Pseudo-Rate Matrices, beyond Dayhoff's model

Claudine Landès, Yolande Diaz-Lazcoz, Alain Hénaut, Bruno Torrésani

Pseudo-Rate Matrices, beyond Dayhoff's model

Claudine Landès, Yolande Diaz-Lazcoz, Alain Hénaut and Bruno Torrésani

Abstract One of the fundamental techniques of biology is sequence alignment, namely transforming one sequence into another with minimal change. Sequence alignment is essential for evolutionary studies and is a source of information for the analysis of the physico-chemical mechanisms which are at the heart of protein activity. Biologists almost exclusively use methods based on a very simple model, although they are aware that this can be quite removed from reality. In fact, the more complex models involve so many variables that they cannot be calculated in practice. This paper presents a method to estimate the quality of the approximation made using simple models, giving a measure of the deviation from reality. It is exclusively based on the analysis of pairwise alignments, without resorting to multiple alignments, and therefore without requiring the construction of trees and the problems associated with it. The paper also describes an approach that allows building trees and clusters from sequences without strongly relying on the choice of a dissimilarity measure. It illustrates the interest and effectiveness of the point of view promoted by Alex: assume as little as possible and try to gather information from the data, before turning to explicit modeling if necessary.

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

Yolande Diaz-Lazcoz
Université d'Evry, LaMME, 91000 Evry, France. e-mail: yolande.diaz2@univ-evry.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

1 Introduction

Alignment-free sequence comparisons make it possible to reconstitute the phylogeny of proteins that have diverged greatly over time (see the introduction of the companion paper [2] in this volume for definitions of the main concepts), but they do not, on their own, enable to model the mechanisms of protein evolution. This requires alignments, that is, the transformation of one sequence into another whilst minimizing the number of changes.

The realization of the alignments has two clearly distinct parts: an alignment algorithm and a transition matrix (also called substitution scoring matrix or matrix of accepted mutation rates, see Definition 5 below). This matrix defines the rate at which amino acids are replaced by others over the course of evolution.

We presented the alignment algorithms in the article [2]. Here we present some of the issues related to the construction of amino acid substitution rate matrices, the related models of sequence evolution, and how Alex approached them. A main aspect of Alex's contributions is the will to adapt models to data, not vice versa. This paper first describes the estimation of observed rate matrices from pairwise sequence alignments, examines connections with popular sequence evolution models and shows how simple multivariate analysis techniques can be applied to these matrices to highlight and quantify departures from such models. It also accounts for an original approach to biological sequence clustering that avoids as much as possible ad-hoc dissimilarity measures that tend to bias results and thus interpretations. Theoretical developments are complemented by numerical results on real data that include topoisomerases already discussed in [2].

2 Classical approaches, scoring matrices

2.1 Dayhoff evolution model - The PAM (Point Accepted Mutation) Matrix

In the late 1960s, Margaret Dayhoff had the excellent idea of collating as many closely related homologous protein sequences as possible, aligning them "by hand" and counting the substitutions, which enabled her to estimate the probability that a given amino acid will be replaced by another when there is about 1% change between two sequences. She thus obtained a 20×20 matrix called PAM1 for "1 Point Accepted Mutation per 100 residues". By construction, the PAM matrices are symmetrical.[1]

Margaret Dayhoff hypothesized that the probability of replacing an amino acid at a given site depends only on the nature of that amino acid. This probability is

[1] In general, the PAM series rather refers to the scoring matrices that are used to weigh the replacements in protein alignment methods. These scoring matrices, calculated from these probabilities, and closely connected to transition matrices studied in section 3.2 below.

independent of what may happen at the other sites and of what may have happened previously at this same site [5]. In other words, repeated mutations over a longer period of evolution follow the same substitution pattern as those observed. We can therefore extrapolate from PAM1 the PAM matrices corresponding to any percentage of change.

Biologists typically use the PAM120 matrix (about 40 % identity). They switch to PAM250 if they find that the sequences are very divergent (about 20 % identity) (see **Fig. 1**).

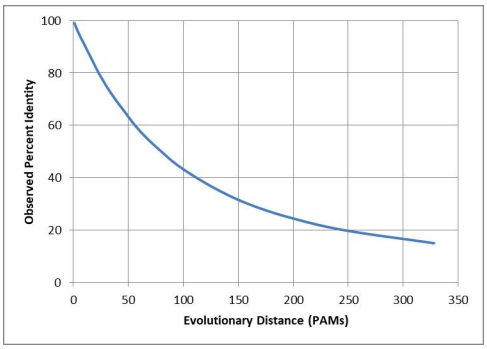

Fig. 1 Correspondence of the observed percent difference and the estimated evolutionary distance in PAM [5].

2.2 BLOSUM, another general purpose substitution matrix

Since then, other matrices based on the same principle have been developed. The most widely used are the *BLOcks of Amino Acid SUbstitution Matrix* (BLOSUM) matrix series. BLOSUM matrices were constructed from the count of substitutions observed in a series of multiple alignments listed in the Blocks database [22]. This database collated alignment blocks without gap within related proteins [2]. These well-conserved regions are believed to have greater functional relevance.

Created twenty years after the PAM matrices, the BLOSUM matrices obviously have the advantage of being constructed from a much higher number of alignments, themselves made from a much larger range of proteins. Biologists generally use the BLOSUM62 matrix (*i.e.* the matrix constructed by retaining in all the aligned pairs of the Blocks database those which present at most 62 % identity).

As BLOSUM matrices are based on structures which have been well conserved during evolution, they are less "lax", which means that amino acids are less easily exchangeable than in PAM matrices. The general consensus is that the BLOSUM matrices are superior in terms of sensitivity and specificity without, however, the difference with the PAM matrices being considerable (see Chapter 11 in [27]).

2.3 Available biological material for the estimation of scoring matrices

Margaret Dayhoff had very little data when she calculated the first PAM matrices (in 1969, 814 substitutions in all were known in pairs of sequences having more than 85% identity). Progress has been very rapid, the matrices were based on 60 000 substitutions by the early 1990s. The number of known sequences has exploded since. Biologists now have a choice of dozens of substitution scoring matrices (see [28] and [23]). They differ among other things by:

- the sequences used in the training set (it can contain a single family of proteins or several hundred);
- some authors use the mutations observed in a global alignment, including both highly conserved regions as well as highly mutable regions (*e.g.* PAM), while others only take into account regions whose structure is well conserved (*e.g.* BLOSUM).

On the other hand, *all authors implicitly assume that the matrices of the substitution rates are homogeneous in the training set*. The latter assumption is not necessary. As we show below, one can simultaneously estimate the evolution rate matrix and a divergence age for each pair of sequences [8].

3 Rate Matrices, beyond Dayhoff's model

The construction of the Dayhoff matrices is very similar to approaches developed in the context of the inference of evolutionary trees that often rely on Markov models on trees (see [9], and Chapter 11 of [14], see also [20] for a different approach, which bears similarities with the techniques presented here).

However, Dayhoff's approach departs from these as it is mainly descriptive and does not involve explicit modeling and corresponding parameter estimation. It only exploits simple counting, converted into scores. The construction in [8], described below, builds on these ideas and attempts to interpret counting matrices in terms of a minimal number of parameters, namely a rate matrix and divergence times whenever possible, or several matrices in more complex cases.

3.1 Definitions and notations

We start by introducing background definitions and notations. Throughout this paper, we work with square matrices with real entries. We refer to [4] for an account of the main aspects of matrix calculus. We use the following notations: for any $m \times m$ matrices $\mathbf{M}, \mathbf{M}'$, their inner product is defined as $\langle \mathbf{M}, \mathbf{M}' \rangle = \sum_{i,j=1}^{m} M_{ij} M'_{ij}$, the corresponding norm is denoted by $\|\mathbf{M}\|_2$, and the trace of $\mathbf{M}$ is $\mathsf{Tr}(\mathbf{M}) = \sum_{i=1}^{m} M_{ii}$. The spectral norm $\|\mathbf{M}\|$ of a matrix $\mathbf{M}$ is its largest singular value.

We will make use of the matrix logarithm, which should be understood as the inverse function of the matrix exponential. While the latter can be defined by its power series which is always convergent, the matrix logarithm raises more difficult questions, and may be defined in various ways (see the note by H.E. Haber [13] for a summary). We will limit ourselves to the definition based on the Mercator series:

Definition 1 (Matrix logarithm)

The logarithm of a matrix $\mathbf{M}$ is defined by the infinite power series expansion

$$\log \mathbf{M} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (\mathbf{M} - \mathsf{I}_m)^k \ , \tag{1}$$

when the latter is convergent. Here, I_m denotes the $m \times m$ identity matrix.

Convergence is ensured whenever $\|\mathbf{M} - \mathsf{I}_m\| < 1$. The matrix logarithm does not satisfy all the usual properties of the numerical logarithm: in general, $\log(\mathbf{M}_1\mathbf{M}_2) \neq \log(\mathbf{M}_1) + \log(\mathbf{M}_2)$. However, the equality holds true when $\mathbf{M}_1$ and $\mathbf{M}_2$ commute, and the relation

$$\log(\mathbf{M}^\tau) = \tau \log(\mathbf{M}) \tag{2}$$

is preserved for all positive integer τ, and more generally when $\mathbf{M}^\tau$ is well defined.

