A DAWSON-LIKE CLUSTERING OF HUMAN MITOCHONDRIAL DNA SEQUENCES BASED ON PROTEIN CODING REGION

INÊS SOARES$^{(*,\#,\dagger,\ddagger)}$, RUI DUARTE$^{(\S,\P)}$,
ANTÓNIO GUEDES DE OLIVEIRA$^{(\dagger,\ddagger)}$, AND ANTÓNIO AMORIM$^{(*,\ddagger)}$

Abstract. In the present paper, our main goal is focused in developing fast algorithms for human mtDNA sequence analyses, requiring *minimum* and emphexplicit assumptions on mutation models and evolutionary pathways. We propose a new approach based on a construction of Dawson, a technique based on the ordering of the variable sites.

In this approach, the first step corresponds to the computation of the order of the positions according to their capacity to separate the sequences into dichotomous groups. Aiming to avoid or at least to minimize the consideration of ambiguous evolutionary events such as insertions/deletions and recurrence, which cause well-known alignment problems, in the present study we only work with the protein coding sequence, the clearly more stable region in human mitochondrial genomes. This method was tested in a small set of 99 human mtDNA comprising representatives of all major haplogroups. The developed approach showed to be a choice to automate the clustering of human mtDNA sequences into broad groups, the output being in agreement with the canonical classification into macro-haplogroups deposited in the Phylotree database.

Keywords. Mitochondrial DNA, sequence alignment, genome comparison, protein coding region, haplogroup assignment, Dawson construction.

1. Introduction

Nowadays, several fields, such as forensics and medicine, have required more and more methodologies capable of describing the relationships among a set of individuals, in order to find out evolutionary signals and/or disease causing variations. During the last years, the number of novel sequence comparison approaches that have been proposed is

(∗) INESCC - Instituto de Engenharia de Sistemas e Computadores de Coimbra

(∗) i3S - Instituto de Investigacão e Inovação em Saúde, IPATIMUP - Instituto de Patologia e Imunologia Molecular da Universidade do Porto

(†) CMUP, Centro de Matemática da Universidade do Porto

(‡) Faculty of Sciences of the University of Porto, Portugal

(§) CIDMA, Centro de Investigação e Desenvolvimento em Matemática e Aplicações

(¶) Department of Mathematics, University of Aveiro, Portugal

E-mail addresses: iness@ipatimup.pt, rduarte@ua.pt, agoliv@fc.up.pt, aamorim@ipatimup.pt.

countless, including alignment, the most widely used, and free-alignment methodologies [18, 20, 1, 12, 4].

The procedures of the former kind are very time consuming and the quality of the results depends on the evolutionary soundness of costs attributed to the observed different mutational events. The methodologies of the latter type are generally much faster than those depending on alignment processes; in contrast, however, they do not interpret the events as evolutionary phenomena. Given this, there is a strong need for novel and efficient approaches in order to solve the sequence comparison problem.

Aiming to compare and to cluster human mitochondrial sequences, we face a bit different reality, with respect to the methodology commonly used. Mammalian mitochondrial DNA is a double-stranded and circular genome with approximately 16500 base pairs, comprising a total of 37 genes, more specifically 13 protein coding genes, 22 tRNAs and 2rRNA [10]. It is a molecule with some remarkable characteristics, such as: (i) has an uniparental mode of transmission, by maternal lineages; (ii) is non-recombining; (iii) presents a high mutation rate [17, 9]. Thus, it passes intact from the ancestors to the offspring, except if some mutation occurs. Consequently, mtDNA genome is currently used to trace back the maternal evolutionary histories in a variety of scales and time depths. Nowadays, in any study of human mtDNA analyses, the common practice has two steps: (a) first mtDNA sequences are aligned against a reference sequence and then (b) a search for diagnostic variable sites in order to classify sequences into haplogroups is performed [19]. These steps, particularly the second one, are done manually in a boring, time-consuming and error-prone process. Although some approaches have been developed aiming to address this problem, it remains challenging [15, 16, 5, 11, 13, 22].

As a consequence, novel and efficient methodologies capable of automatedly clustering mtDNA sequences are must welcome. In the present work, we propose a new approach, based on a construction of Dawson using mtDNA protein coding region, which is a very stable and length-conserved region in humans. Our application uses the same design of variable sites as standard manual technique. However, it is made by a computer and so is much faster, friendlier, and human error-avoiding.

