Four billion years: the story of an ancient protein family

Gilles Didier, Claudine Landès, Alain Hénaut, Bruno Torrésani

Four billion years: the story of an ancient protein family

Gilles Didier, Claudine Landès, Alain Hénaut and Bruno Torrésani

Abstract The comparison of protein sequences has been for long a very effective tool in producing biological knowledge. It was initially based on the alignment of sequences, that is to say organizing the set of sequences in columns (of a spreadsheet) of sites which have evolved from a common site of the ancestral sequence. Alignments are generally obtained by minimizing an evolution or an edition cost. Sequence comparisons are now often performed without alignments by comparing the N-mer compositions of the sequences. We present here the most popular methods used by biologists to compare sequences and place emphasis on an approach to augment the alphabet of a set of sequences in order to ease their comparison. The family of DNA topoisomerases, a set of ancient proteins whose history can be traced back 4 billion years, is used to illustrate this approach.

1 Introduction

The DNA double helix has a very stable structure. For example, dissociating the two strands results in the creation of supercoils that stop their separation, which is clearly a bonus for the conservation of the genetic inheritance. It poses a problem, however, when the two strands need to be dissociated in order to be copied [45]. This topological problem is settled in vivo by proteins, the DNA topoisomerases,

Gilles Didier
IMAG, Univ Montpellier, CNRS, Montpellier, France. e-mail: gilles.didier@umontpellier.fr

Claudine Landès
Univ Angers, Institut Agro, INRAE, IRHS, SFR QUASAV, F-49000 Angers, France. e-mail: claudine.landes@inrae.fr

Alain Hénaut
Université Publique Française, France. e-mail: alainhenaut@yahoo.fr

Bruno Torrésani
Aix Marseille Univ, CNRS, I2M, Marseille, France. e-mail: bruno.torresani@univ-amu.fr

that modify the supercoils. They trigger a transient cut of the DNA on one of the strands for type I topoisomerases, on both strands for type II topoisomerases.

There are several classes of topoisomerases. Those belonging to class IA are found among all living species and have probably existed for 4 billion years. They form a multigenic family that has been undergoing a series of duplications in the course of evolution [20, 18, 1, 4], see **Fig. 1**.

How can biologists trace back the history of such an ancient family of proteins? In other words, how can they draw up the phylogeny[1] of the DNA topoisomerases IA?

It has been known since 1965 that it is possible to reconstruct phylogenies through sequence comparisons[2]. This is due to the fact that those proteins which possess the same function in closely related species do exhibit very similar sequences. In two closely related sequences the amino acids will be generally the same at a given position but they can also differ – the result of a mutation. The total number of differences (mutations) between two sequences being as a first approximation proportional to the time of divergence between the two species [49], they provide a useful information to reconstruct their phylogeny. Things become more complicated when the sequences have diverged a long time ago since several mutations may have occurred over time at the same site. Deciphering the phylogeny of species being a central concern in biology, it is not surprising that a number of different methods would have been developed in this respect.

This article describes the principles of these methods and shows how original Alex's approach is.

2 Sequence comparisons with alignment

2.1 Pairwise alignment of two sequences

Take two sequences. Modify (edit) the first sequence so that you end up with the second sequence and count the minimal number of modifications that are necessary to go from the first to the second: you have performed the pairwise alignment of the two sequences. The modifications that are relevant in biology are 1) the substitution of one letter by one another (the substitution of an amino acid by another one at a given site) and 2) the insertion or the deletion of one or several letters at a given site (which results in the creation of a gap at that position in one of the two sequences).

---

[1] Phylogeny is the study of the degree of relationship between living organisms, which enables to reconstruct their evolution. In a phylogenetic tree, the nodes represent the common ancestors. The greater the number of nodes between two taxa, the more ancient is their common ancestor and the farther they are in the tree of life – the length of the branches is approximately proportional to the time of divergence between the taxa [22, 40].

[2] Proteins are macromolecules composed of a linear string of amino acids. They are generally made of several hundreds of the 20 different amino acids.

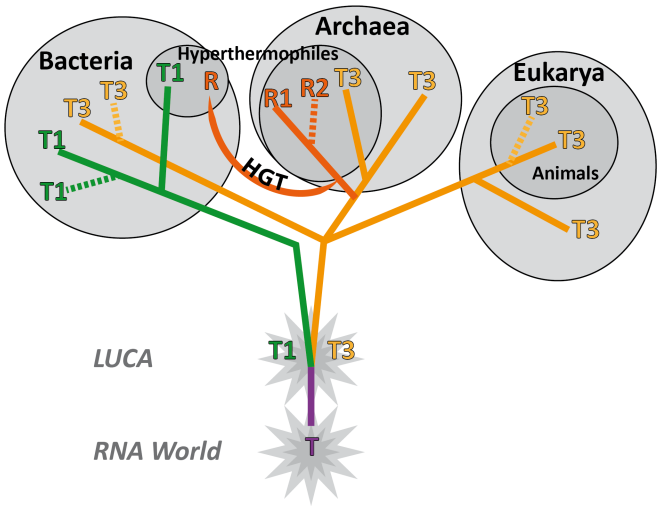


Fig. 1 Hypothetical origin and distribution of Topoisomerases IA within the tree of life. T1, T3, and R indicate the TopoI, TopoIII and reverse gyrase, respectively. The dashed lines indicate that only a part of the organisms belonging to that branch possess such enzymes. LUCA: *Last Universal Common Ancestor* is a theoretical construct - it might or might not have been something we today would call an organism. HGT: *horizontal gene transfer* or *lateral gene transfer*; the movement of genes between distantly related organisms. T represents the ancestor of all of the TopoIAs (Figure taken from [20]).

.

A cost is associated with each modification, which enables to obtain a score when the alignment is completed.

The first algorithm explicitly devoted to the pairwise alignment of biological sequences was devised by Needleman and Wunsch in 1970 [31]. This dynamic programming algorithm aims at aligning two sequences over their whole lengths (a global alignment) and guarantees that the resulting alignment score is maximal. However, a global alignment is not relevant if the similarity spreads only over a limited length. Smith and Waterman suggested in 1981 to look only for the regions of greatest similarity between two sequences and report only those regions (a local alignment) [41]. Their algorithm is a derivative of that by Needleman and Wunsch. The cost optimized by these algorithms mainly relies on two parameters:

- a scoring matrix giving, for each pair of amino acids, a score that was estimated from:
  - the substitutions that have been observed in well-known protein families or
  - the chemical properties of the amino acids [37, 34]
- a penalty for the creation of a gap (and possibly another one for its extension).

A scoring matrix between amino acids is normally built from experimental data [11]. Unfortunately, the gap penalties are totally empirical – one may say that it's just been cobbled together [35, 42, p. 56].

The programs based on dynamic programming are slow and can hardly be used to perform pairwise alignments of hundreds or thousands of sequences. Thus some heuristics were devised at the end of the 1980s to speed up the comparisons without sacrificing too much from sensitivity or specificity. The result are local alignment programs that proceed in two steps: 1) look for all the "regions of similarity" between two sequences, 2) if possible, join together all those regions in order to obtain a longer alignment. Many of these programs look for segments that are identical in both sequences and keep from this list only those pairs of segments that are equidistant in both sequences – thus allowing no gaps between them. The idea here is that two closely related sequences deriving from a common ancestor will not suffer from insertions/deletions in conserved regions. These segments are then used as anchoring points for a further local alignment.