Definition 2 (Transition matrices, rate matrices)

1. A *transition matrix* is an $m \times m$ matrix $\mathbf{P}$ such that for all i, $\sum_{j=1}^{m} P_{ij} = 1$.
2. A *pseudo rate matrix* is a $m \times m$ matrix $\mathbf{Q}$ such that $\sum_{j=1}^{m} Q_{ij} = 0$ for all i and $Q_{ii} \leq 0$ for all i.
3. A *rate matrix* is a pseudo rate matrix $\mathbf{Q}$ such that $Q_{ij} \geq 0$ for all $i \neq j$.

Transition matrices are sometimes called *stochastic matrices*, or *Markov transition matrices*. Transition matrices are naturally associated with finite state Markov chains, *i.e.* random processes such that the probability of moving from state i to state j in one time step is given by the matrix element P_{ij}. In general, the eigenvalues of a transition matrix are complex numbers of modulus smaller than or equal to 1.

Definition 3 (Markov semigroup)

A *Markov semigroup* is a family of transition matrices $t \in \mathbb{R}^+ \to \mathbf{P}(t)$ satisfying the *Chapman–Kolmogorov equation*

$$\mathbf{P}(t)\mathbf{P}(t') = \mathbf{P}(t + t') \ , \quad t, t' \in \mathbb{R}^+ \ , \tag{3}$$

and such that for all i, j, $P_{ij}(0) = \delta_{ij}$ and $\lim_{t\to 0} P_{ii}(0) = 1$.

Given a Markov semigroup, there always exists a matrix $\mathbf{Q} = \mathbf{P}'(0)$ such that $\mathbf{P}(t) = e^{t\mathbf{Q}}$ and $\mathbf{Q}$ is a rate matrix. Conversely, if $\mathbf{Q}$ is a rate matrix, then the exponentials $e^{t\mathbf{Q}}$, where $t \in \mathbb{R}^+$, form a Markov semigroup.

Definition 4 (Embeddable transition matrices)

1. A transition matrix $\mathbf{P}$, is *embeddable* into a Markov semigroup, or simply embeddable if there exists a corresponding Markov semigroup $t \to \mathbf{P}(t)$ such that $\mathbf{P} = \mathbf{P}(1)$ or, equivalently, if there exist a rate matrix $\mathbf{Q}$ such that $\mathbf{P} = e^{\mathbf{Q}}$. $\mathbf{Q}$ is called a generator.
2. A set of transition matrices $\{\mathbf{P}^{(1)}, \ldots \mathbf{P}^{(K)}\}$ is jointly embeddable if there exist a rate matrix $\mathbf{Q}$ and positive numbers $\tau_1, \ldots \tau_K$ such that for all k, $\mathbf{P}(k) = e^{\tau_k \mathbf{Q}}$

When $\mathbf{P}$ is embeddable, the corresponding rate matrix $\mathbf{Q}$ coincides with the matrix logarithm of the transition matrix $\mathbf{P}$.

These notions are at the heart of the maximum likelihood approach developed by Felsenstein [9] and followers for the estimation of evolutionary trees. Evolutionary trees model the evolution of a set of current time data, called taxa, from their most recent common ancestor. In the situation of interest here, taxa are protein sequences. The latter are represented as symbolic sequences, namely sequences $\{x(k),\ k = 1 \ldots K\}$ with values in a finite alphabet $\mathcal{A}$ of cardinality $m = \#\mathcal{A}$ (for protein sequences, $m = 20$, letters label amino-acids). A pairwise alignment of two sequences x, y is an ordered pair $(x, y) = \{(x(k), y(k)),\ k \in I(x, y)\}$ where $I(x, y)$ is a set of indices. Pairwise alignments are used to identify regions of similarity between two sequences of interest. Multiple alignments involve more than two sequences. Algorithms for aligning sequences have been described in the companion paper [2].

Many models have been proposed for describing a family of aligned sequences (a multiple alignment) from an evolutionary perspective (see [25] for a review). Among these the Markov Chain on a Tree (MCT) model has received considerable attention. The MCT model assumes that the sites of the sequences are independent, identically distributed, random variables (see *e.g.* [26] for a discussion of the consequences of such assumptions), whose time evolution is mainly described by two parameters:

- a binary tree, whose leaves are the sequences considered (present time), nodes are ancestor sequences, and branches represent Markovian evolution,
- a family of transition matrices associated with the branches of the tree that characterize evolution along the branch, modeled by a Markov chain.

At each node a sequence gives rise to two different sequences, each one evolving according to its own Markov chain. A couple of examples is provided in **Fig. 2**. A rooted tree with K leaves has $2K - 3$ edges ($2K - 2$ for unrooted trees), thus $2K - 3$ transition matrices. It is worth mentioning at this point that the number of different tree topologies is equal to $(2K - 3)!!$, which makes the tree identification problem extremely difficult for large numbers of sequences.

Among MCT models, the F81 model of [9] also assumes that the evolutionary process is stationary, homogeneous, and reversible. In the context of interest here, stationarity means that the probabilities of amino-acids are the same at all nodes of the tree; homogeneity essentially means that all transition matrices are embeddable (local homogeneity) or jointly embeddable (global homogeneity); reversibility means that for all pair (i, j) of amino-acids, substitutions $i \to j$ and $j \to i$ have equal probabilities.

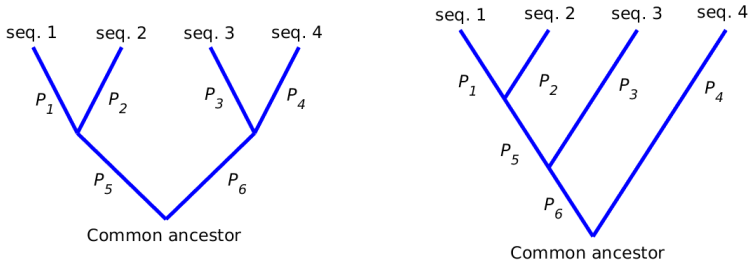

Fig. 2 Two examples of rooted evolutionary trees with 4 taxa, with different topologies. In the most general form of MCT models a transition matrix is associated with each branch.

The parameters of the model (transition matrices and tree) are sufficient to compute probabilities of all possible multiple alignments and evaluate them numerically. The associated estimation problem is: infer parameter values of the multiple alignment, which was done in [9] using a maximum likelihood approach.

Parameter estimation turns out to be computationally heavy for large families of sequences, even under the assumptions of homogeneity, stationarity and reversibility and additional simplifications are often made. In addition, the comparison of likelihoods for different tree topologies raises other difficult questions [1]. Finally, these assumptions are often violated by data, so that the inferred evolutionary trees have to be taken cautiously. An alternative approach avoiding these assumptions has been proposed in [3] and more recently [16], it attempts to estimate a transition matrix for each branch of the tree using again maximum likelihood. Still, the problem is extremely difficult when large sequence families are considered, and the statistical significance of results obtained with such a large number of parameters may be rather questionable.

3.2 From pairwise alignments to rate matrix

The approach developed in [8] departs from these general models and attempts to find simpler and more versatile descriptions for multiple alignments. The starting point is a multiple alignment of *sufficiently related* and *sufficiently close* protein sequences (to be introduced in Definitions 6 and 7 below). As stressed above, unlike most probabilistic approaches to phylogeny (see [9], and [14] and references therein), the approach of [8] does not use the full multiple alignment, and only focuses on pairwise alignments. From each pairwise alignment (x, y), an observed transition matrix $\mathbf{P}^{(x,y)}$ is computed (see Definition 5 below), and the question is: to which extend can the so-obtained family of observed transition matrices can be gathered (embedded in the sense of Definition 4) into a common framework. For that, the objects of interest will be the matrix logarithms $\mathbf{L}^{(x,y)}$ of the observed transition matrices $\mathbf{P}^{(x,y)}$.

3.2.1 Observed transition and rate matrices

From now on, we consider a set of sequences denoted by X, and a set $\{(x,y), x,y \in X\}$ of pairwise alignments.

Definition 5 (Counts, frequencies)

Given an ordered pairwise alignment (x,y) of length $c(x,y)$,

1. The matrix of transition frequencies $\mathbf{F}^{(x,y)}$ is defined by its elements

$$F_{ij}^{(x,y)} = \frac{1}{c(x,y)} \sharp\big\{k : x(k) = i \text{ and } y(k) = j\big\} , \quad i,j = 1,\dots m . \tag{4}$$

2. The vectors of frequencies $\pi^{(x)} = (\pi_1^{(x)}, \cdots \pi_m^{(x)})$ are given by

$$\pi_i^{(x)} = \frac{1}{c(x,y)} \sharp\big\{k : x(k) = i\big\} , \quad i = 1,\dots m . \tag{5}$$

We also denote by $\Pi^{(x)} = \mathsf{diag}(\pi^{(x)})$ the corresponding diagonal matrices.