2. Methods

2.1. **Dawson construction.** Let $\mathcal{B}$ be a collection of subsets of $[n] := \{1, 2, \ldots, n\}$ for a positive integer n. For example, let $n = 7$ and $\mathcal{B}$ be the set formed by the following 15 subsets of [7]:

$\{1,2,4,5\}$	$\{1,2,4,6\}$	$\{1,2,4,7\}$	$\{1,2,5,6\}$	$\{1,2,6,7\}$
$\{1,3,4,5\}$	$\{1,3,4,6\}$	$\{1,3,6\}$	$\{1,3,6,7\}$	$\{1,4,6,7\}$
$\{1,5,6,7\}$	$\{2,3,4,5\}$	$\{2,5,6,7\}$	$\{3,5,6,7\}$	$\{4,5,6,7\}$

From $\mathcal{B}$, we construct a rooted binary labeled tree D, that we call the *Dawson tree*. This graph is a *tree*, so that any pair of vertices is connected by a unique sequence of adjacent vertices without repetition [1]. Both the vertices and the edges of D are labeled. We label the vertices of D either with integers between 1 and n or with (all) the elements of $\mathcal{B}$. The edges are labeled with v or with $\overline{v}$ for $v \in [n]$.

In fact, from each vertex labeled with $v \in [n]$ there are two edges leaving the vertex [2] which are labeled v and $\overline{v}$, (see Figure 1). No edge leaves from the vertices labeled with the elements of $\mathcal{B}$.

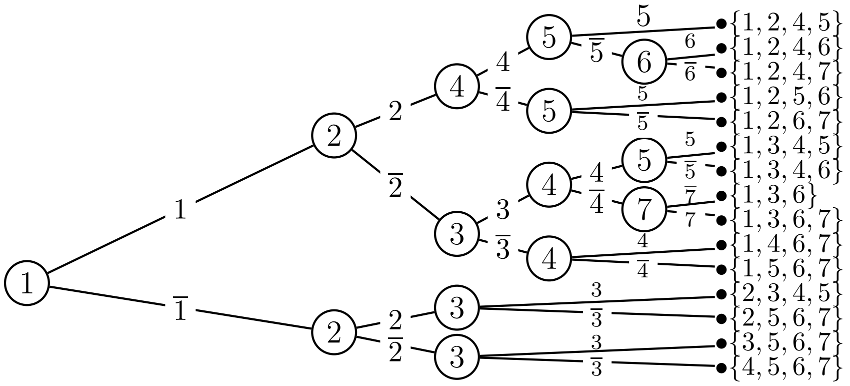


FIGURE 1. Dawson's tree D

This tree is easy to build. We first sort every element of $\mathcal{B}$ in increasing order and then sort $\mathcal{B}$ *in lexicographical order.*

Let u be the least element of $\widetilde{\mathcal{B}} := (\cup\mathcal{B})\backslash(\cap\mathcal{B})$. We start the tree with a vertex labeled u. This vertex is the *root.* In our case, $u = 1$. Now, we split $\mathcal{B}$ in two sets: the set $\mathcal{B}_u$ of those elements containing u and the set $\mathcal{B}_{\overline{u}}$ of those elements not containing u; note that each element of the first set precedes every element of the second set in lexicographical order. In our case $\mathcal{B}_1$ is formed by the first eleven elements and $\mathcal{B}_{\overline{1}}$ by the remaining four.

Now, we look for the least element v_1 of $\cup\mathcal{B}_u \setminus \cap\mathcal{B}_u$ and for the least element v_2 of $\cup\mathcal{B}_{\overline{u}} \setminus \cap\mathcal{B}_{\overline{u}}$; in our example, $v_1 = v_2 = 2$. We create an

[1] Two vertices are *adjacent* if they belong to the same edge

[2] We say that the edge $\{vw\}$ *leaves* v if w does not belong to the (unique) sequence of non-repeating adjacent vertices connecting v with the root.

edge labeled u connecting u to a new vertex labeled v_1 and an edge labeled $\overline{u}$ connecting u to another new vertex labeled v_2.

We iterate this process while either of the new sets corresponding to $\mathcal{B}_1$ and $\mathcal{B}_{\overline{1}}$ contains more than one element.