Some programs search for strictly identical segments of a given length, some others allow for some flexibility and accept "almost identical" segments. "Almost identical segments" are precisely defined through the use of a scoring matrix or a reduced amino acid alphabet. In the last case the number of amino acids is reduced from twenty to six or even four: acidic, basic, polar (hydrophilic) and apolar (hydrophobic) [34]. The use of a reduced alphabet speeds up the alignment process at the expense of specificity.

The most famous programs for massive pairwise comparisons are FASTA [33] and BLAST [2], created in 1988 and 1990, respectively. Even if BLAST was inspired by FASTA, the two programs present an important difference: BLAST searches for similar words of three consecutive amino acids while FASTA searches for strictly identical words of two consecutive amino acids.

BLAST is by far the most popular. It is generally used to look for similarities between a given sequence and all those that are gathered in databanks (at the end of 2020, the protein sequence data banks contained 186 billion sequences).

2.2 Simultaneous alignment of several sequences

In theory, the Smith and Waterman algorithm allows for the simultaneous alignment of several sequences (which is called a multiple alignment) but in practice this is feasible for only a very limited number of sequences (the multiple alignment problem is NP-hard [25]). One has therefore to resort to heuristics [6], the vast majority of them following the same path: 1) computation of the similarity between each pair of sequences through pairwise alignments; 2) creation of an ascending hierarchical tree using the similarity matrix based on the scores of the pairwise alignments. This tree then sets the order in which the sequences will be aggregated: i) choose the pair of closest sequences in the tree; ii) align these sequences; iii) replace in the tree the pair of sequences by this alignment. Hence, during the course of the algorithm, one

may align a sequence with a group of sequences that have already be aligned, or even align a group of sequences with another group. The root of the tree harbors the alignment of all the sequences.

There are two major steps for the alignment of one sequence with a group of sequences or a group of sequences with another group:

- the group is represented by a position sensitive substitution matrix (PSSM) where each element depends on the nature of the two amino acids which are in the same column as well as on the position of the column in the multiple alignment [39, 21]; the alignment is performed in a classical way
- one aligns the "regions of similarity" that were found in the first step. If a gap must be introduced, it is placed at the same position in all the sequences of the group

Eventually, after completion of the multiple alignment, the matrix of similarity between all the sequences is edited as a phylogenetic tree. Some relevant algorithms are available for this purpose [43, 32].

To sum up, the biologist is offered a host of options to perform a multiple alignment: global or local alignment, strict identity or mere similarity, scoring matrices or reduced alphabet, single-, average- or complete-linkage hierarchical classification, to name a few (see [3] for a review of the most popular programs). If the sequences are closely similar, all the options will provide essentially the same results. But if the sequences are rather dissimilar, the alignments will heavily depend on the chosen options. In such cases the biologists will take advantage of additional information provided by a deep knowledge of the protein family (some adjustments may be done by hand) or by the 3D structures of some of the proteins in the set (if available) or even by a tentative and approximate prediction of the foldings of the proteins – a computationally intensive task [24, 7].

3 Alignment-free sequence comparison and local decoding

The concerns raised in the section above have motivated the development of approaches to compare sequences without aligning them. A natural way to do this is to compare sequences with regard to their composition. Since considering the frequencies of the 20 amino acids is not discriminating enough, one rather considers their composition in words of a given length N (so-called *N-mers*), i.e., by counting the number of times each word of length N occurs in each sequence to compare.

3.1 *N*-mers and *N*-local decoding

Though the N-mers-based comparisons may provide accurate approximations of evolutionary distances, they generally lack definition due to the fact that they do not

distinguish between the situation where two N-mers differ in only a few positions and the situation where they are completely unrelated. This point can be improved by allowing mismatches between words. This leads to the question of deciding to what extend two different N-mers can be considered "equals" in a sequence comparison context. Note that this question is still being actively investigated (e.g., [27, 47, 48, 16]).

The approach presented in [15] relies on the fact that the sequence of the overlapping successive N-mers of a given sequence s may generally be obtained as the sequence of the overlapping successive N-mers of many sequences (up to relabelling the N-mers), among which one is maximal in the sense that all the others (including s) may be obtained from it through letter-to-letter applications (not necessarily one-to-one, **Fig. 2**). It follows that the alphabet of this maximal sequence is greater than that of any other sequences whose overlapping successive N-mers sequence is the same as that of s, still up to relabelling. In particular, it is greater or equal to the alphabet of s. The *N-local decoding* of a sequence is this maximal sequence.

(1):	a	b	a	c	d	e	a	b	d	a	c	f	a	b	g
(2):	[ab]	[ba]	[ac]	[cd]	[de]	[ea]	[ab]	[bd]	[da]	[ac]	[cf]	[fa]	[ab]	[bg]	
(3):	0	1	2	3	4	5	0	6	7	2	8	9	0	10	
(4):	[ab]	[ba']	[a'c]	[cd]	[de]	[ea]	[ab]	[bd']	[d'a']	[a'c]	[cf]	[fa]	[ab]	[bg]	
(5):	a	b	a'	c	d	e	a	b	d'	a'	c	f	a	b	g

Fig. 2 (1): a sequence over $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}, \mathtt{f}, \mathtt{g}\}$; (2): the sequence of overlapping successive 2-mers of the sequence (1); (3): the sequence (2) with all letters replaced by the position of their first occurrence; (4): the sequence of overlapping successive 2-mers of a sequence different from sequence (1), which is the sequence (5) over $\{\mathtt{a}, \mathtt{a'}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{d'}, \mathtt{e}, \mathtt{f}, \mathtt{g}\}$. It can be proved that the sequence (5) is maximal in the sense that all the sequences with overlapping successive 2-mers sequence corresponding to (3) can be obtained from (5) by letter-letter applications [12]. For instance, going from (5) to (1) is done by the letter-to-letter application which discards the 'primes'.

The N-local decoding can alternatively be defined (and is computed) by considering the equivalence relations between the positions of the sequence(s) presented below. Let S be a set of sequences over a finite alphabet. Its *site space* Σ is the set of all pairs (s, p) where s is a sequence of S, and p a position in it, namely,

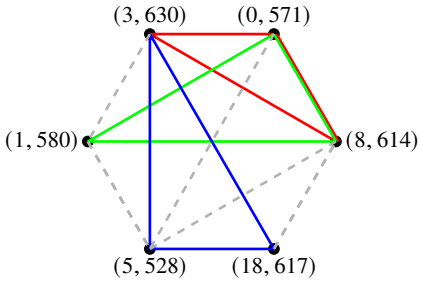
$$\Sigma = \big\{(s, p) \mid s \in S, 1 \leq p \leq \ell(s)\big\},$$

where $\ell(s)$ is the length of sequence s. For all positive integers N, we define the following relations between sites of S.