From these quantities, observed transition and rate matrices can be introduced.

Definition 6 (Observed transition and rate matrices)

Given an ordered pairwise alignment (x,y) of length $c(x,y)$,

1. The associated *observed transition matrix* is defined as

$$\mathbf{P}^{(x,y)} = \Pi^{(x)^{-1}} \mathbf{F}^{(x,y)} , \qquad P_{ij}^{(x,y)} = \frac{F_{ij}^{(x,y)}}{\pi_i^{(x)}} . \tag{6}$$

2. The sequences x, y are *sufficiently related* if the corresponding observed transition matrices admit a logarithm in the sense of Definition 1. In this case, the matrices

$$\mathbf{L}^{(x,y)} = \log \mathbf{P}^{(x,y)} \tag{7}$$

are called *observed rate matrices*. The diagonals of these matrices are called *mutabilities*, and denoted by

$$\mu^{(x,y)} = \mathsf{diag}\left(\mathbf{L}^{(x,y)}\right) . \tag{8}$$

Algorithms for reconstructing evolutionary trees from multiple alignments can be based upon "evolutionary distances". Such distances can be constructed from the above data. For example, the *LogDet distance* proposed in [3]

$$\mathsf{ldet}(x,y) = \log(\det(\mathbf{P}^{(x,y)})) , \tag{9}$$

is a natural choice in situations where the observed transition matrices are jointly embeddable, *i.e.* of the form $\mathbf{P}^{(x,y)} = e^{\tau(x,y)\mathbf{Q}}$ for some rate matrix $\mathbf{Q}$ (called

generator). In such a case $\mathsf{ldet}(x,y) = \mathsf{Tr}(\mathbf{L}^{(x,y)}) = \tau(x,y)\mathsf{Tr}(\mathbf{Q})$ is proportional to the divergence time $\tau(x,y)$. Although termed "distance", ldet is not a metric, in particular is not symmetric, and is therefore not a suitable quantity for most tree reconstruction methods that require tree metrics[2]. Nevertheless, ldet may be an interesting quantity to look at, precisely because it doesn't force symmetry. We will analyze a biologically relevant example in section 4.1 below.

Remark 1 (Symmetric LogDet distances)

Several alternative *LogDet* distances have also been proposed and studied in the literature. Among these, the quantity $d(x,y) = -\log(\det(F^{(x,y)}))$ proposed in [18], where it was shown that this distance allows identification of the tree topology, but not edge lengths. Another alternative is $\delta(x,y) = \frac{1}{2}\log(\det(\mathbf{P}^{(x,y)}\mathbf{P}^{(y,x)}))$ which possesses the desired symmetry property and interesting interpretations in the context of reversible MCT models [25].

3.2.2 The symmetrized case

In the above setting a pairwise alignment (x,y) gives rise to two observed rate matrices $\mathbf{L}^{(x,y)}$ and $\mathbf{L}^{(y,x)}$, which complicates the analysis (although a strong discrepancy between these two would indicate a strong departure from the above model). Simplification can be achieved by averaging these two matrices, however it appears more natural to introduce symmetrization directly in the counting procedure. With the above notations, we introduce the symmetrized matrices $\widetilde{\mathbf{F}}^{(x,y)}$ and vectors $\tilde{\pi}^{(x,y)}$

$$\widetilde{\mathbf{F}}^{(x,y)} = \frac{1}{2}\left(\mathbf{F}^{(x,y)} + \mathbf{F}^{(y,x)}\right) , \qquad \tilde{\pi}^{(x,y)} = \frac{1}{2}\left(\pi^{(x,y)} + \pi^{(y,x)}\right) , \tag{10}$$

and define as before the diagonal matrix $\widetilde{\Pi}^{(x,y)} = \mathsf{diag}(\tilde{\pi}^{(x,y)})$.

Definition 7 (Sufficiently close sequences)

Two sequences (x,y) are *sufficiently close* when the corresponding matrix $\widetilde{\mathbf{F}}^{(x,y)}$ is positive definite.

As stated in [8], for sufficiently close sequences (x,y), the matrix $\widetilde{\Pi}^{(x,y)}$ is nonsingular. This motivates the introduction of corresponding transition and rate matrices:

Proposition 1 *Let (x,y) be a pairwise alignment of sufficiently close sequences. Then the following symmetrized observed transition and rate matrices are well defined*

$$\widetilde{\mathbf{P}}^{(x,y)} = (\widetilde{\Pi}^{(x,y)})^{-1}\widetilde{\mathbf{F}}^{(x,y)} , \qquad \widetilde{\mathbf{L}}^{(x,y)} = \log \widetilde{\mathbf{P}}^{(x,y)} . \tag{11}$$

Assuming, for argument's sake, that alignments were generated according to the Markov tree model of [9], the following observations can be made:

[2] A tree metric is a map $(x,y) \to \delta(x,y)$ that satisfies the requirements of a dissimilarity map (it is non-negative, symmetric, and such that $\delta(x,x) = 0$ for all x) and an additional condition called the four points condition, see *e.g.* chapter 11 in [14].

- Observed transition matrices can be expected to be powers $\mathbf{P}^{\tau(x,y)}$ of a unique transition matrix $\mathbf{P}$ (up to statistical fluctuations).
- Therefore observed rate matrices can be expected to be (up to fluctuations) proportional to a unique rate matrix $\mathbf{Q} = \log(\mathbf{P})$.

In such a situation, simple tools such as linear regression may be expected to yield estimates for the rate matrix $\mathbf{Q}$ and divergence times $\tau(x, y)$. To resolve the scaling indeterminacy, a normalisation condition has to be imposed on either $\mathbf{Q}$ or the divergence times, for example $\mathsf{Tr}(\mathbf{Q}) = -1$ (or $\|\mathbf{Q}\|_F = 1$ as in [8]).

3.3 Multivariate analysis of observed rate matrices

We now address the problem of comparing sufficiently related sequences using observed rate matrices and without additional assumptions. Consider a set of p pairwise alignments (x, y) of sufficiently related sequences as defined in Definition 6. To each pair (x, y) is associated an observed rate matrix $\mathbf{L}^{(x,y)}$, which provides an m^2-dimensional representation of the alignment.

In the biological applications described below we mainly focus on two aspects, namely the symmetry and the adequacy of models associated with a unique rate matrix $\mathbf{Q}$. For that we will resort to multivariate analysis techniques, in particular adaptations of principal component analysis (PCA for short, see *e.g.* [17] for a recent review), which we briefly outline here.

PCA is a very simple and routinely used tool for exploratory data analysis. Given an $n \times p$ data matrix $\mathbf{X}$, PCA provides orthonormal bases of the space of rows and the space of columns of $\mathbf{X}$, denoted respectively by $\{\mathbf{V}_k\}$ and $\{\mathbf{U}_\ell\}$. These are eigenvectors of the matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ respectively[3]. Eigenvalues are real and non-negative, and conventionally ordered in decreasing order (basis vectors are sorted accordingly). They represent the standard deviations of the projections of rows and columns of $\mathbf{X}$ onto the axes generated by corresponding basis vectors.

The coordinates of rows and columns with respect to these bases can often be given a sensible interpretation. It is customary to represent graphically projections onto subspaces spanned by the first eigenvectors (for convenience two-dimensional subspaces are chosen, the so-called first *factorial planes*). Also of interest are the weights of the expansion of basis vectors $\{\mathbf{V}_k\}$ (resp. $\{\mathbf{U}_\ell\}$) as linear combinations of rows (resp. columns) of $\mathbf{X}$, which we will call *contributions*.

Remark 2 (Mutabilities, ldet) Besides observed rate matrices, other simple quantities are also worth investigating. For example, the mutabilities (diagonals of observed rate matrices) defined in (8) can be analysed in the same way and provide similar or complementary conclusions. While rate matrix elements are labelled by pairs of amino acids, mutabilities are vectors labeled by amino acids. Biologists prefer to

[3] Standard PCA often involves prior centering of the columns of X, and sometimes an additional normalization.

use mutabilities because they are easier to interpret. This is what we will do in the examples below.

Also, traces of observed rate matrices (*i.e.* sum of diagonal elements) coincide with the ldet distance defined in (9), and therefore provide information relative to divergence times of sequences.

4 Biological validation of observed rate matrices

The above approach is very general, and provides a representation of alignments which enables in particular to check if $\mathbf{L}^{(x,y)}$ is indeed equal to $\mathbf{L}^{(y,x)}$ or if the rate matrix $\mathbf{Q}$ is the same for all the sequences in the sample. These are necessary checks because, as we show below, this is not always the case. They provide a rational basis for the choice of a model of molecular phylogenies.