This means that we split $\mathcal{B}_u$ [resp. $\mathcal{B}_{\overline{u}}$] in two subsets, $\mathcal{B}_{uv_1}$ and $\mathcal{B}_{u\overline{v_1}}$ [resp. $\mathcal{B}_{\overline{u}v_2}$ and $\mathcal{B}_{\overline{uv_2}}$], and chose the least element corresponding to each case (in our example, we obtain $c_1 = 4$ and $c_2 = c_3 = c_4 = 3$, etc.).

More precisely, the procedure is as follows, where the initial values are $C = \mathcal{B}$ and $o = 1$.

```
proc($C, o$)
[Inputs a collection $C$ of subsets of $[n]$ sorted in lexicographical order and
an integer $o$ (for order) and returns a vertex of the growing graph $G$]
  (1) If $|C| = 1$, let $C = \{B\}$,
      (a) Add a new vertex $v$ to $V$, labeled with $B$;
      (b) Return($v$);
  (2) Let $C_2$ be the ordered set of the $o$.th integers of the elements of $C$;
      (a) If $|C_2| = 1$,
            (i) Return(proc($C, o+1$));
      (b) If $|C_2| > 1$, let $a$ be the least element of $C_2$,
          let $C(a, o)$ be the set of elements of $C$ with $o$.th integer equal to $a$, and
          let $D(a, o) = C \setminus C(a, o)$. Then:
            (i) Add a new vertex $v$ to $V$, labeled with $a$;
            (ii) $E \leftarrow E \cup \big\{\{v, \mathbf{proc}(C(a,o), o+1)\}\big\}$; label this edge with $a$.
            (iii) $E \leftarrow E \cup \big\{\{v, \mathbf{proc}(D(a,o), o)\}\big\}$; label this edge with $\overline{a}$.
            (iv) Return($v$);
```

The effect of this procedure can be described as follows: Consider, for an element B of $\mathcal{B}$, the "word" $\mathrm{W}(B)$ formed by the consecutive labels of the unique sequence of adjacent edges that starts at the root and ends with the vertex labeled with B. In the example,

$$\begin{aligned}\mathrm{W}\big(\{1,2,4,5\}\big) &= 1\,2\,4\,5\\ \mathrm{W}\big(\{1,3,4,6\}\big) &= 1\,\overline{2}\,3\,4\,\overline{5}\\ \mathrm{W}\big(\{2,5,6,7\}\big) &= \overline{1}\,2\,\overline{3}\end{aligned}$$

Note that $\mathrm{W}(A) = \mathrm{W}(B)$ if and only if $A = B$, by construction. Note also that if $\mathrm{W}(B) = a_1\,a_2 \cdots a_k$ (where some of the integers may be overlined) then, by construction, $1 \le a_1 < \cdots < a_k \le n$)

We can easily verify that $\{1, 3, 4, 6\}$ is the only element of $\mathcal{B}$ to which 1, 3 and 4 belong, but neither 2 nor 5. It can be shown that the same happens in general.

Theorem 2.1. *Let, for an element $B \in \mathcal{B}$,* $\mathrm{W}(B) = a_1\, a_2\, \cdots\, a_k$ *, and let $P_B = \{a_{i_1}, a_{i_2}, \ldots, a_{i_\ell}\}$ be the set of integers that are* not *overlined in* $\mathrm{W}(B)$ *and $N_B = \{a_1, \ldots, a_k\} \setminus P_B$ be the set of integers that are overlined in* $\mathrm{W}(B)$*. Then*

$$\{B\} = \Big\{A \in \mathcal{B} \colon P_B \subset A, A \cap N_B = \varnothing\Big\} \tag{1}$$

Proof. Clearly, $P_B \subset B$ and $B \cap N_B = \varnothing$. For the converse, suppose that, for some $A \in \mathcal{B}$ with $A \neq B$, $P_B \subset A$ and $A \cap N_B = \varnothing$, and that $\mathrm{W}(A)$ coincides with $\mathrm{W}(B)$ up to a certain step. The edges of this common part (if any) connect the root to a vertex labeled with integer i, say. Then, the first difference of $\mathrm{W}(A)$ and $\mathrm{W}(B)$ is that i is overlined in one of the words and not in the other. Suppose that i is overlined in A but not in B. This means that $i \notin A$ but $i \in P_B$, contrary to our assumptions. If i is overlined in B but not in A then $i \in N_B \cap A$, also contrary to our assumptions. Hence it must be $A = B$. □