1. Two sites (s, p) and (s', p') in Σ are directly related if there exists a word w of length N occurring at positions $p - i$ and $p' - i$ of sequences s and s' respectively, with $i < N$. In other words, w overlaps both sites (s, p) and (s', p') with the same offset i. If two sites (s, p) and (s', p') are directly related, we write $(s, p) \simeq_N (s', p')$. Note that determining if the sites (s, p) and (s', p') are directly related only requires to consider their centered neighbourhoods of length $2N - 1$.

2. The equivalence relation $\sim_N$ is then defined as the transitive closure of $\simeq_N$. In other words, we say that $(s, p) \sim_N (s', p')$ if there is a chain of directly related sites connecting (s, p) and (s', p').

In order to illustrate these relations on an example, let us consider a set of protein sequences and examine one of the equivalence classes associated to the relation $\sim_N$ with $N = 7$ (**Fig. 3**). This class contains 6 sites. The first site is described by the pair $(0, 571)$: this means that it lies at position 571 of the sequence number "0", and similarly for the other five sites. Since $N = 7$, the neighborhoods around these sites used to determine the relations above are of length $2N - 1 = 13$ and are displayed in the figure with their central letter underlined.



(0, 571)	EVLREI<u>D</u>EDRKKV
(3, 630)	SILREI<u>D</u>EDREKV
(8, 614)	SVLREI<u>D</u>EDRKKV
(1, 580)	KLIKEI<u>D</u>EDRKRA
(5, 528)	ELIKKI<u>D</u>EDREKI
(18, 617)	EILRQI<u>D</u>EDREKV

LREI<u>D</u>ED EI<u>D</u>EDRK I<u>D</u>EDREK

Fig. 3 Graphical representation of relatedness within an $\sim_N$-class, with $N = 7$. For each one of six sites, the word occupying its centered neighborhood is listed on the right of the figure. Directly related sites are connected by solid lines: each color corresponds to a word of length 7 shared by at least two neighborhoods and displayed at the bottom-right. Dashed gray lines connect sites that are related by $\sim_N$ but not directly related by $\simeq_N$.

Directly related sites are connected by solid lines. For instance, the sites $(0, 571)$, $(3, 630)$ and $(8, 614)$ share the word LREI<u>D</u>ED starting at the third position of their neighbourhood. The sites that are related without being directly related are connected by dashed lines. For instance, the sites (1, 580) and (5, 528) are connected by the chain (1, 580) $\rightarrow$ (0, 571) $\rightarrow$ (3, 630) $\rightarrow$ (5, 528).

Theorem 1 ([15]) *Let S be a set of sequences and N a positive integer. Any set of sequences having the same successive overlapping N-mers sequences as S (up to a relabelling) can be obtained by a letter-to-letter application from the set of sequences obtained from S by putting at each site the (ident of the) class to which it belongs in the partition associated to the relation $\sim_N$.*

In other words, the N-local decoding of a set of sequences can be obtained by replacing the letter at each site of the set by the class of this site in the $\sim_N$-partition.

This approach can be applied to compare a set of sequences by considering their composition in symbols of its N-local decoding. The N-local decoding of a set of sequences can be seen as an intermediate level of information between this set and its sequences of N-mers. From a practical point of view, the N-local decoding of a

(set of) sequence(s) can be computed with a complexity linear with the length of the sequence(s) both in time and memory space whatever N and thus with the same complexity as that required to compute its N-mers [15, 13, 9, 8].

Both the N-mers and the N-local decoding approaches require the selection of a suitable value of the parameter N in order to compare a given set of sequences. For large values of N, each N-mer (and each symbol of the local decoding) tends to occur at most once in the set of sequence whereas small values of N, the N-mers occur at too many positions in the set to be actually informative for comparison purposes. A basic solution is to try several values of N and to select the value which seems the most relevant with regard to the results observed.

The MS4 approach, presented in Section 3.2, was designed to tackle this issue (MS4 stands for *Multi-Scale Selector of Sequence Signatures*). MS4 selects for a given site the smallest N such that the average number of occurrences per sequence of the equivalence class of this site is smaller than a given threshold γ. The resulting values of N are allowed to differ between sites in order to adapt to the context of each site of the set of sequences that need to be compared. The parameter γ has an intuitive interpretation since it reflects the average number of repetitions in the sequences.

The necessity to adapt the size of the context considered around a site of a set of sequences to compare has also motivated the development of the *variable length local decoding* which generalizes the N-local decoding by considering not N-mers but words of various lengths to code and decode the set of sequences to be compared. This approach is briefly presented in Section 3.3.

3.2 MS4

The N-local decoding is used in order to produce partitions of the set of all sites in the sequences under study [15]. The MS4 approach relies on an object that describes the embedding of the successive results of the N-local decoding as N increases. The tree structure of this object is essential, since it can easily be parsed to select "relevant" (according to a certain criterion) classes of sites, which may occur at several values of N.

3.2.1 The partition tree

A recurring problem of N-mers-based methods is the lack of an objective criterion to tune the parameter N to a suitable value in order to compare a set of sequences. There is actually no reason to believe that a single "optimal" value of N will always be meaningful since a given set of sequences can contain parts that are very well conserved between sequences while others may vary a lot among them.

The MS4 approach combines different N-local decoding equivalence classes for various values of N by using an original construction, the *partition tree*, which allows

us to choose a set of "relevant" N-local decoding-classes. Let $\mathcal{E}^N$ be the partition of Σ, the site space of a set of sequences S, induced by the equivalence relation $\sim_N$.

Lemma 1 *For all $N \geq 0$, the partition $\mathcal{E}^N$ is coarser than $\mathcal{E}^{N+1}$ (i.e., each class of $\mathcal{E}^{N+1}$ is included in a class of $\mathcal{E}^N$).*

This lemma (see [9] for proof) is crucial, and corresponds to the intuitive idea that it is harder to group together large words than small ones. We are now ready to define the partition tree.

Definition 1 Let us set $\mathcal{E}^0 = \{\Sigma\}$ and $V = \cup_{i \geq 0} \mathcal{E}^i$ (i.e., V contains all the equivalence classes of all the partitions $\mathcal{E}^i$ for $i \geq 0$). The partition tree $\mathbf{P} = (V, E^{\mathbf{P}})$ is the tree where the vertices are the equivalence classes of V and the set of edges $E^{\mathbf{P}}$ is defined by

$$E^{\mathbf{P}} = \{(u, v) \in \mathcal{E}^N \times \mathcal{E}^{N+1} \mid v \subset u\}.$$

In other words, the vertices of $\mathbf{P}$ are the equivalence classes associated to the relations $\sim_N$ for all values of N. The edges are drawn between pairs of classes that correspond to successive values of N and such that one is a subset of the other. By Lemma 1, any two sites that are $(N+1)$-equivalent are also N-equivalent. On the other hand, two sites that are N-equivalent are not necessarily $(N+1)$-equivalent. In other words, the N-classes split as N increases. The edges are drawn precisely between any N-class C and all the $(N+1)$-classes into which C splits. Since any vertex of $\mathbf{P}$ has at most one ancestor by construction, $\mathbf{P}$ is a tree.