4.1 Rate matrices and subfunctionalization in reverse gyrases

For some genes, multiple copies are present in the genome. The different copies can have the same function, the duplication simply increasing the amount of protein produced, or different functions while keeping the same type of enzymatic activity; this is what biologists call a subfunctionalization. Subfunctionalization involves a modification of the protein sequence that goes beyond conservative substitutions, *i.e.* those where one amino acid is replaced by another which will play the same role. The associated mutation matrix must therefore be different from those usually observed since the latter were calculated on sequences whose function was conserved during evolution. The whole question is whether the differences come out of the noise enough to be visible.

We approached the problem by studying the duplication of a gene, the reverse gyrase[4], because we know from biological data that there are cases of subfunctionalization. (See the introduction and section 4 of [2] where these points are discussed in more detail). In *Sulfolobus* for example, the two genes encoding the reverse gyrase are essential, the two proteins have different enzymatic properties and have a specific regulatory pathway [11, 10].

The set of sequences that we used contained 17 reverse gyrases representative of the biodiversity of hyperthermophiles and the duplicated genes *topR1* and *topR2* from three *Sulfolobus*, namely *S. acidocaldarius*, *S. solfataricus* and *S. tokodaii*. We performed a pairwise alignment of the 23 reverse gyrases. The sequences are close enough (according to Definition 6) to allow the computation of the ldet divergences, following equation (9).

[4] The reverse gyrase exists mostly in bacteria and archaea whose growth optimum is above 80 °C; it protects DNA from the denaturation that normally occurs at such high temperatures [11].

The distribution of the asymmetry $|\mathsf{ldet}(x, y) - \mathsf{ldet}(y, x)|$ is displayed in **Fig. 3**. A first remark to be made: the main mode of the asymmetry distribution is not located at the zero value (it is located between 0.4 and 0.5); in addition, the 22 values greater than 2.5 forming the second mode all correspond to the alignments involving the reverse gyrase of the bacteria *Thermus thermophilus*.

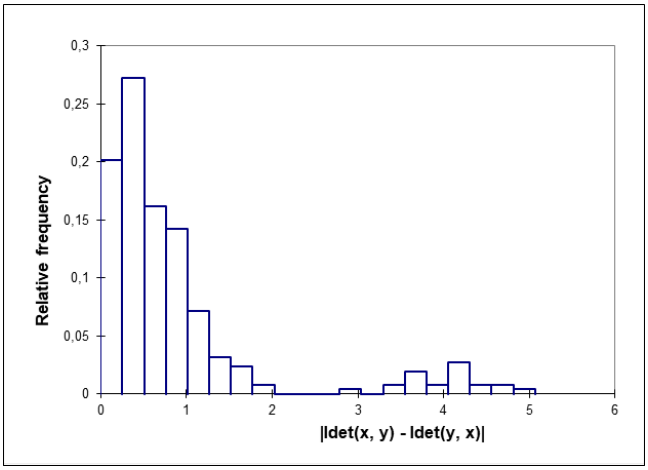

Fig. 3 Distribution of the asymmetry $|\mathsf{ldet}(x, y) - \mathsf{ldet}(y, x)|$ for the 253 alignments of the 23 reverse gyrases. The 22 values larger than 2.5 correspond to alignments involving the reverse gyrase of the bacteria *Thermus thermophilus*.

Here we will only discuss the comparisons between *topR1* and *topR2* in *Sulfolobus*. As shown in **Table 2**, the asymmetry is much weaker in their case, but it is not zero (the identifiers are defined in **Table 1**).

Species	Gene	Identifier	Species	Gene	Identifier
S. acidocaldarius	*topR1*	A1	*S. acidocaldarius*	*topR2*	A2
S. solfataricus	*topR1*	B1	*S. solfataricus*	*topR2*	B2
S. tokodaii	*topR1*	C1	*S. tokodaii*	*topR2*	C2

Table 1 Definition of identifiers used in **Table 2** and **Table 3**.

Table 3 displays the differences $\mathsf{ldet}(x_1, y_2) - \mathsf{ldet}(x_2, y_2)$ between a pair (x_1, y_1) of sequences *topR1* and the peer pair (x_2, y_2) of sequences *topR2*. The average value of this difference approximately equals -2.3. Since $\mathsf{ldet}(x, y) = \tau(x, y)\mathsf{Tr}(\mathbf{Q})$ (equation (9)) and as τ is the same for *topR1* and *topR2* for any given couple of species (this is the time since the two species evolved separately), this means that a unique rate matrix $\mathbf{Q}$ cannot describe both *topR1* and *topR2*. This proves that the subfunctionalization is associated with a modification of the observed rate matrix $\mathbf{L}$.

We now turn to multivariate analysis of mutabilities $\mu^{(x,y)}$ (*i.e.* diagonals of observed rate matrices, see (8)). Although the set of such vectors may be seen geo-

Alignment	$\mathsf{ldet}(x, y)$	Alignment	$\mathsf{ldet}(y, x)$	$\mathsf{ldet}(x, y) - \mathsf{ldet}(y, x)$
A1.B1	13.03	B1.A1	13.07	-0.04
A1.C1	9.79	C1.A1	9.89	-0.10
B1.C1	11.90	C1.B1	11.91	-0.01
A2.B2	15.11	B2.A2	15.03	0.08
A2.C2	12.22	C2.A2	12.08	0.14
B2.C2	14.47	C2.B2	14.36	0.11

Table 2 Reverse Gyrases: asymmetry of the ldet "distance" between *topR1* and *topR2* in *Sulfolobus*.

Alignment	$\mathsf{ldet}(x_1, y_1)$	Alignment	$\mathsf{ldet}(x_2, y_2)$	$\mathsf{ldet}(x_1, y_1) - \mathsf{ldet}(x_2, y_2)$
A1.B1	13.03	A2.B2	15.11	-2.08
A1.C1	9.79	A2.C2	12.22	-2.43
B1.A1	13.07	B2.A2	15.03	-1.95
B1.C1	11.90	B2.C2	14.47	-2.57
C1.A1	9.89	C2.A2	12.08	-2.19
C1.B1	11.91	C2.B2	14.36	-2.45

Table 3 Reverse Gyrases: differences between ldet values for type R1 proteins (denoted by $\mathsf{ldet}(x_1, y_1)$) and homologous type R2 proteins (denoted by $\mathsf{ldet}(x_2, y_2)$). The average value approximately equals -2.3.

metrically as a cloud of points in a 20-dimensional space, they actually lie (to some extent) in a subspace of much smaller dimension. Performing an uncentered PCA (see section 3.3 above) on this dataset gives a satisfactory image of this dimension reduction: the first two axes here account for 59 % (35% + 24%) of the variance. There is an almost perfect homothety between the values of $\mathsf{ldet}(x, y)$ and the coordinates on the first axis: a linear regression gives $\text{axis1}(x, y) \approx 0.24\ \mathsf{ldet}(x, y)$ ($R^2 = 0.97$).

However, the observation of the projection onto the first factorial plane (*i.e.* the plane generated by the first two principal components) displayed in **Fig. 4** provides more information than the simple calculation of $\mathsf{ldet}(x, y)$. We see for example that the second axis separates $\mu^{(x,y)}$ and $\mu^{(y,x)}$ and therefore outlines the asymmetry mentioned above. In addition, **Fig. 5** shows that the projection onto the plane 2-3 both outlines the asymmetry (axis 2) and separates $\mu^{(x_1,y_1)}$ and $\mu^{(x_2,y_2)}$ (axis 3, which accounts for 7% of the variance), *i.e.* matrices corresponding to pairs of sequences *topR1* and the peer pairs *topR2*. We could not observe clear structures in the higher dimensions.

The analysis of the contribution makes it possible to give biological significance to these observations by highlighting the amino acids whose mutability varies according to the matrices $\mathbf{Q}$.

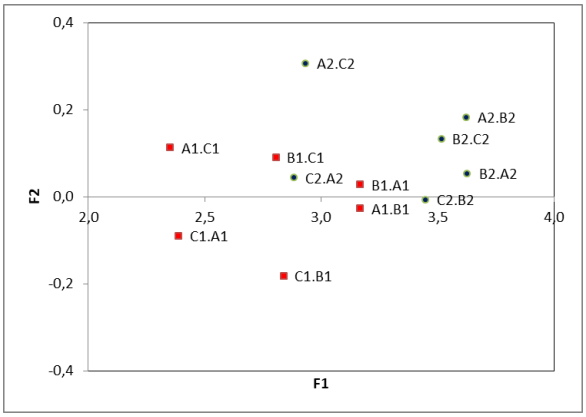

Fig. 4 Reverse gyrases: projections of mutabilities $\mu^{(x,y)}$ onto the first factorial plane of the principal component analysis. Each point represents an alignment between *topR1* or *topR2*. See Table 1 for the definitions of identifiers.