This is the base of Dawson's construction in [3]. For a recent update, see [2]. To be more precise, denote by $A\Delta B$ the *symmetric difference* of A and B which consists of the elements that belong exactly to one of these sets (i.e., $A\Delta B := (A \setminus B) \cup (B \setminus A)$), and define for $A \subset [n]$ and for $B \in \mathcal{B}$:

$$\begin{aligned} I_B &:= \Big\{A \subset [n] \colon P_B \subset A, N_B \cap A = \varnothing\Big\} \\ \omega(A) &:= \sum_{i \in A} 2^{n-i} \end{aligned} \tag{2}$$

Then, the collection $\{I_B\}_{B \in \mathcal{B}}$ forms the partition of the set $\mathcal{P}([n])$ of subsets of $[n]$ introduced in [3] or, in other words,

Theorem 2.2 ([3])**.**

$$I_B = \big\{A \subset [n] \colon \forall_{B' \in \mathcal{B}},\ \textit{if } B' \neq B \textit{ then } \omega(A\Delta B') > \omega(A\Delta B)\big\}.$$

Proof. First we prove that, for any $A \subset [n]$ and for any $B' \in \mathcal{B}$ different from B, $P_B \subset A$ and $N_B \cap A = \varnothing$ implies that $\omega(A\Delta B') > \omega(A\Delta B)$. Now, similarly to the proof of Theorem 2.1, consider the least integer i relatively to which $\mathrm{W}(B)$ and $\mathrm{W}(B')$ differ, say, i is overlined in $\mathrm{W}(B)$ but not in $\mathrm{W}(B')$, that is, $\underline{i \in N_B \cap P_{B'}}$. Then $i \notin A$, $i \notin B$ and $i \in B'$. Hence, $i \in A\Delta B'$, $i \notin A\Delta B$ and i is the least element in $B\Delta B'$. Therefore,

$$\omega(A\Delta B') = \sum_{j \in A\Delta B'} 2^{n-j} \underline{\geq 2^{n-i} > 2^{n-i} - 1} \geq \sum_{j \in A\Delta B} 2^{n-j} = \omega(A\Delta B)\,.$$

The case where i is overlined in $\mathrm{W}(B')$ but not in $\mathrm{W}(B)$ follows similarly.

For the converse, suppose that $\omega(A\Delta B') > \omega(A\Delta B)$ for all $B' \in \mathcal{B}$ such that $B' \neq B$, but $A \notin I_B$ because P_B is not a subset of A or because $N_B \cap A \neq \varnothing$. Consider the first element i either in $P_B \setminus A$ or in $A \cap N_B$ and note that $B \setminus A \supset P_B \setminus A$ and $A \cap N_B \subset A \setminus B$. Hence, $i \in A \Delta B$. Now, since $i \in P_B \cup N_B$, overlined or not it belongs to $\mathrm{W}(B)$, and there exists at least one element $B' \in \mathcal{B}$ such that $\mathrm{W}(B')$ differs from $\mathrm{W}(B)$ in that i is overlined in B but not in B' or vice-versa but does not differ in the subwords previous to i. Note that $i \in N_{B'}$ if and only if $i \in P_B$. Then, the elements of $B \Delta B'$ are greater than i. But this is in contradiction to the fact that $\omega(A \Delta B') > \omega(A \Delta B)$. □

It follows from (1) that Dawson's construction (through Dawson's tree) is a general "dichotomous key" for the elements of $\mathcal{B}$. Of course, changing the order in $[n] = \{1, \dots, n\}$ changes necessarily the key.

Our purpose in here is to adapt this construction for using it in the classification of genetic haplotypes. The invention of dichotomous keys in Biology is quite old: it is generally credited to Jean-Baptiste Lamarck, who used one in the first edition of *Flora Française*, in 1778; but in fact it goes back at least to March 1689, when Richard Waller presented in the Royal Society an image-based dichotomous key for the "English herbs" [6].