3.2.2 Classes selection

When we examine N-equivalence classes for all possible N, we face a deluge of information, moreover altogether redundant. We shall now use the partition tree to alleviate this problem. Given any set C of sites, let us define $|C|$, the *size* of C as the number of sites in C and the *spread* of C as the number of sequences which contain at least one element of C. We shall consider the quantity $\kappa(C)$ defined as the ratio between the size and the spread of C:

$$\kappa(C) = \frac{|C|}{|\{s \in S \mid \exists p, (s, p) \in C\}|} \geq 1.$$

For a given value $\gamma \geq 1$, the condition $\kappa(C) \leq \gamma$ means that the average number of occurrences of class C per sequence where it does occur is less or equal than γ. In particular, $\kappa(C) = 1$ means that no sequence contains more than one element of C (of course we take here C to be an N-local decoding-class). We call the parameter γ the *maximum average repetitivity*. We use this parameter to select nodes in the partition tree that satisfy $\kappa(C) \leq \gamma$.

This condition is not sufficient to make these classes relevant. Indeed, the bottom of the partition tree is occupied by classes corresponding to large N, which occur in only one sequence. Such classes are of no interest. In order to find relevant classes,

we have to "climb upward" (towards smaller values of N). Since any vertex of a tree has only one ancestor, the following definition does make sense.

Definition 2 An N-local decoding-class C is γ-relevant if it satisfies $\kappa(C) \leq \gamma$ while its ancestor does not.

The MS4 method selects all and only the relevant classes in a set of sequences (and ignores all the others).

3.2.3 The dissimilarity matrix

At the end of the MS4 procedure, each sequence can be rewritten, by replacing the letter originally found at a given site by the identifier of the relevant MS4-class to which the site belongs. We use the number of MS4 classes shared by two sequences to define a similarity index in a similar way as described in [14]. This measure is closely related to the percentage of identity classically used for sequence comparison.

Given any two sequences s_i and s_j, we compute their dissimilarity level d_{ij} as follows. For a class C, let $n_i(C)$ be the number of occurrences of C in s_i. Denote by C_{ij} the set of relevant classes that have representatives both in s_i and s_j. Since these two sequences may contain a different number of occurrences, we put $n_{ij} = \Sigma_{C \in C_{ij}} \min\{n_i(C), n_j(C)\}$. We define the dissimilarity level d_{ij} by

$$d_{ij} = 1 - \frac{n_{ij}}{\min\{\ell(s_i), \ell(s_j)\}},$$

where $\ell(s_i)$ and $\ell(s_j)$ are the lengths of s_i and s_j, respectively.

When $\gamma = 1$, n_{ij} is simply the number of relevant classes having representatives in both s_i and s_j. The dissimilarity matrix $(d_{ij})_{1 \leq i,j \leq S}$ can be given as input to a phylogenetic reconstruction software [23, 32].

3.3 Variable length local decoding of sequences

The variable length local decoding of sequences extends the local decoding of a given (and fixed) order presented above in the same way as variable length Markov models extend Markov models of a given order in the sense that the size of the window used to "decode" a position depends on the symbols in its neighbourhood in a way similar to its "local order" (i.e., the memory size at this position) under a variable length Markov model [38].

A variable length decoding scheme is defined from a prefix code $\mathcal{P}$. Let us first recall that a *prefix code* is a set $\mathcal{P}$ of words on a given alphabet which is such that no word in $\mathcal{P}$ is prefix of another word of $\mathcal{P}$ but itself. For instance, $\mathcal{P} = \{A, CA, CCA, CCC\}$ is a prefix code over the alphabet $\{A, C\}$. By construction, the words of a prefix code $\mathcal{P}$ are the leaves of the prefix tree storing the words of $\mathcal{P}$

(i.e., the tree where the nodes are the prefixes of words of $\mathcal{P}$ and where the direct ancestor of a node is obtained by discarding its last letter). A tree is a convenient representation of a prefix code (**Fig. 4**).

Note that from the property defining a prefix code, there is at most one word of a prefix code which occurs at a given position of a sequence.

We say that a prefix code $\mathcal{P}$ is *compliant* w.r.t. a given sequence (or a set of sequences[3]) if whatever the position of the sequence one picks, there is a word of $\mathcal{P}$ occurring at this position. It follows that if a prefix code $\mathcal{P}$ is compliant w.r.t. a sequence, there is one and only one word of $\mathcal{P}$ occurring at all positions of the sequence, except possibly at its last positions for which the corresponding words of $\mathcal{P}$ may be truncated.

The (variable length) coding of a sequence w.r.t. a given $\mathcal{P}$ where each word is associated to a unique identifier, is the sequence of identifiers of the overlapping words of $\mathcal{P}$ occurring along the sequence (**Fig. 4**). Given the coding of a sequence and the sequence of the corresponding lengths of the words of the prefix code (e.g., the two last rows of the table at the bottom-left of **Fig. 4**), there exists an antecedent which is maximal in the sense that (i) it has the greatest alphabet possible among the antecedents obtained with prefix codes with the same sequence of lengths of words and (ii) all the other antecedents can be obtained from it by letter-to-letter applications.

Note that the set of all N-mers over a given alphabet is a compliant prefix code w.r.t. any sequence over this alphabet. The maximal antecedent obtained from the prefix code made of all N-mers is exactly the N-local decoding presented above and the variable length decoding does generalize the (standard) local decoding.

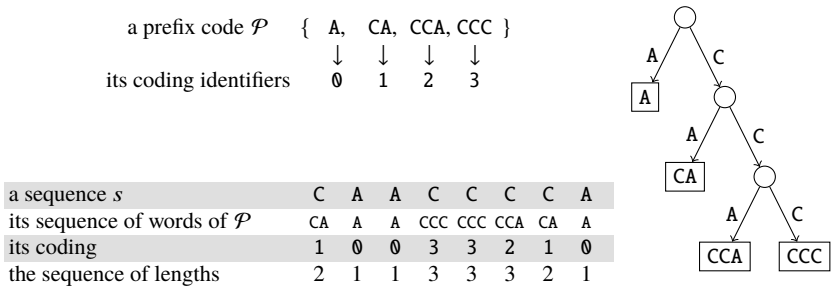
The variable length local decoding of the sequence may be equivalently defined by considering the equivalence relation between the positions of the sequences defined as the transitive closure of relation connecting two positions if there is a same word from the prefix code covering them with the same offset. An important point here is that there is an algorithm performing the variable length local decoding of a sequence which is linear with the size of the sequence both in time and memory space. In other words determining the variable length local decoding is not more expensive than dealing with the N-mers or the N-local decoding from a computational point of view.

Determining a (somewhat) relevant prefix code in order to perform an alignment free comparison of a given set of sequences is not obvious. In order to perform this task, we first remark that pruning the suffix tree of a sequence [43] leads to a compliant prefix code (reciprocally, the useful part of any compliant prefix code of a sequence may be obtained by this way). In [13], the suffix tree is pruned at the shallowest nodes corresponding to words having a probability smaller than a given threshold t to appear more than once in the whole set of sequence, under a Markov model of order 1 estimated on the set of sequences (in the current implementation, this probability is approximated from a binomial distribution). The probability threshold

[3] All the statements of this section still hold by replacing “sequence” with “set of sequences”

t is determined according to a heuristic criterion involving the number occurrences of the words of the prefix code in the set of sequences.