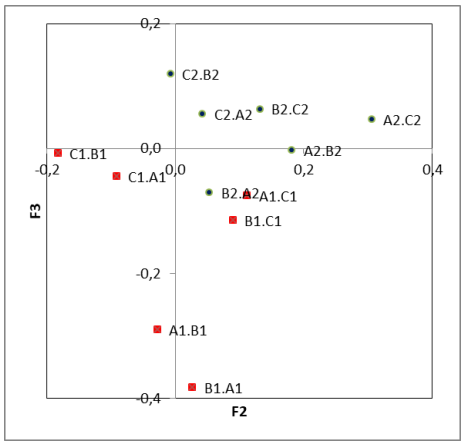

Fig. 5 Reverse gyrases: projections of mutabilities $\mu^{(x,y)}$ onto the plane generated by principal components 2 and 3. Each point represents an alignment between *topR1* or *topR2*. See Table 1 for the definitions of identifiers.

4.2 Several rate matrices within a protein family: the case of mitochondrial proteins

Mitochondria are organelles found in almost all eukaryotic cells (*i.e.* in all living organisms except bacteria and archaea). They contain the respiratory chain. Mitochondria have their own genome, which encodes for a subset of the proteins of the respiratory chain (they are called *mtDNA-encoded proteins* in the following).

We compared the *mtDNA-encoded proteins* of 120 representative species of animals: arthropods, tetrapods, echinoderms, molluscs and roundworms. These are very

different groups: 500 million years ago the ancestors of today's arthropods were already totally different from the ancestors of vertebrates. Arthropods have diversified throughout geological time: the ancestors of spiders and scorpions already existed 500 million years ago while the ancestors of insects appeared 400 million years ago, at the same time as the first tetrapods [19]. Mammals are much more recent [29]. Just before the disappearance of the dinosaurs (65 million years ago) and especially during the ten million years that followed, mammals underwent an explosive diversification. The sample also contains pairs of species that have diverged for several million years (*e.g.* man and chimpanzee, different species of Drosophila - small fruit flies). It therefore allows us to analyze the evolution over a very large time scale.

We aligned the 12 *mtDNA-encoded proteins* that are present in all the species considered here. Transition matrices $\mathbf{P}^{(x,y)}$ were computed for all the 14 280 pairs, after summation over the 12 proteins of two species x and y. Sequences were sufficiently related (according to Definition 6), to allow the computation of the observed rate matrices $\mathbf{L}^{(x,y)}$.

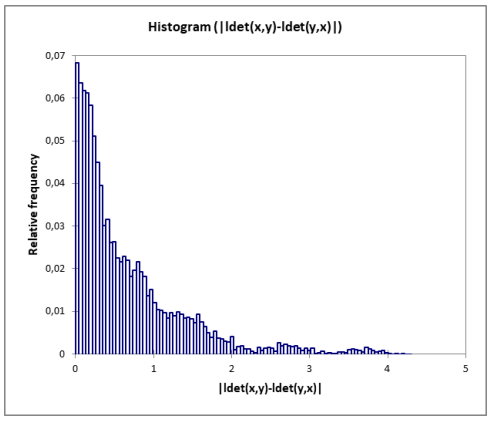

Fig. 6 Distribution of the asymmetry $|\mathsf{ldet}(x,y) - \mathsf{ldet}(y,x)|$ for the 14 280 alignments of the *mtDNA-encoded proteins* of 120 species. Values between 3 and 4 all correspond to alignments involving bees.

The distribution of the asymmetry $|\mathsf{ldet}(x,y) - \mathsf{ldet}(y,x)|$ is displayed in **Fig. 6**. The main mode of the asymmetry distribution is around the zero value and 95% of the values are less than 2. However, the difference can be as high as 4. Values between 3 and 4 all correspond to alignments involving bees.

As before, the mutabilities $\mu^{(x,y)}$ can be viewed geometrically as points in a 20 dimensional space. As the cloud of points is very elongated, the PCA gives a satisfactory image. The first axis represents 66% of the variance, the second 15% and the third 8%, all others are below 3%. The dominance of axis 1 over axis 2 means, as already mentioned, that a MCT model with a single generator is approximately valid for the proteins in our set, axis 1 roughly corresponding to the divergence time τ.

Fig. 7 displays the projections of the 22×21 mutabilities $\mu^{(x,y)}$ corresponding to alignments within insects excluding bees (in blue) and the 21×20 vectors $\mu^{(x,y)}$ corresponding to alignments within mammals (in red).

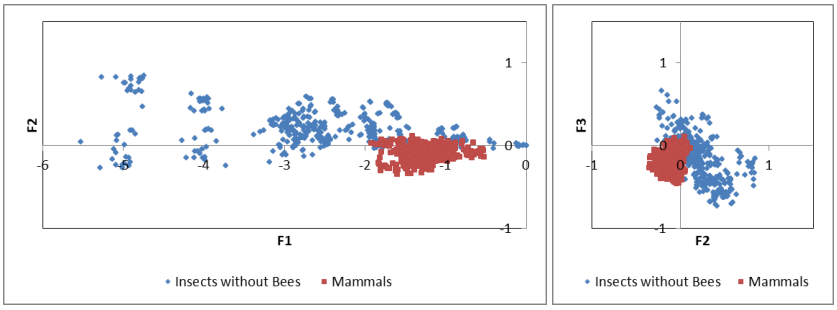

Fig. 7 *mtDNA-encoded proteins*: projections of the vectors $\mu^{(x,y)}$ onto the first factorial planes of an uncentered principal component analysis. Each point represents an alignment between two species. Points corresponding to alignments are identified by blue dots (insect–insect) or red dots (mammal-mammal). The projection strongly suggests the existence of two different generators **Q** for insects and mammals.

The staggering of the alignments along axis 1 reflects the time that has elapsed since the species have diverged: mammals are grouped to the right near the origin of the cloud, while alignments involving insects are far to the left of the cloud.

However, a closer examination of the projection on the second and third axis shows biologically significant deviations. The points corresponding to the vectors $\mu^{(x,y)}$ mammal-mammal (red) form a cluster clearly disjointed from that of the vectors $\mu^{(x,y)}$ insect-insect (blue). Clearly, a unique rate matrix **Q** cannot account for these two clusters.

It is perhaps not surprising that groups that have evolved separately for more than 500 million years correspond to different **Q** matrices. It is very surprising, however, to observe a difference between bees and other insects (**Fig. 8**). This observation is consistent with other works on the mitochondrial genome of bees which have shown that its evolution presents peculiarities [6, 30]. It should be noted that these authors took into account other criteria than the protein sequence. This difference may be due to the extremely high A+T/G+C ratio and the very small number of reproductive individuals in bees. Studies on other organisms have shown high genome instability under these conditions [21].

5 The influence of dissimilarity measures on sequence clustering and phylogeny reconstruction

The analysis of multiple alignment based upon rate matrices is interesting in several respects. On the one hand, it assumes as little as possible about data, of which

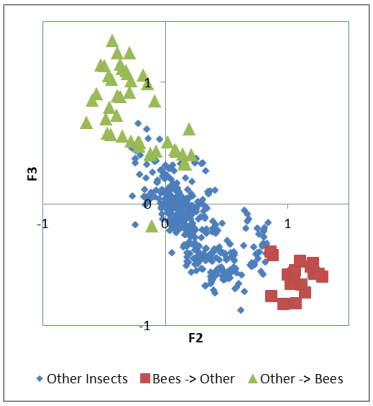

Fig. 8 *mtDNA-encoded proteins*: projections of the mutabilities $\mu^{(x,y)}$ onto the factorial plane 2-3 of an uncentered principal component analysis. Each point represents an alignment between two species. Points corresponding to alignments are identified by blue dots (insects other as bees), red dots (bees → other) and green dot (other → bees). The projection strongly suggests the existence of two different generators **Q** for bees and other insects.

it provides model-free representations (even though largely inspired by Markov evolution ideas). On the other hand, these representations turn out to provide valuable information about evolution. It is a good example which shows the existence of two distinct generators for the sequence family under consideration. Such information may in turn be used to fine tune models.

However, such an approach alone does not directly meet the expectations of biologists, who often seek a tree or a clustering that summarizes the information contained in dissimilarity matrix. Even though there exist many techniques that can generate such trees or clustering, the demand is less trivial than it appears. Indeed, the available, biologically relevant dissimilarity measures are numerous and varied (including the ldet distance introduced above), not to mention that some of them do not satisfy the necessary assumptions that allow using these techniques. This is actually an inexhaustible source of debate for biologists.

5.1 An iterative-rank based clustering

To overcome the dependence on the choice of a dissimilarity measure (and a tree building algorithm), Alex proposed an alternative solution [7]: assume that what really matters in dissimilarities between sequences is their relative order, not their numerical value. The choice of a particular measure then no longer has any importance since they will all give the same result as long as they are all related by a monotonic transformation. As for the dependence on tree building/clustering technique, the solution of [7] rely on an iterative procedure which we detail below.