In fact, *bifurcating* phylogenetic trees are dichotomous trees that approximate Dawson's trees in the following sense: suppose that a bifurcating phylogenetic tree of various human mitochondrial ADN, for example, exists, in which each position of the "coding region" either does not change or changes between only two possible bases. As an example, suppose the first position of the coding region is in some haplogroups, say, A, and is T, say, in the remaining haplogroups. Furthermore, suppose that the mutations occur at most once in each position and that we know in which order they have occurred. Finally, suppose that, for each position that has changed, we fix one of the two bases it has changed within and that we represent the haplogroups (even "missing links") by the positions for which the base is the given one, according to the order in which the mutations occurred. Then, clearly, the bifurcating phylogenetic tree and the Dawson's tree coincide.

Our interest in the Dawson's construction lies not only on an approximation of this, that we explore below, but on using some of the known properties of the construction. For example, in (2), the elements of the partition of $\mathcal{P}([n])$ are defined based on the elements $B \in \mathcal{B}$, that may be called the *seeds* of the partition. In [7], a characterization of the sets of seeds that produce the same partition is given. It follows from this characterization that we know, at least in theory, how to narrow down the number of positions that have to be compared. At the moment, however, this has not yet been thoroughly studied.

2.2. **A Dawson-like classification of human mitochondrial DNA.** The method we report in this paper may be shortly described as being in the line of the procedure that was previously described, but, contrary to $\{1, \dots, n\}$, of course, in a set without a clear order. Hence, in step 2.(b) of the description of the procedure, we have to define the element that we take in the place of the integer a.

In fact, with our method the elements of $\mathcal{B}$ are replaced by the mitochondrial DNA of the coding region of different human haplogroups and the vertices are to be labeled by different positions of the coding region.

For this collection,

(1) We have considered all positions where at least two bases were present (in fact, only very rarely a third base was found in a given position in our 99 samples).
(2) Then, by counting the number of occurrences of each base in the given position for the 99 samples, we arranged the data in a 0/1 table, where a zero in a given haplotype in a given position meant that in that specific position the haplotype presented the most frequent base. As in the previous example, in this construction a given haplotype is seen as the set of positions for which the table entry is 1 (without distinction as to whether there are two or more bases for the position).
(3) The vertices of the tree are either labeled by positions or by the 99 haplotypes, one for each haplotype. The graph separates, for every vertex labeled with position p, the haplotypes where the base in the given position is not the most frequent (the 1's) from the ones with the most frequent base in the position (the 0's). The edges are labeled accordingly, p in the first case and $\overline{p}$ in the second one. By definition, after this separation two new procedures start, one for each of these sets of haplotypes, that form the two new initial values of C.
(4) Hence, in each step (for a given C and p), we must try to find the position that better separates the haplotypes under consideration (that form C), coresponding to the "oldest mutation" in the previous example. This is the crux of all construction and is implemented according to the procedure **Choose**(C, D, ℓ, o) described below. Shortly, this is how this is done.

    Note that, by definition, haplogroups that are "really different" must be separated by various positions. So, we count for each position the number of positions that separate groups in C in a similar manner, and choose the one for which this number is maximum. The precise way that we use will be explained later. It suffices to tell now that we use two variables, ℓ and D in this calculation. Their meaning is as follows: suppose that the position we are looking for will label the vertex v; ℓ is the number of vertices in the sequence of adjacent edges that goes from the root to v, including the root. As for D, it is the set of all haplotypes with a 1 in any of these vertices. Note that ℓ

and D are automatically constructed by **proc2**, also described below.

So, we define the new procedure.

```
proc2($C, D, \ell, o$)
[Inputs a collection $C$ of subsets of $[n]$, a subset $D$ of $[n]$, two integers, $\ell$ and $o$, and returns a vertex of the growing graph $G$.]
  (1) If $|C| = 1$, say, $C = \{B\}$,
      (a) Add a new vertex $v$ to $V$, labeled with $B$;
      (b) Return($v$);
  (2) Let $C_2$ be the ordered set of the $o$.th integers of the elements of $C$;
      (a) If $|C_2| = 1$,
          (i) Return(proc($C, o+1$));
      (b) If $|C_2| > 1$, choose a suitable position $a = $ Choose($C, D, \ell$);
          let $C(a,o)$ be the set of elements of $C$ with $o$.th integer equal to $a$ and
          let $D(a,o) = C \setminus C(a,o)$. Then:
          (i) Add a new vertex $v$ to $V$, labeled with $a$;
          (ii) $E \leftarrow E \cup \big\{\{v,$ proc2$(C(a,o), D \cup C(a,o), \ell+1, o+1)\}\big\}$; label this edge with $a$.
          (iii) $E \leftarrow E \cup \big\{\{v,$ proc2$(D(a,o), D \cup C(a,o), \ell+1, o)\}\big\}$; label this edge with $\overline{a}$.
          (iv) Return($v$);
```

The initial values are $C = \mathcal{B}$, $D = \varnothing$, $\ell = 1$ and $o = 1$.