Finally, a dissimilarity matrix is computed from the variable length local decoding of the set of sequences to be compared in a way similar to that described in Section 3.2.3. This dissimilarity matrix can then be given as input to a phylogenetic reconstruction software.



a sequence s	C	A	A	C	C	C	C	A
its sequence of words of $\mathcal{P}$	CA	A	A	CCC	CCC	CCA	CA	A
its coding	1	0	0	3	3	2	1	0
the sequence of lengths	2	1	1	3	3	3	2	1

Fig. 4 Top-left: a prefix code $\mathcal{P}$ and the corresponding coding identifiers. Right: the tree representation of $\mathcal{P}$. Bottom-left: a sequence, its coding w.r.t. $\mathcal{P}$ and the corresponding sequence of words lengths.

4 Results: a brief look at the history of DNA topoisomerases IA

The necessity for a mechanism that would change the DNA topological state seemed obvious as soon as the structure of DNA was deciphered in 1953. The first DNA topoisomerase to be identified was isolated from a bacterium in 1971 and its sequence established in 1986 [44]. Any living organism possesses the two types of DNA-topoisomerases, at least one of each type, and several subtypes can often be found in the same species.

Among the different types, topoisomerases IA are the only ones to be present in all the living organisms. They are obviously proteins that have existed for a very long time (see **Fig. 1**).

4.1 The evolutionary history of topoisomerases IA

How can we decipher the evolutionary history of topoisomerases IA and trace the ancient duplications and horizontal transfers that led to the current state of affairs?

One first problem comes from their presence in every living organism. In late 2020, nearly 100 000 different organisms were represented in the sequence databanks (**Tbl. 1**). This is too much; it would be practically impossible to make an

exhaustive analysis of such an enormous set. In addition, the information is not always relevant, for example:

- numerous strains or individual of some model species were sequenced plenty of times. While important for understanding the intra-species polymorphism, this information is not relevant in terms of phylogeny.
- a comparison of man and chimpanzee is useless if one is interested in events that occurred way before the dinosaur era. One had better select species that diverged a long time ago, thus being far apart in the tree of life.

Prokarya	
Archaea	1 337
Bacteria	63 237
Eukarya	
Protozoa	573
Fungi	13 970
Plant	5 684
Vertebrate mammalian	1 294
Vertebrate other	4 544
Invertebrate	4 321
Total	93 209

Table 1 Number of species represented in RefSeq release 203 as of 9 november 2020

For the following study we selected 2 651 sequences: 2 135 sequences of DNA topoisomerases IA from bacteria, 268 from archaea and 68 from eukaria. The dissimilarity index between two sequences is measured with the *Variable length local decoding* (*VLD*, see above).

There are 3 subtypes of topoisomerases IA: TopoI, TopoIII and reverse gyrase. In the reverse gyrase subtype, the protein contains a helicase domain in addition to the topoisomerase domain.

- the TopoI subtype is present only among bacteria,
- the TopoIII subtype is present among bacteria, archaea and eukaryotes,
- the reverse gyrase subtype is present in hyperthermophilic bacteria and archaea[4].

Within the subtypes, one can see groups that reflect some physiological differences (for example, thermophile vs halophile). There are also some sequences that cannot be confidently linked with a given subtype. When submitted to automatic classification algorithms, the topoisomerases IA tend to cluster into 9 groups that are biologically coherent:

- with less than 9 groups, some markedly different proteins are clustered together and some characteristics are not highlighted,

[4] The reverse gyrase exists mainly in bacteria and archaea whose growth optimum is above 80 °C; it protects DNA from the denaturation that normally occurs at such high temperatures

- with more than 9 groups, the groups are no longer coherent.

The complete tree bears 2 651 leaves, its analysis would be beyond the scope of this article. We present here the tree of hyperthermophiles (**Fig 5**). The leaves correspond to the genera. The dissimilarity index of a genus is the average of the dissimilarity index of the species it groups. In order to simplify the reading, the identifier of a leaf is the name of the corresponding class (a class is a taxonomic unit grouping several genera, *e.g.* Mammalia is a class) with R1 / R2 if the reverse gyrase is duplicated. The number following the name of the class allows to go back to the sequences constituting the leaf. The names are in lower case for the reverse gyrases and in upper case for the other types of topoisomerases IA.

The tree of Archeae TopoIII is consistent with the taxonomy except perhaps for Thermoprotei 65 (*Pyrobaculum aerophilum* str. IM2 + *Pyrobaculum ferrireducens*).

The tree of Bacteria TopoI is consistent with the taxonomy except for Aquificae. Aquificae 75 and Aquificae 89 (*Desulfurobacterium thermolithotrophum* DSM 11699 and *Thermovibrio ammonificans* HB-1, respectively) are clearly disjoint from other Aquificae. Both groups of Aquificae have their own Thermodesulfobacteria (*Thermodesulfatator indicus* DSM 1528 – Aquificae 85 – in the first group and *Thermodesulfobacterium geofontis* OPF15 – Aquificae 86 – in the second).

The tree of **Fig. 5** contains in addition the TopoIII of vertebrates (labeled with a red star). It is very clearly related to the archaea TopoIII. This is true for all eukaryotes. The place of archaea in the evolutionary history of eukaryotes remains, however, an open question [29].

The tree of reverse gyrases is more complicated. In the Thermoprotei archaea, two types of reverse gyrases are clearly distinct and are separated from the bottom of the tree (noted R1 and R2, respectively). They correspond to a duplication of the reverse gyrase gene.

Species	R1 / R2 (identifiers in **Fig. 5**)
Aeropyrum pernix + A. camini	Thermoprotei 2 / Thermoprotei 3
Desulfurococcus amylolyticus + D. mucosus	Thermoprotei 6 / Thermoprotei 7
Hyperthermus butylicus	Thermoprotei 10 / Thermoprotei 11
Pyrolobus fumarii	Thermoprotei 22 / Thermoprotei 23
Saccharolobus solfataricus	Thermoprotei 24 / Thermoprotei 25
Sulfolobus islandicus	Thermoprotei 28 / Thermoprotei 27
Sulfurisphaera tokodaii	Thermoprotei 29 / Thermoprotei 30

Table 2 Archeae of our sample possessing a duplication of the reverse gyrase gene with the corresponding identifier in **Fig. 5**

A similar but less marked dichotomy is observed in bacteria.

An analysis of the differences between the two types of reverse gyrases is presented in the following section.