The agreement of judgments

Consider a set X of species (*i.e.* sequences) of size $n = \sharp X$, and a matrix D on X, seen as a function $D : X \times X \to \mathbb{R}_+$ such that $D(x,x) = 0$ for all $x \in X$. D may be chosen symmetric, but not necessarily. Associate with D a family $\{D_x,\ x \in X\}$ of evaluation maps defined by

$$D_x(y) = D(x,y)\ . \tag{12}$$

These maps provide a description of how each $x \in X$ (hereafter called *judge*) evaluates all elements of X (the *candidates*). To achieve the announced goal, *i.e.* get rid of numerical values and preserve ordering, each map D_x is replaced with the corresponding rank $r_{D_x} : y \to r_{D_x}(y) \in \mathbb{N}$ defined by

$$r_{D_x}(y) = \sharp\{z \in X,\ D_x(z) \leq D_x(y)\}\ . \tag{13}$$

It may be shown that for all $x, y, z \in X$, $r_{D_x}(y) \leq r_{D_x}(z)$ is equivalent to $D(x,y) \leq D(x,z)$, *i.e.* ordering is preserved.

The *agreement of judgments* of two judges $x, y \in X$ may be measured using any mapping $T : \mathbb{N}^n \times \mathbb{N}^n \to \mathbb{R}_+$, by evaluating $T(r_{D_x}, r_{D_y})$.

Definition 8 With the above notations, the T-derivate of the matrix $D : X \times X \to \mathbb{R}_+$ is the map $\partial_T D : x, y \in X \longrightarrow \partial_T D(x,y) \in \mathbb{R}_+$ defined by

$$\partial_T D(x,y) = T\left(r_{D_x}, r_{D_y}\right)\ . \tag{14}$$

$\partial_T D(x,y)$ thus provides a quantitative measure of the agreement of judges x and y on X. In [7], the squared Euclidean distance $T_2 : (u,v) \to \sum_{x \in X}(u(x) - v(x))^2$ is used (also studied in Spearman rank statistics).

Iteration

Clearly enough, if T is a symmetric map, the T-derivate $\partial_T D$ of the matrix D is always a dissimilarity matrix. This also suggests to iterate the procedure, thus deriving a whole family of dissimilarities $\partial_T^\ell D : X \times X \to \mathbb{R}$ ($\ell = 0, 1, \ldots$) from any matrix D, defined recursively by

$$\partial_T^0 D := D \quad \text{and} \quad \partial_T^{\ell+1} D := \partial_T\left(\partial_T^\ell D\right) \tag{15}$$

Since X is finite, the sequence $(\partial_T^\ell D)_\ell$ necessarily runs into a cycle, but there is no reason to expect that this iteration should converge. However, the authors of [7] observe that in their experiments, for most distance data obtained either by comparing biological sequences or by random simulation, there was always some integer i_0 of about the same order of magnitude as $\sharp X$ such that $(\partial_{T_2})^{i_0} D$ is a fixed point of ∂_{T_2}.

Clustering and tree construction

In cluster analysis, a standard task is to associate, to any dissimilarity matrix D, a *Linnean hierarchy* $\mathcal{H} = \mathcal{H}(D)$, *i.e.* a collection $\mathcal{H}$ of subsets $A, B, \ldots$ of X such that $A \cap B \neq \emptyset$ implies $A \subseteq B$ or $B \subseteq A$.

Introduce the collection

$$\mathcal{A}_D = \mathcal{A}_D(X) := \{A \subseteq X | a, b \in A, x \in X \backslash A \Rightarrow D(a,b) < D(a,x)\} \tag{16}$$

of subsets A of X such that, for any $a \in A$, any other $b \in A$ is "closer" to a than any $x \in X$ outside A. These subsets always form a Linnean hierarchy, and are therefore natural candidates for being *clusters* in applications.

Further, denote by $B_D(x : y)$ the smallest ball (with respect to D) with center x containing y or, in other words, the set of all z in X that are, relative to D, at least as "close" to x as y. One has [7] for all $x, y \in X$

$$r_{D_x}(y) = \#B_D(x : y) \ . \tag{17}$$

So, while the actual values of D might be debatable, one only needs to trust that one can use D to decide, for any three distinct objects x, u, v in X, whether u or v is more similar to x. And only the resulting *rankings* of the objects $u, v, \ldots$ in X relative to the objects $x, \ldots$ in X are needed to define $\mathcal{A}_D$.

Remark 3 (Practical considerations)

1. Given the way the rank is defined, the maximal rank value in a cluster C equals the cluster size $\sharp C$. This simplifies significantly the determination of clusters. Indeed, in order to find all the clusters of size N (groups of N judges having the same view on the candidates), it suffices to browse the rows (which correspond to judges) of the final rank matrix and find all values with rank less than N.
2. The chosen definition of ranks also facilitates the computation of the the distance between two objects (two judges) for the construction of the tree. This distance turns out to be equal to the maximum rank between the two objects, minus 1.
   This distance is ultrametric, *i.e.* the two largest distances of a triplet are equal to each other, which defines a hierarchy that can be represented by a dendogram.

5.2 Application to mtDNA-encoded proteins of tetrapods

We present below a study of the phylogeny of tetrapods based on *mtDNA-encoded proteins*, already discussed in section 4.2. As this phylogeny is very firmly established at the scale considered here, it makes it possible to assess the reliability of phylogeny reconstruction programs. A common pitfall of the latter is that they separate species which actually have a common ancestor. This apparent non-monophyly is due to poor management of the differences between the transition matrices of the different

branches. The *Iterative-Rank Clustering* described above turns out to give results of a quality quite comparable to that of the most frequently used software suites.

The sample is composed of 49 species representative of Amphibians, Reptiles, Birds and Mammals. It covers about 400 million years. The species belong to 12 clearly separated monophyletic categories but the divergence can be significant in some branches of a given category. However, Prototheria (platypus) is relatively close to Metatheria (marsupial) and Tubulidentata (aardvark) to Cetartiodactyla (ruminant, cetacean).

The 12 *mtDNA-encoded proteins* that are present in the 49 species under consideration have been aligned, and the matrices $\mathbf{P}^{(x,y)}$ have been calculated for all the pairs of sequences. The resulting observed rate matrices could be obtained, since the sequences are sufficiently related (according to Definition 6). The sample is homogeneous (a single matrix $\mathbf{Q}$) and asymmetry $|\mathsf{ldet}(x,y) - \mathsf{ldet}(y,x)|$ is weak, the mode is around 0.15 and 95% of values are less than 0.45. In the study below, we use the average $\frac{1}{2}(\mathsf{ldet}(\mathsf{x},\mathsf{y}) + \mathsf{ldet}(\mathsf{y},\mathsf{x}))$, but the same results are obtained using $\mathsf{ldet}_{\mathsf{min}}(\mathsf{x},\mathsf{y})$ or $\mathsf{ldet}_{\mathsf{max}}(\mathsf{x},\mathsf{y})$.

	e.g.	Nb Species	ldet Rank	ldet NJ	ProtDist NeighborNet	ProtDist NJ
Amphibia	frog	4	4	4	4	4
Testudines	turtle	4	4	4	4	4
Squamata	snakes	3	3	**2 | 1**	3	**2 | 1**
Paleognathae	ostrich	7	7	7	7	7
Neognathae	chicken	7	7	7	**4 | 3**	**5 | 2**
Crocodylidae	alligator	2	2	2	**1 | 1**	2
Prototheria	platypus	1	1	1	1	1
Metatheria	marsupial	2	2	2	2	2
Tubulidentata	aardvark	1	1	1	1	1
Cetartiodactyla	ruminant	5	5	5	5	5
Lagomorpha	rabbit	2	2	2	2	2
Primata	monkey	6	**4 | 2**	**4 | 2**	**4 | 2**	6
Rodentia	mouse	5	**3 | 2**	5	5	5

Table 4 Reliability of the reconstitution of the phylogeny of tetrapods from *mtDNA-encoded proteins*. The program makes a mistake when it splits the species of a category into several groups. This is the case for example with ldet + BioNJ and ProtDist + BioNJ which distinguish two groups of snakes (one with 2 species and the other with 1) while they are monophyletic.

Tetrapod phylogenies have been constructed using various standard programs. These are based upon two ingredients:

1. distance matrices: we used ldet (defined in equation (9)) and a distance calculated using the ProtDist[5] software, see [24].

[5] ProtDist provides a distance measure for protein sequences, using maximum likelihood estimates based on amino acid scoring matrices. It uses the multiple sequence alignment provided by the user.

2. a tree or network reconstruction algorithm: we used BioNJ [12] (a variant of Neighbor Joining[6]), iterative rank (section 5.1) and NeighborNet[7] (see [15]).

Table 4 summarizes the groupings proposed by the phylogeny construction programs. It should also be noted that all the programs show the proximity of Prototheria to Metatheria and of Tubulidentata to Cetartiodactyla. All the programs made at least two errors (three in the case of ProtDist + NeighborNet) in finding non-monophyly, but not necessarily for the same species.

The nature of the errors depends both, on the distances and on the method used to construct the tree. One way to detect them is therefore to compare the results of several independent analyzes. In this context, the iterative rank clustering approach developed by Alex is of great interest in several respects. On the one hand, it is not redundant with the commonly used methods. In addition, the user is not likely to give in to the temptation to tweak the options until he obtains the result he wants.