Let us explain now the procedure **Choose**(C, D, ℓ): as we said, it chooses the position that better separates the elements of C, in a certain sense. In fact, it evaluates for each position for which the haplotypes in C vary a value that we call the weight of p, and afterwards it chooses the position with highest weight. Two positions with ones and zeros in the same haplotypes count with one to each other weight; if the set of ones of a position is contained in the set of ones in another one, it counts *more* for the weight of the second one than the second one counts for the weight of the first one. Finally, every position has an initial value that measures how close it is to the previous divisions. More precisely:

Let $\mathbf{1}_q$ be the set of haplotypes with position q equal to one. For a given position p, let

$$w(p) = \min\Big(\ell, \frac{|D \cap \mathbf{1}_p|}{|C \cap \mathbf{1}_p|}\Big);$$

and, for each postion q, add to $w(p)$

$$\max\left(0, \frac{|\mathbf{1}_p \cap \mathbf{1}_q| - |\mathbf{1}_q \setminus \mathbf{1}_p| - \frac{|\mathbf{1}_q \setminus \mathbf{1}_p|}{2}}{|C \cap (\mathbf{1}_p \cup \mathbf{1}_q)|}\right)$$

2.3. **Sequences.** A training set consisting of 99 human mtDNA protein coding sequences, comprising representatives of all major haplogroups [14], was used. The algorithm was written in Mathematica and tested on a Linux operative system.

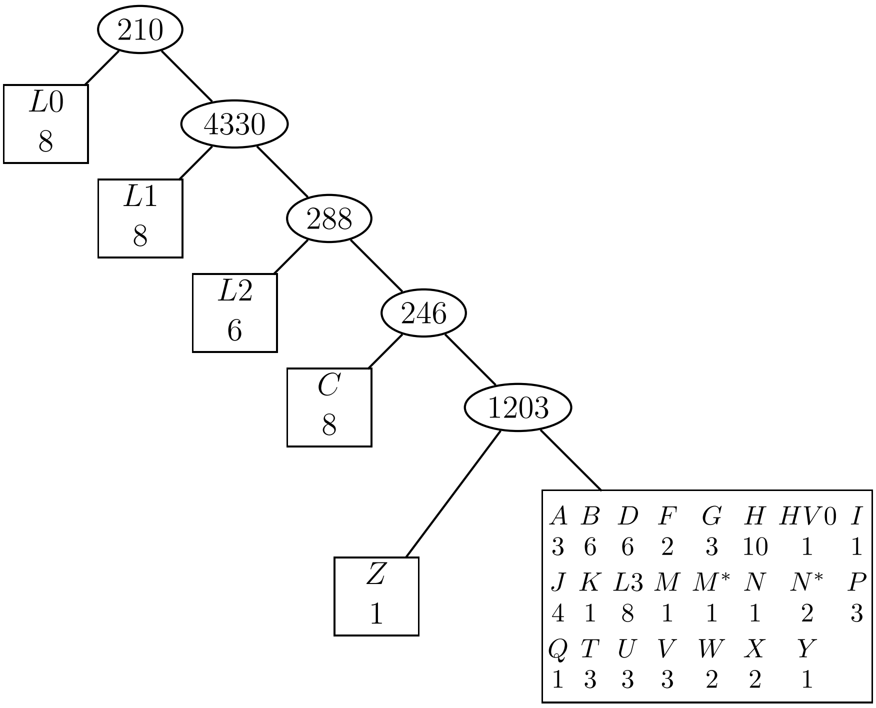
3. Results

Our program was applied to a collection of 99 different samples of human mtDNA protein coding sequences, encompassing representatives of all known macro-haplogroups [8, 14]. The generated Dawson-like tree, constructed with base on the computed order explained with detail in the Methods section, can be seen in Figure 2. The Dawson-like tree produced by our method show that the sequences are clustered into major haplogroups according to the canonical organization stipulated by Phylotree (http://www.phylotree.org) [19].