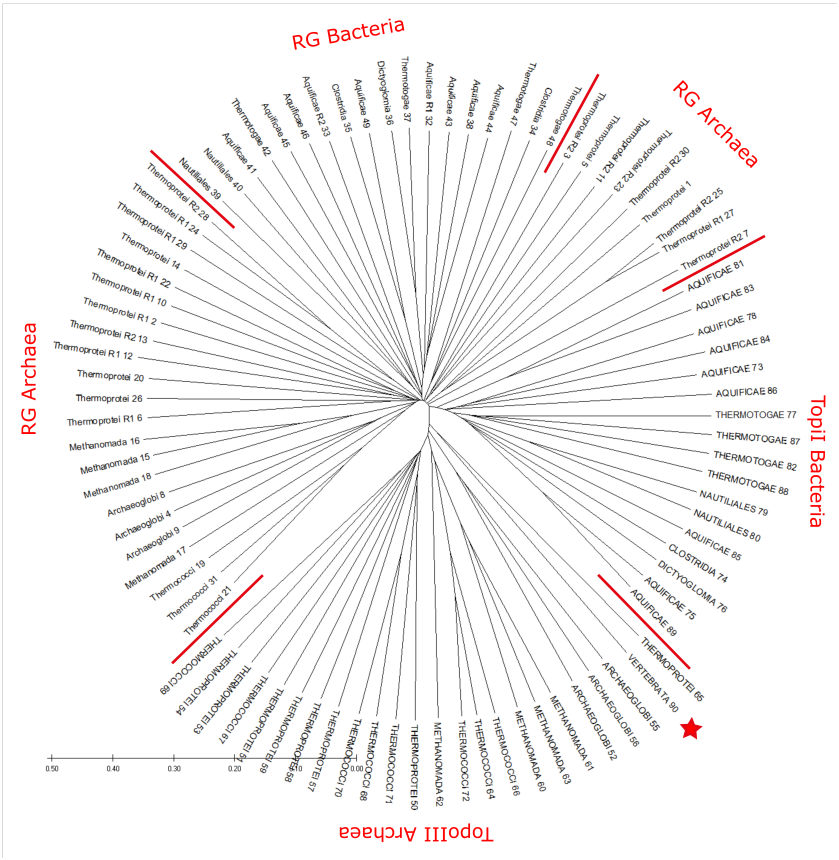


Fig. 5 Topoisomerases IA tree in hyperthermophiles. The leaves correspond to the genus. Identifier is the name of the corresponding class (a class is a taxonomic unit grouping several genera, *e.g.* Mammalia is a class) with R1 / R2 if the reverse gyrase is duplicated. The number following the name of the class allows to go back to the sequences constituting the leaf. The names are in lower case for the reverse gyrases and in upper case for the other types of topoisomerases IA. The tree also contains the TopoIII of vertebrates (labeled with a red star); the branch is clearly related to the archaea TopoIII.

4.2 The subfunctionalization of reverse gyrases

The reverse gyrase is the only protein (hence the only gene) to be quasi specific to hyperthermophilic organisms. It is systematically present among them and almost wholly absent in mesophilic cells. The reverse gyrase gene results from the fusion of a topoisomerase gene with a helicase gene [19]. It is possible that hyperthermophilic organisms may have existed before the advent of reverse gyrases, but the selective advantage provided by this gene is such that it must have been incorporated very

quickly in the genomes of all the hyperthermophilic organisms, bacteria as well as archaea [5]. A similar phenomenon can be observed nowadays with antibiotic resistance. The genes providing this resistance were rarely present among the bacterial populations – they were definitely not necessary – but with the current massive use of antibiotics those bacteria that possess the genes have now an obvious, tremendous selective advantage. As a result, the resistance genes are now quite common within pathogenic bacteria. They have been gained through horizontal transfer (HGT).

The revere gyrase gene is duplicated in several organisms, notably *Sulfolobus*. It has been shown that the two copies present some functional differences in *Sulfolobus* [19, 20]. The biologist now needs to identify the positions in the proteins which distingish the two copies and are responsible for those differences. He must first identify the potentially interesting sites, as experiments (in the "wet lab") are long and costly.

The approach presented below provides an answer, since it establishes a list of words (in the sense of Figure 3) that are characteristic of a given group of sequences.

Fig. 6 gives an example of the results obtained by studying 35 reverse gyrases that are representative of the biodiversity of hyper-thermophilic organisms, comprising 8 pairs of duplicated genes *topR1* and *topR2*. Among all the classes (in the sense of **Fig. 3**), we looked for those that were lacking in *topR1* but present in *topR2*, and vice versa. **Fig. 6** – drawn with the WebLogo software [10] – provides two examples of classes that distinguish *topR1* and *topR2*. The sequence conservation at a particular position in the alignment is defined as the difference between the maximum possible entropy and the entropy of the observed symbol distribution:

$$R_{seq} = S_{max} - S_{obs} = \log_2 m - \left(-m \sum_{n=1}^{m} p_n \log_2 p_n \right)$$

(with m=20 amino-acids). The maximum sequence conservation per site is 4.32 bits. Amino acids are given colors according to their chemical properties: polar amino acids (G, S, T, Y, C) show as green, hydrophobic (A, V, L, I, P, W, F, M) yellow, basic (K, R, H) blue, acidic (D, E) red and their amide (N, Q) purple [10].

Class N143_11 is observed in 19 reverse gyrases (among the 35 of the set) including the 8 *topR1* and class T23_8 in 16 reverse gyrases of the same set including the 8 *topR2*. Class A26_6 is present in 17 reverse gyrases including the 8 *topR1* and class G156_8 in 15 reverse gyrases including the 8 *topR2* (3 reverse gyrases do not fall into either category). The discriminating amino acids are at the center of the motifs. Some positions in T23_8 are strictly conserved -VESP on the left of the central T and KA on its right- while the other positions are more versatile.

Of course, the fact that classes N143_11 and T23_8 discriminate *topR1* and *topR2* does not prove that these sequences are responsible for the functional differences between the two genes. It is, however, an observation of interest to the biologist as it could give him a clue on where to start searching. The results given by the computer do not bring any proof but they enable to optimize the experimental work, which is important since experiments are long, extensive and expensive (the actual experiments had not yet been completed when this article was written).