5.3 The impact of symmetry assumptions

Experience shows that given a pair of sequences (x, y), there is always a more or less significant gap between $\mathsf{ldet}(x, y)$ and $\mathsf{ldet}(y, x)$. This gap is mainly due to the asymmetry of the matrix of transition frequencies $\mathbf{F}^{(x,y)}$ as defined in equation (4). Since the early work of Margaret Dayhoff, the problem has been evaded by symmetrizing the matrices $\mathbf{F}^{(x,y)}$ and the vectors $\pi^{(x,y)}$ (see equation (10)). This choice has no consequences if the difference is small, as in the case presented in **Table 4**, it is, however, very questionable when the asymmetry is important as in the case of **Fig. 3** and **Fig. 8**. It amounts to attempting to characterize a bimodal distribution by its mean, which is hardly not the most relevant characteristic value in this case!

With the exception of *iterative-rank based clustering*, the methods generally used, notably those cited in **Table 4**, assume that the data have been symmetrized previously. This is a drawback because the exploitation of asymmetry opens new possibilities, which can highlight different aspects of the alignments.

We consider here four symmetric measures that can be derived from asymmetric quantities, in this case ldet. Given an alignment (x, y), we consider $\mathsf{ldet}_{\mathsf{min}}(x, y) = \min(\mathsf{ldet}(y, x), \mathsf{ldet}(x, y))$ and $\mathsf{ldet}_{\mathsf{max}}(x, y) = \max(\mathsf{ldet}(y, x), \mathsf{ldet}(x, y))$. In the spirit of section 5.1, we also use the rank-based dissimilarities $D_1(x, y) = \partial_{T_2}\mathsf{ldet}(x, y)$ and the similar quantity $D_2(x, y) = \partial_{T_2}\mathsf{ldet}^t(x, y)$ built from the transposed matrix ldet^t of ldet. Here, ∂_{T_2} is the T_2-derivate (see section 5.1), T_2 being the squared Euclidean distance. These four solutions are not mutually exclusive (one may also consider

[6] Neighbor joining is an agglomerative (*i.e.* aggregation from leaves to root) clustering method for the creation of phylogenetic trees that only requires the knowledge of the distance between each pair of taxa (*e.g.*, sequences) to form the tree. It evaluates branch lengths so that the distances deduced from the tree are closest to the values in the distance table.

[7] NeighborNet is similar to Neighbor Joining, except that it can lead to overlapping clusters which do not form a hierarchy, and are represented using a type of phylogenetic network called a splits graph.

others, for example higher order T_2 derivates of ldet); in fact they turn out to provide complementary points of view on the set of observed rate matrices.

We used this approach to analyze the phylogeny of duplicated genes in reverse gyrases (see section 4.1) using the Neighbor Joining method to build the trees. The $\mathsf{ldet}_{\mathsf{max}}$ matrix shows, as already known, that the *topR1* and *topR2* genes have a common origin. The matrix $\mathsf{ldet}_{\mathsf{min}}$ and the two dissimilarity matrices D_1 and D_2 highlight the phylogenetic relationships of *Sulfolobus* and *Aeropyrum pernix*, which are both Archaea of the class *Thermoprotei*. The tree groups the genes *topR1* of *Sulfolobus* and the gene *topR1* of *Aeropyrum pernix* on one branch and the genes *topR2* of *Sulfolobus* and the gene *topR2* of *Aeropyrum pernix* on another. We display in **Fig. 9** the reverse gyrase phylogenetic tree constructed from $\mathsf{ldet}_{\mathsf{max}}(x, y)$, and in **Fig. 10** the corresponding tree constructed from $\mathsf{ldet}_{\mathsf{min}}(x, y)$.

6 Discussion - Conclusion

One of the fundamental techniques of biology is sequence alignment, namely transforming one sequence into another with minimal change. Sequence alignment is essential for the study of evolution and is a source of information for the analysis of the physico-chemical mechanisms which are at the heart of protein activity.

Almost all multiple alignment programs use a guide tree to reduce complexity. Advanced programs proceed by iteration:

1. A first estimation of the distance between the sequences is made
2. A guide tree is computed based on these estimates
3. The sequences are progressively aligned following the order given by the guide tree
4. A new distance between the sequences is calculated after alignment
5. The program iterates at point 2 as long as the procedure improves the alignment score

A refinement consists in splitting the guide tree and proceeding in the same way in each sub-tree, the stopping criterion remaining the improvement of the alignment score (see Chapter 25 in [27]).

The procedure gives an optimal *Alignment - Tree pair* for a given measure of distance between sequences. The topology is almost frozen by the alignment program. The biologist may then use Monte-Carlo-type methods to get an idea of trees with roughly equivalent scores. It consists in randomly modifying the matrices **P** to identify the TRULY robust parts in the tree.

This approach differs considerably from the one we propose as we do not modify the matrices **P** at all, we simply change the angle under which we look at the matrices **L** in order to better perceive the proximities between the sequences.

Indeed, one can think that in many situations a single tree is not enough to faithfully summarize the information contained in an alignment, it can therefore be interesting to build several trees, exploiting different points of view, rather than

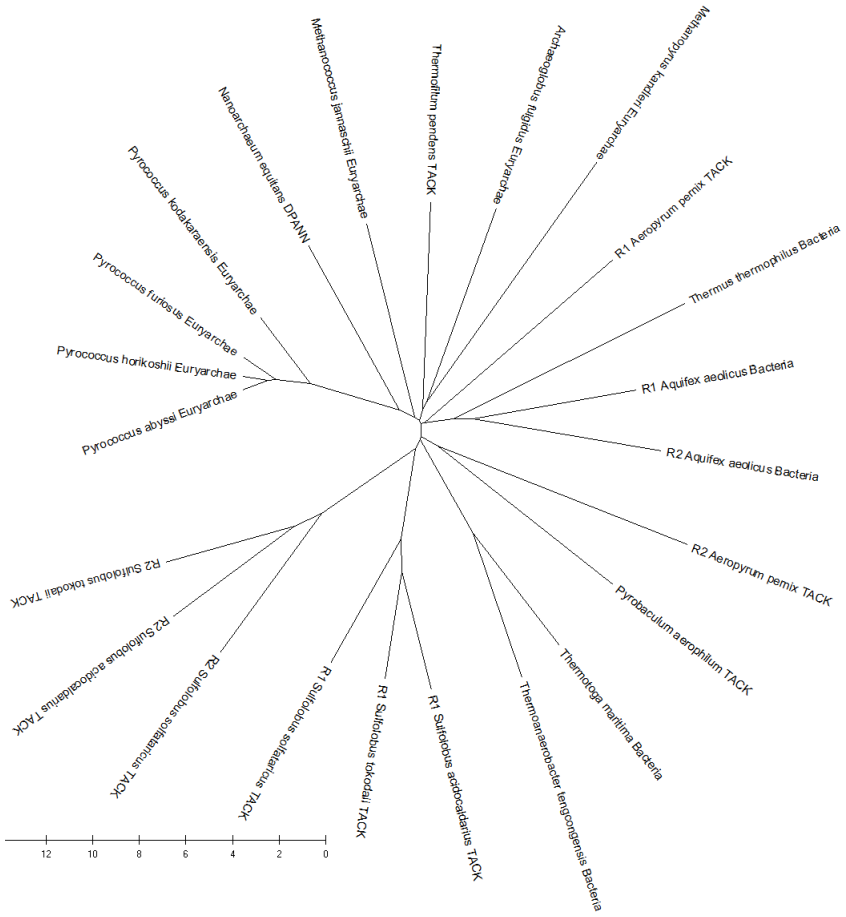

Fig. 9 Phylogenetic tree of reverse gyrase constructed with NJ from $\mathsf{ldet}_{\max}(x, y)$. The *topR1* and *topR2* proteins of *Sulfolobus* are clustered in this tree. In contrast, the proteins *topR1* and *topR2* of *Aeropyrum pernis* are separated. The identifiers consist of genus, species and group. They are archaea when not specified otherwise.

trying to make a single one and trying to show that it is significant by bootstrap methods or similars.

The point of view promoted by Alex was to assume as little as possible and try to collect information from data, before turning to explicit modeling if needed "*We must look for a model that fits the data and not twist the data to fit the model*".

The starting point here is to avoid modeling multiple alignments (and therefore introducing trees) and try to see which information could be obtained from pairwise alignments. Another originality was to compare pairwise alignments, not only sequences.