4. Discussion

The proposed methodology represents a new and fast approach capable of clustering mtDNA sequences according their macro-haplogroup not very different from the frequently used manual ideas but without human intervention. The construction projected by Dawson represents a simple and fast technique to automatically cluster a set of subsets, just indicating the presence/absence of each element in the data set. To be possible the application of this promising methodology to mtDNA sequences, we chose the protein coding region because it is the most stable region of human mtDNA sequences not presenting length variation, and thus we just need to compare homologous positions not requiring explicit assumptions to interpret the noised and subjective events of control region.

The set of variable positions of coding region are the only relevant ones which are capable of distinguishing and of separating the sequences among them. Thus, the methodology developed was based only in these positions, interpreting just substitution events, which are the only mutational phenomena occurring in this region. The method does not require knowing which base is present in every variable position, instead, it just needs to know if the base is the most frequent or not in a specific position of the sequences. This is the obvious motif why the bases A, C, G and T were ignored and replaced by digits 0 and 1, in the case of the base is the most frequent or not in the scanned position. Assuming this, we circumvented the circular assumptions behind the



A	B	D	F	G	H	$HV0$	I
3	6	6	2	3	10	1	1
J	K	$L3$	M	M^*	N	N^*	P
4	1	8	1	1	1	2	3
Q	T	U	V	W	X	Y	
1	3	3	3	2	2	1	

FIGURE 2. Dawson-like tree of 99 human mtDNA protein coding sequences. This is the result of recursively using **proc2**(C, D, ℓ, o). First, $a = 210$ is returned by **Choose**(C, D, ℓ); $C(a, o)$ contains 8 elements that turn out to be of the same haplotype ($L0$) according to Phylotree. Now, **proc2**(C, D, ℓ, o) proceeds on $D(a, o) = C \setminus C(a, o)$, where $a = 4330$ is chosen, etc.

estimations of transitions and transversions mutation rates, which are frequently evaluated based on the sample. In fact, the present study relies exclusively in the presence/absence of a series of equally weigh substitution events, not biasing the comparison and cluster results.

The construction proposed by Dawson has as foundation a lexicographical order of the elements and, thus, the quality of the construction has a direct correlation with the order adopted to perform the analyses. Therefore, the admitted order for variable positions, in the present adaptation of Dawson's construction to cluster mtDNA sequences, is the unique assumption imposed, being the most relevant computation capable of processing with good results. The order proposed and used to perform the presented analyses is straightaway and the dependence with the sample, if exist, is present in an inconsequential scale, being thus a good choice to perform the analyses leading to well satisfactory results.

In summary, our approach showed a good and fast way, devoiced of manual data manipulation, capable of mimicking the standard practice of assignment human mtDNA sequences into haplogroups, clustering the sequences according their classifications into major groups, just using the reliable information of the protein coding regions, being thus a good suggestion when aiming to group human mtDNA sequences into broad related sets.

In the future, we intend to test our Dawson's approach with a large dataset and also in subsets arbitrarily sampled from it, in order to understand and analyze the performance of the positions order on the clusters obtained. Furthermore, we aim to be capable of getting the subset of positions that better characterize the different groups in the entire clustering, independently of the sample used. Moreover, the study of more heterogeneous groups or sequences (including various species) would be essential to evaluate the usefulness of this approach.

Acknowledgements. This work was partially supported by IPATIMUP/i3S (UID/BIM/04293/2019), CMUP (UID/MAT/00144/2019), and CIDMA (UID/MAT/04106/2019), which are funded by FCT (Portugal) with national (ME) and European structural funds through the programs FEDER, under the partnership agreement PT2020. IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT in the framework of the project Institute for Research and Innovation in Health Sciences (POCI-01-0145-FEDER-007274).

References

[1] S. Batzoglou: The many faces of sequence alignment, *Brief Bioinform* **6.1** (2005) pp. 6–22.

[2] J. Brunat, A. Guedes de Oliveira, M. Noy: Partitions of a finite Boolean lattice into intervals, *European J. Combin.* **30** (2009) pp. 1801–1809.