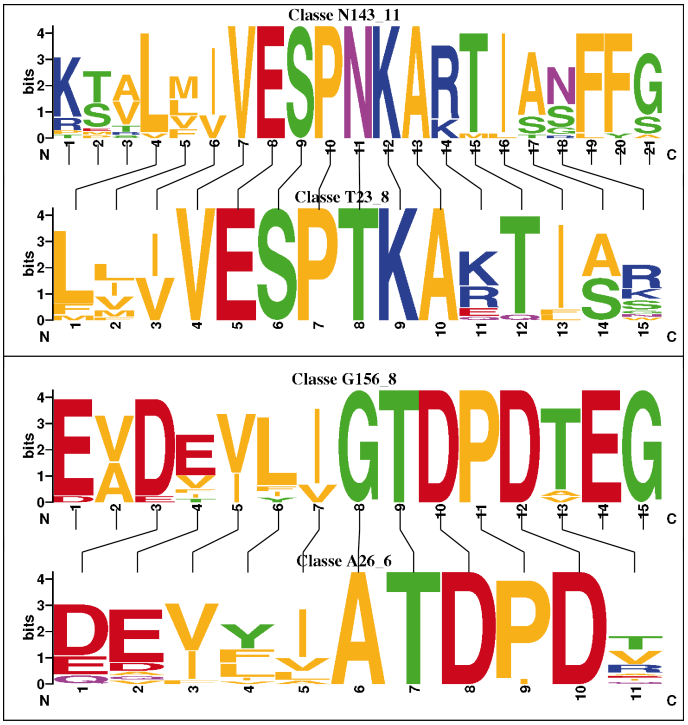


Fig. 6 Examples of classes (in the sense of **Fig. 3**) that discriminate reverse gyrases *topR1* and *topR2*. Top: alignment of classes N143_11 and T23_8. Bottom: alignment of classes G156_8 and A23_6. For example, N143_11 is one of the classes within the 35 gyrases given by the *N-local decoding* with N=11. With an asparagine N at its center, it is 21 amino acids long. The sequence conservation at a particular position in the alignment is defined as the difference between the maximum possible entropy and the entropy of the observed symbol distribution $R_{seq} = S_{max} - S_{obs} = \log_2 m - (-\sum_{n=1}^{m} p_n \log_2 p_n)$ (with m=20 amino-acids). The maximum sequence conservation per site is 4.32 bits. Amino acids have colors according to their chemical properties: polar amino acids (G, S, T, Y, C) show as green, hydrophobic (A, V, L, I, P, W, F, M) yellow, basic (K, R, H) blue, acidic (D, E) red and their amide (N, Q) purple [10].

5 From molecular phylogenies to the tree of life

Several insights can be gained from the topoisomerases phylogeny, without necessarily being able to establish the tree of life: the TopoI subtype is specific to the bacterial world, the TopoIII subtype enables to distinguish the bacteria from the archaea and from the eukaryotes while the reverse gyrases group together all the hyperthermophilic species. This is a common feature of molecular phylogenies being due on the one hand to the duplication of genes in the course of evolution and on the other hand, within bacteria and archaea, to the transfer of genes between widely divergent species -the so-called horizontal gene transfer or HTG. It is estimated

that 97% of the genes in bacteria and archaea have been the subject of horizontal transfers [46].

These horizontal transfers, however, seem to be randomly distributed. No obvious species are either donors or receptors. In other words, the HTGs blur the image we have of the tree of life, but without introducing any systematic bias [36]. As a result, the topologies of the phylogenetic trees are generally convergent. The evolutionary histories of the genes that are present in (almost) all the bacteria and archaea, as deduced from their phylogenetic trees, are coherent. This is also the case for the phylogeny of the DNA-topoisomerases IA. The link created by the reverse gyrases between the bacteria and the hyperthermophilic archaea does not call into question the validity of the bacterial and archaeal branches. This link is observed exclusively in the phylogeny of the reverse gyrases which are "modern" enzymes resulting from the fusion of two pre-existing genes, a DNA topoisomerase and a helicase. By contrast, the separation of bacteria and archaea into two different branches, which is observed in the phylogenies of the TopoI and TopoIII isomerases, is also found in most of the molecular phylogenies [46].

The relative position of the branches that are situated near the root of the tree, however, is controversial [17, 30]. Considering that those events occurred four billion years ago, this is not surprising. It is possible that, at that time, the genetic material might have been RNA and not DNA (this is still the case for many viruses) [28]. Interestingly, most of the DNA-topoisomerases IA possess an RNA-topoisomerase activity which appears important for untangling long RNA that forms pseudoknots. It has been hypothesized that this RNA-topoisomerase activity could be crucial in the RNA world, suggesting that the type IA is one of the most ancient enzymes [1, 20, 19].

Why four billion years? Molecular phylogenies show how the various evolutionary events are linked together, but provide no clue as to the date they occurred. The chronology is given by other disciplines. Conventional fossils trace the history of animals over a period of ca. 600 million years. Microfossils, stromatolites, remains of lipids and isotopic ratios[5] provide information on microorganisms and biogeochemical cycles in the Proterozoic oceans (2 500-540 My). They can be roughly interpreted in terms of extant organisms and metabolic processes. Archean rocks (more than 2 500 My) provide proof of the presence of life as far as 3 500 My ago, maybe even more. The phylogenetic and functional details are, however, quite limited [30, 26] (see **Fig. 7**).

6 Conclusions

Mathematics are now at the heart of biology. They are absolutely necessary to extract relevant information from the gigantic mass of data coming from the sequencing of

[5] As an example, let us take the bias in the isotopic composition of carbon. Atmospheric CO_2 is made up of a mixture of ^{12}C and ^{13}C. Since the photosynthetic organisms have a preference for the ^{12}C-containing CO_2, the biological fossil sediments will be richer in ^{12}C than the abiotic sediments. This corresponds to the ^{13}C *depleted reduced carbon* in **Fig. 7**.

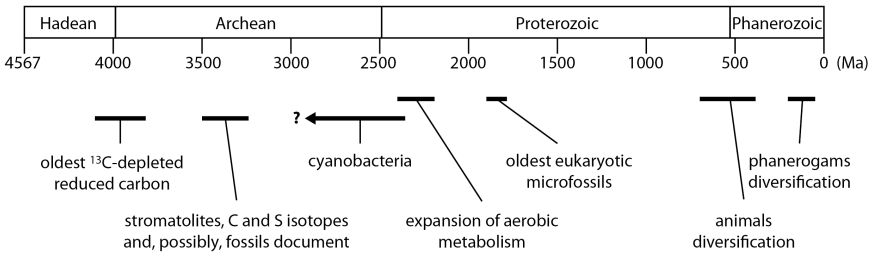


Fig. 7 Time table for Earth's early history (Figure taken from [26])

numerous genomes and other high throughput techniques. In 2020, 186 million protein sequences, belonging to 105 000 organisms (including viruses) had been determined.

Alex had anticipated this evolution and had become interested in the analysis of biological sequences as early as the 90's. However, while the general tendency in biology is to develop tools and their *ad hoc* tweaks, Alex systematically looked for non-trivial but simple solutions. Which lead him to constantly ask the question "What are the fundamental principles?". Probably a legacy of his career in theoretical physics.

7 Acknowledgments

The authors would like to sincerely thank Marc Nadal and Jean-Loup Risler for their constructive criticism and Alessandra Riva for proofreading the article.

References

1. M. Ahmad, Y. Xue, S. Lee, J. Martindale, W. Shen, W. Li, S. Zou, M. Ciaramella, H. Debat, M. Nadal, F. Leng, H. Zhang, Q. Wang, G. Siaw, H. Niu, Y. Pommier, M. Gorospe, T.-S. Hsieh, Y.-C. Tse-Dinh, and W. Wang. RNA topoisomerase is prevalent in all domains of life and associates with polyribosomes in animals. *Nucleic Acids Res.*, 44:gkw508, 06 2016.
2. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403 – 410, 1990.
3. P. Bawono, M. Dijkstra, W. Pirovano, K. A. Feenstra, and S. Abeln. *Multiple Sequence Alignment*, volume 1525, pages 167–189. Humana Press Inc, 11 2017.
4. A. H. Bizard and I. D. Hickson. The many lives of type IA topoisomerases. *J. Biol. Chem.*, 295(20):7138–7153, 2020.
5. R. J. Catchpole and P. Forterre. The evolution of reverse gyrase suggests a nonhyperthermophilic last universal common ancestor. *Mol Biol Evol*, 36(12):2737–2747, Dec. 2019.
6. M. Chatzou, C. Magis, J.-M. Chang, C. Kemena, G. Bussotti, I. Erb, and C. Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, 11 2015.