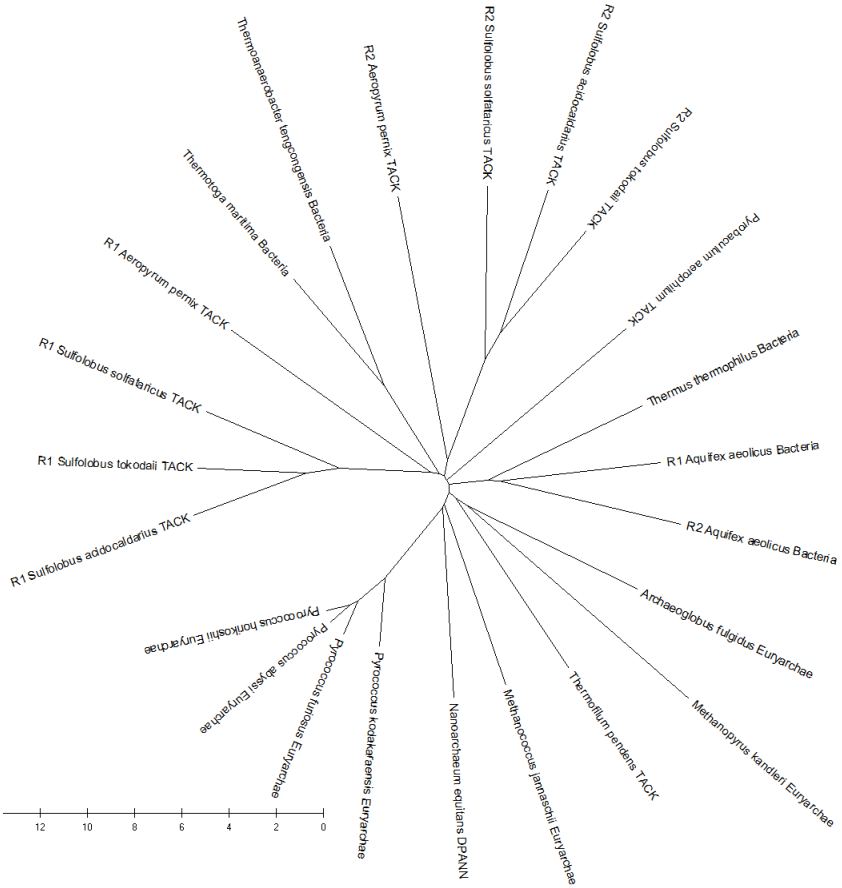

Fig. 10 Phylogenetic tree of reverse gyrase constructed with NJ from $\mathsf{ldet}_{\mathsf{min}}(x, y)$. The *topR1* proteins of *Sulfolobus* are grouped with the *topR1* proteins of *Aeropyrum pernis* in this tree (and similarly for the *topR2* proteins of all four species) but *topR1* and *topR2* are separate. The identifiers consist of genus, species and group. They are archaea when not specified otherwise.

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. J. Adachi and M. Hasegawa. Amino acid substitution of proteins coded for in mitochondrial DNA during mammalian evolution. *Idengaku zasshi*, 67:187–97, 07 1992.
2. l. j. Bande. Four billion years: the story of an ancient protein family. This volume, 2021.
3. D. Barry and J. A. Hartigan. Statistical analysis of hominoid molecular evolution. *Statist. Sci.*, 2(2):191–207, 05 1987.
4. R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.
5. M. Dayhoff, R. Schwartz, and B. Orcutt. A model of evolutionary change in proteins. In M. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, pages 345–352. National Biomedical Research Foundation, Washington, D. C., 1978.
6. de Paula Freitas, C. Flávia, A. P. Lourenço, F. M. F. Nunes, A. R. Paschoal, F. C. P. Abreu, F. O. Barbin, L. Bataglia, C. A. M. Cardoso-Júnior, M. S. Cervoni, S. R. Silva, F. Dalarmi, M. A. Del Lama, T. S. Depintor, K. M. Ferreira, P. S. Gória, M. C. Jaskot, D. C. Lago, D. Luna-Lucena, L. M. Moda, L. Nascimento, M. Pedrino, F. R. Oliveira, F. C. Sanches, D. E. Santos, C. G. Santos, J. Vieira, A. R. Barchuk, K. Hartfelder, Z. L. P. Simões, M. M. G. Bitondi, and D. G. Pinheiro. The nuclear and mitochondrial genomes of frieseomelitta varia - a highly eusocial stingless bee (meliponini) with a permanently sterile worker caste. *BMC Genomics*, 21(1):386, June 2020.
7. C. Devauchelle, A. W. M. Dress, A. Grossmann, S. Grünewald, and A. Henaut. Constructing hierarchical set systems. *Annals of Combinatorics*, 8(4):441–456, Jan. 2005.
8. C. Devauchelle, A. Grossmann, A. Hénaut, M. Holschneider, M. Monnerot, J. Risler, and B. Torrésani. Rate matrices for analyzing large families of protein sequences. *J. Comput. Biol.*, 8(4):381–399, 2001. PMID: 11571074.
9. J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17(6):368–376, Nov. 1981.
10. F. Garnier, M. Couturier, H. Débat, and M. Nadal. Archaea: a gold mine for topoisomerase diversity. *Front. Microbiol.*, 2021. In press.
11. 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.
12. O. Gascuel and M. Steel. Neighbor-Joining Revealed. *Molecular Biology and Evolution*, 23(11):1997–2000, 07 2006.
13. H. E. Haber. Notes on the matrix exponential and logarithm. online, May 2019.
14. D. M. Hillis, C. Moritz, and B. K. Mable, editors. *Molecular Systematics*. Sinauer Associates Inc., 1996.
15. D. H. Huson and D. Bryant. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23(2):254–267, 10 2005.
16. V. Jayaswal, L. S. Jermiin, and J. Robinson. Estimation of phylogeny using a general Markov model. *Evolutionary bioinformatics online*, 1:62–80, Feb. 2007.
17. I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.
18. P. J. Lockhart, M. A. Steel, M. D. Hendy, and D. Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, 11(4):605–612, 07 1994.
19. B. Misof, S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, P. B. Frandsen, J. Ware, T. Flouri, R. G. Beutel, O. Niehuis, M. Petersen, F. Izquierdo-Carrasco, T. Wappler, J. Rust, A. J. Aberer, U. Aspöck, H. Aspöck, D. Bartel, A. Blanke, S. Berger, A. Böhm, T. R. Buckley, B. Calcott, J. Chen, F. Friedrich, M. Fukui, M. Fujita, C. Greve, P. Grobe, S. Gu, Y. Huang, L. S. Jermiin, A. Y. Kawahara, L. Krogmann, M. Kubiak, R. Lanfear, H. Letsch, Y. Li, Z. Li, J. Li, H. Lu, R. Machida, Y. Mashimo, P. Kapli, D. D. McKenna, G. Meng, Y. Nakagaki, J. L. Navarrete-Heredia, M. Ott, Y. Ou, G. Pass, L. Podsiadlowski, H. Pohl, B. M. von Reumont, K. Schütte, K. Sekiya, S. Shimizu, A. Slipinski, A. Stamatakis, W. Song, X. Su, N. U. Szucsich,

M. Tan, X. Tan, M. Tang, J. Tang, G. Timelthaler, S. Tomizuka, M. Trautwein, X. Tong, T. Uchifune, M. G. Walzl, B. M. Wiegmann, J. Wilbrandt, B. Wipfler, T. K. F. Wong, Q. Wu, G. Wu, Y. Xie, S. Yang, Q. Yang, D. K. Yeates, K. Yoshizawa, Q. Zhang, R. Zhang, W. Zhang, Y. Zhang, J. Zhao, C. Zhou, L. Zhou, T. Ziesmann, S. Zou, Y. Li, X. Xu, Y. Zhang, H. Yang, J. Wang, J. Wang, K. M. Kjer, and X. Zhou. Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767, 2014.

20. T. Müller and M. Vingron. Modeling amino acid replacement. *J. Comput. Biol.*, 7(6):761–776, 2000. PMID: 11382360.
21. D. T. Nguyen, B. Wu, S. Xiao, and W. Hao. Evolution of a Record-Setting AT-Rich Genome: Indel Mutation, Recombination, and Substitution Bias. *Genome Biology and Evolution*, 12(12):2344–2354, 09 2020.
22. S. Pietrokovski, J. G. Henikoff, and S. Henikoff. The Blocks Database—A System for Protein Classification. *Nucleic Acids Research*, 24(1):197–200, 01 1996.
23. 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.
24. Protdist. Program to compute distance matrix from protein sequences. online, 1993. https://evolution.gs.washington.edu/phylip/doc/protdist.html.
25. C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford lecture series in mathematics and its applications*. Oxford University Press, 2003.
26. M. Steel. Reconstructing evolutionary trees under a variety of Markov-style models. Proceedings of Phylogeny Workshop 95-48, DIMACS, Princeton University, 1995. 51-54.
27. D. Tagu and J.-L. Risler. *Bioinformatique ; Principes d'utilisation des outils*. Editions Quae, Paris, 2010.
28. R. Trivedi and H. A. Nagarajaram. Substitution scoring matrices for proteins - an overview. *Protein Sci.*, n/a(n/a), 2020.
29. N. S. Upham, J. A. Esselstyn, and W. Jetz. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol*, 17(12):e3000494–e3000494, Dec. 2019.
30. S.-j. Wei, M. Shi, M. J. Sharkey, C. van Achterberg, and X.-x. Chen. Comparative mitogenomics of braconidae (insecta: Hymenoptera) and the phylogenetic utility of mitochondrial genomes with special reference to holometabolous insects. *BMC Genomics*, 11(1):371, June 2010.