[3] J.E. Dawson: A construction for a family of sets and its application to matroids, *Combinatorial Mathematics VIII (Geelong, 1980)*, in: L.L. MacAvaney (Ed.) Lect. Notes Math. vol. 884 Springer, Berlin-New York (1981) pp. 136–147.

[4] M. Domazet-Loso and B. Haubold: Efficient estimation of pairwise distances between genomes, *Bioinformatics* **25.24** (2009) pp. 3221–3227.

[5] L. Fan, Long, Y-G Yao: An update to MitoTool: Using a new scoring system for faster mtDNA haplogroup determination, *Mitochondrion* **13** (2013) pp. 360–363.

[6] L. R. Griffing: Who invented the dichotomous key? Richard Waller's watercolors of the herbs of Britain , *Am. J. Bot.* **98** (2011) pp. 1911–1923.

[7] A. Guedes de Oliveira and D. Oliveira e Silva: Note on the integer geometry of bitwise XOR, *European J. Combin.* **26** (2005) pp. 755–763.

[8] D. Mishmar: Natural selection shaped regional mtDNA variation in humans, *Proceedings of the National Academy of Sciences* **100.1** (2003) pp. 171–176.

[9] B. Nabholz, S. Glemin and N. Galtier: Strong Variations of Mitochondrial Mutation Rate across Mammals — the Longevity Hypothesis, *Molecular Biology and Evolution* **25.1** (2007) pp. 120–130.

[10] A. K. Reeve, K. J. Krishnan and D. Turnbull: Mitochondrial DNA Mutations in Disease, Aging, and Neurodegeneration, *Annals of the New York Academy of Sciences* **1147.1** (2008) pp. 21–29.

[11] A. W. Röck, A. Dür, M. van Oven, W. Parson: Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA), *Forensic Science International: Genetics* **7** (2013) pp. 601–609.

[12] G. E. Sims, S. R. Jun, G. A. Wu and S. H. Kim: Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions, *Proceedings of the National Academy of Sciences* **106.8** (2008) pp. 2677–2682.

[13] I. Soares and A. Amorim: Sequence Comparison, Classification and Phylogeny: Avoiding Hidden Assumptions and Speeding up Analyses *J. Comput. Eng. Inf. Technol.* **5.3** (2016).

[14] I. Soares, A. Amorim, and A. Goios: A new algorithm for mtDNA sequence clustering, *Forensic Science International: Genetics Supplement Series* **3.1** (2011) pp. e315–e316.

[15] I. Soares, A. Amorim, and A. Goios: mtDNAoffice: a software to assign human mtDNA macro haplogroups through automated analysis of the protein coding region *Mitochondrion 12.6* (2012) pp.666–8.

[16] I. Soares, A. Goios, and A. Amorim: Sequence Comparison Alignment-Free Approach Based on Suffix Tree and L-Words Frequency *The Scientific World Journal* **2012** (2012) Art. ID 450124, 4 pages.

[17] R. W. Taylor, M. T. McDonnell, E. L. Blakely, P. F. Chinnery, G. A. Taylor, N. Howell, M. Zeviani, E. Briem, F. Carrara and D. M. Turnbull: Genotypes from patients indicate no paternal mitochondrial DNA contribution, *Annals of Neurology* **54.4** (2003) pp. 521–524.

[18] J. D. Thompson, D. G. Higgins and T. J. Gibson: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research* **22.22** (1994) pp. 4673–4680.

[19] M. van Oven, and M. Kayser: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, *Human Mutation* **30.2** (2009) pp. E386–E394.

[20] S. Vinga and J. Almeida: Alignment-free sequence comparison–a review, *Bioinformatics* **19.4** (2003) pp. 513–523.

[21] H. Weissensteiner Hansi, L. Forer, C. Fuchsberger, B. Schpf, A. Kloss-Brandsttter, G. Specht, F. Kronenberg, and S. Schnherr: mtDNA-Server: Next-Generation Sequencing Data Analysis of Human Mitochondrial DNA in the Cloud *Nucleic Acids Research* (2016).

[22] H. Weissensteiner, D. Pacher, A. Kloss-Brandsttter, L. Forer, G. Specht, H. Bandelt, F. Kronenberg, A. Salas, and S. Schnherr: HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing, *Nucleic Acids Research* (2016).