(6):708–715, 2006. URL http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics22.html#CatanzaroPM06.

[9] D. S. Stoffer and T. Panchalingam. A walsh-fourier approach to the analysis of binary time series. *Time Series and Econometric Modelling*, 1987.

[10] Shawhin Talebi. The wavelet transform : An introduction and examples. URL https://towardsdatascience.com/the-wavelet-transform-e9cfa85d7b34.

[11] G. J. Janacek, A. J. Bagnall, and M. Powell. A Likelihood Ratio Distance Measure for the Similarity Between the Fourier Transform of Time Series. pages 737–743. 2005.

[12] Anastasios Tsonis, James Elsner, and Panagiotis Tsonis. Periodicity in DNA coding sequences: Implications in gene evolution. *Journal of Theoretical Biology*, 151:323–331, 1991.

[13] Hadamard transform. URL https://en.wikipedia.org/wiki/Hadamard_transform.

[14] Charles Constantine Gumas. A century old, the fast hadamard transform proves useful in digital communications. URL https://web.archive.org/web/20060513150253/http://archive.chipcenter.com/dsp/DSP000517F1.html.

[15] Fast walsh–hadamard transform. URL https://en.wikipedia.org/wiki/Fast_Walsh%E2%80%93Hadamard_transform.

[16] Walsh matrix. URL https://en.wikipedia.org/wiki/Walsh_matrix.

[17] Prof. Vaibhav Pandit. Haar transform - image transforms - image processing and machine vision. URL https://www.youtube.com/watch?v=s02VrcLObvU4.

[18] Catherine Bénéteau. Discrete haar wavelet transforms. URL http://www.unm.edu/~vassilev/talk-CB.pdf.

[19] Haar wavelet. URL https://en.wikipedia.org/wiki/Haar_wavelet.

[20] Discrete wavelet transform. URL https://en.wikipedia.org/wiki/Discrete_wavelet_transform.

[21] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

[22] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Suderland, USA, 2004.

[23] A. W. F. Edwards and L. L. Cavalli-Sforza. The reconstruction of evolution. *Ann. Hum. Genet.*, 27:104–105, 1963.

[24] Mike A. Steel, Michael D. Hendy, and David Penny. Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results. *Discret. Appl. Math.*, 88(1-3): 367–396, 1998.

[25] T. Jukes and C. Cantor. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, volume III, chapter 24, pages 21–132. Academic Press, New York, 1969.

[26] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[27] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 78:454–458, 1981.

[28] J Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[29] S. Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[30] National center for biotechnology information (ncbi). URL `https://www.ncbi.nlm.nih.gov/`.

Abstract : For the last decades, phylogenetics have become more and more important in genetics researches and of prime interest for DNA sequences analysis. Alignment-free methods have overtaken multiple sequence alignment methods, quite complex and memory consuming. But several of these alignment-free methods are still complex and do not use all the information available (loss of information content). To overcome this issue, new methods have been developed using content of the signals in another domain, using transforms preserving the energy levels, and so, the information). Transforms such as Fourier or Ramanujan are now used for that purpose. DNA sequences (composed of the four types of nucleotides) are often represented using binary code to be usable. The goal of this paper is to introduce new transforms and distances (applied on transforms), not used in this context before, and more convenient for binary signals or to compute distances between spectra. The accuracy of the new methods is assessed using phylogenetic trees to evaluate the quality of the new distance computation. They are built based on artificial and real datasets. The results show that some combinations of new transforms and metrics produce accurate distance computation and phylogenetic trees.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique

www.uclouvain.be/lsm