7. J. Chen and S. W. I. Siu. Machine learning approaches for quality assessment of protein structures. *Biomolecules*, 10(4):626, Apr. 2020.
8. E. Corel, R. Fegalhi, F. Gérardin, M. Hoebeke, M. Nadal, A. Grossmann, and C. Landès-Devauchelle. Local similarities and clustering of biological sequences: New insights from N-local decoding. *The First International Symposium on Optimization and Systems Biology*, 01 2007.
9. E. Corel, F. Pitschi, I. Laprevotte, G. Grasseau, G. Didier, and C. Landès-Devauchelle. MS4 - multi-scale selector of sequence signatures: An alignment-free method for classification of biological sequences. *BMC Bioinf*, 11:406, 07 2010.
10. G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. Weblogo: A sequence logo generator. *Genome Res.*, 14(6):1188–1190, 2004.
11. C. Devauchelle, Y. Diaz, G. Didier, A. Hénaut, and B. Torrésani. Pseudo-rate matrices, beyond Dayhoff's model. This volume, 2021.
12. G. Didier. Caractérisation des n-écritures et application à l'étude des suites de complexité ultimement n+ cste. *Theoretical computer science*, 215(1-2):31–49, 1999.
13. G. Didier, E. Corel, I. Laprevotte, A. Grossmann, and C. Landès-Devauchelle. Variable length local decoding and alignment-free sequence comparison. *Theoretical Computer Science*, 462:1 – 11, 2012.
14. G. Didier, L. Debomy, M. Pupin, M. Zhang, A. Grossmann, C. Devauchelle, and I. Laprevotte. Comparing sequences without using alignments: application to HIV/SIV subtyping. *BMC Bioinf.*, 8(1):1, Jan. 2007.
15. G. Didier, I. Laprevotte, M. Pupin, and A. Hénaut. Local decoding of sequences and alignment-free comparison. *Journal of computational biology : a journal of computational molecular cell biology*, 13:1465–76, 11 2006.
16. T. Farkaš, J. Sitarčík, B. Brejová, and M. Lucká. SWSPM: A novel alignment-free DNA comparison method based on signal processing approaches. *Evolutionary bioinformatics online*, 15:1176934319849071, 2019.
17. P. Forterre. The universal tree of life: an update. *Front. Microbiol.*, 6:717, 2015.
18. P. Forterre and D. Gadelle. Phylogenomics of DNA topoisomerases: Their origin and putative roles in the emergence of modern organisms. *Nucleic Acids Res*, 37:679–92, 03 2009.
19. F. Garnier, M. Couturier, H. Débat, and M. Nadal. Archaea: a gold mine for topoisomerase diversity. *Front. Microbiol.*, 2021. In press.
20. F. Garnier, H. Debat, and M. Nadal. Type IA DNA topoisomerases: A universal core and multiple activities. In M. Drolet, editor, *DNA Topoisomerases*, volume 1703 of *Methods in Molecular Biology*, chapter 1, page 1:20. Springer, 2018.
21. G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577, 07 1999.
22. D. M. Hillis, C. Moritz, and B. K. Mable, editors. *Molecular Systematics*. Sinauer Associates Inc., 1996.
23. D. Huson. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23(2):254–267, 01 2006.
24. L. Jaroszewski. Protein structure prediction based on sequence similarity. *Methods in molecular biology (Clifton, N.J.)*, 569:129–56, 02 2009.
25. W. Just. Computational complexity of multiple sequence alignment with SP-Score. *J. Comput. Biol.*, 8(6):615–623, 2001. PMID: 11747615.
26. A. H. Knoll, K. D. Bergmann, and J. V. Strauss. Life: the first two billion years. *Philos Trans R Soc Lond B Biol Sci*, 371(1707):20150493, Nov. 2016.
27. C.-A. Leimeister, S. Sohrabi-Jahromi, and B. Morgenstern. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*, 33(7):971–979, 01 2017.
28. W. Ma. What does "the rna world" mean to "the origin of life"? *Life (Basel, Switzerland)*, 7(4):49, Nov. 2017.
29. F. MacLeod, G. S. Kindler, H. L. Wong, R. Chen, and B. P. Burns. Asgard archaea: Diversity, function, and evolutionary implications in a range of microbiomes. *AIMS microbiology*, 5(1):48–61, Jan. 2019.

30. W. F. Martin and F. L. Sousa. Early microbial evolution: The age of anaerobes. *Cold Spring Harbor Perspect. Biol.*, 8(2), 2016.
31. S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443 – 453, 1970.
32. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.*, 32(1):268–274, 11 2014.
33. W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85:2444–8, 05 1988.
34. E. L. Peterson, J. Kondev, J. A. Theriot, and R. Phillips. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics (Oxford, England)*, 25(11):1356–1362, June 2009.
35. V. Polyanovsky, A. Lifanov, N. Esipova, and V. Tumanyan. The ranging of amino acids substitution matrices of various types in accordance with the alignment accuracy criterion. *BMC Bioinf.*, 21(11):294, Sept. 2020.
36. P. Puigbò, Y. I. Wolf, and E. V. Koonin. Seeing the tree of life behind the phylogenetic forest. *BMC Biol.*, 11(1):46, Apr. 2013.
37. J. Risler, M. Delorme, H. Delacroix, and A. Henaut. Amino acid substitutions in structurally related proteins a pattern recognition approach: Determination of a new and efficient scoring matrix. *J. Mol. Biol.*, 204(4):1019 – 1029, 1988.
38. J. Rissanen. A universal data compression system. *IEEE Transactions on Information Theory*, 29(5):656–664, 1983.
39. T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415 – 431, 1986.
40. C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford lecture series in mathematics and its applications*. Oxford University Press, 2003.
41. T. Smith and M. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195 – 197, 1981.
42. D. Tagu and J.-L. Risler. *Bioinformatique ; Principes d'utilisation des outils*. Editions Quae, Paris, 2010.
43. E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, Sept. 1995.
44. J. C. Wang. DNA topoisomerases: why so many ? *J Biol Chem*, 266(11):6659–62, 1991.
45. J. D. Watson and F. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.
46. M. C. Weiss, M. Preiner, J. C. Xavier, V. Zimorski, and W. F. Martin. The last universal common ancestor between ancient earth chemistry and the onset of genetics. *PLos Genet*, 14(8):1–19, 08 2018.
47. A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. Röhling, J. J. Choi, M. S. Waterman, M. Comin, S.-H. Kim, S. Vinga, J. S. Almeida, C. X. Chan, B. T. James, F. Sun, B. Morgenstern, and W. M. Karlowski. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.*, 20(1):144, July 2019.
48. A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.*, 18(1):186, Oct. 2017.
49. E. Zuckerlandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97 – 166. Academic Press, 1965.