HAL
open science

Rate matrices for analyzing large families of protein sequences

Claudine Devauchelle, Alex Grossmann, Alain Hénaut, Matthias Holschneider, Monique Monnerot, Jean-Loup Risler, Bruno Torrésani

To cite this version:

Claudine Devauchelle, Alex Grossmann, Alain Hénaut, Matthias Holschneider, Monique Monnerot, et al.. Rate matrices for analyzing large families of protein sequences. Journal of Computational Biology, 2004, 8 (4), pp.381-399. 10.1089/106652701752236205 . hal-01300316

HAL Id: hal-01300316

https://hal.science/hal-01300316v1

Submitted on 10 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rate matrices for analyzing large families of protein sequences

C. Devauchelle* A. Grossmann†* A. Hénaut* M. Holschneider‡
M. Monnerot§ J.L. Risler* B. Torrésani¶

May 24, 2001

Key words: protein evolution, rate matrices, LogDet distances, mitochondrial evolution.

The names of the authors are given in alphabetical order

*Laboratoire Génome et Informatique, Tour Evry2, 523 Place des Terrasses, 91034 Evry Cedex. France. devauchelle@genopole.cnrs.fr

†Centre de Physique Théorique, CNRS Luminy, Marseille, France. grossman@cpt.univ-mrs.fr

‡Géoscience Rennes, Bât. 15, Campus de Beaulieu, Université de Rennes 1, 263 av. du maréchal Leclerc, CS 74205, 35042 Rennes Cedex. Matthias.Holschneider@univ-rennes1.fr

§Centre de Génétique Moléculaire, CNRS Gif-sur -Yvette, France. Monique.Monnerot@cgm.cnrs-gif.fr

¶Laboratoire d'Analyse, Topologie et Probabilités, CMI, Université de Provence, 39 rue Joliot-Curie, 13453 Marseille Cedex 09. Bruno.Torresani@sophia.inria.fr

Abstract

We propose and study a new approach for the analysis of families of protein sequences. This method is related to the *LogDet* distances used in phylogenetic reconstructions; it can be viewed as an attempt to embed these distances into a multi-dimensional framework.

The proposed method starts by associating a Markov matrix to each pairwise alignments deduced from a given multiple alignment. The central objects under consideration here are matrix-valued logarithms $\mathbf{L}$ of these Markov matrices, which exist under conditions that are compatible with fairly large divergence between the sequences. These logarithms allow us to compare data from a family of aligned proteins with simple models (in particular, continuous reversible Markov models) and to test the adequacy of such models. If one neglects fluctuations arising from the finite length of sequences, any continuous reversible Markov model with a single rate matrix $\mathbf{Q}$ over an arbitrary tree predicts that all the observed matrices $\mathbf{L}$ are multiples of $\mathbf{Q}$. Our method exploits this remark, without relying on any tree estimation.

We test this prediction on a family of proteins encoded by the mitochondrial genome of 26 multicellular animals, which include vertebrates, arthropods, echinoderms, molluscs and nematodes. A principal component analysis of the observed matrices $\mathbf{L}$ shows that a single rate model can be used as a rough approximation to the data, but that systematic deviations from any such model are unmistakable, and related to the evolutionary history of the species under consideration.

1 Introduction

We develop in this paper a new approach for analyzing large families of protein sequences. The proposed method is based upon the comparison of the Markov matrices associated with the pairwise alignments of the sequences under consideration, and allows us to analyze their compatibility with standard Markov models.

The use of Markov chain models (which was at least implicit in the early work of Dayhoff and collaborators (Dayhoff et al., 1972; Dayhoff et al., 1983)), was advocated by many authors in the context of the construction of phylogenetic trees from DNA sequences. Among them, let us quote (Felsenstein, 1981) and (Tavaré, 1986), and refer to the chapter 11 of (Hillis et al., 1996) for a systematic account of most significant contributions. Let us also mention (Müller and Vingron, 2000) for a different approach, which bears similarities with the techniques presented in this paper.
The aim of this approach is to provide a probabilistic evolution model describing a family of aligned sequences (a multiple alignment). All sites of the sequences are treated as independent identically distributed random variables (see for example (Steel, 1995) and (Tavaré, 1986) for a discussion of the consequences of such assumptions). In its most general form, Felsenstein's model is based on two ingredients: a (rooted) tree, whose leaves are the considered sequences, and a family of stochastic matrices (see below for a definition) associated with the branches of the tree. Along each branch of the tree, a sequence is supposed to undergo an evolution governed by a Markov chain. At each node, a sequence gives rise to two different sequences, each one continuing with its own Markov chain evolution.
These ingredients (the *parameters* of the model) are sufficient to compute the probabilities of all possible multiple alignments. The practical problem is the estimation: infer the values of the parameters from data at hand. Felsenstein's method is a standard maximum likelihood approach: the likelihood function (i.e. the probability of the multiple alignment under consideration), which depends upon the parameters of the model, is maximized. The parameters which realize the maximum of the likelihood function are the maximum likelihood estimators, and may be used for further studies. In the case of a Markov chain on a tree, the parameters are the stochastic matrices and the topology of the tree.
The maximization, which is to be performed numerically, turns out to become a difficult problem for large families of sequences, and some simplifications are often made (see however (Barry and Hartigan, 1987) for a general discussion of "parameter rich" models). In addition, the comparison of likelihoods for different tree topologies may be difficult (see the discussion in (Adachi and Hasegawa, 1992)).
The most common simplification amounts to assume that all the stochastic matrices associated to the branches of the tree are powers $\mathbf{P}^\tau$ of a single stochastic matrix $\mathbf{P}$, associated to some "universal" Markov chain. With such a simplification, the complexity of the model and of the estimation problem is reduced considerably: for a given tree topology, the remaining parameters are now the matrix $\mathbf{P}$ (a 20×20 matrix in the case of proteins) and the exponents τ of all the branches, interpreted as time parameters (the "ages" of the branches). However, even with such simplifications, the estimation problem is still difficult to solve for large families (say, for families of more than 20 sequences).
A further simplification consists in assuming that the Markov chain is "reversible", which again reduces by a factor 2 the number of parameters to be estimated, and allows a direct connection between data and model parameters. For this and other reasons, all models of sequence evolution used in the analysis of data -including the formalizations of the Dayhoff's pioneering work on protein sequences (Dayhoff et al., 1983)- are, as far as we know, continuous reversible single-generator Markov models.

As a preliminary to parameter estimation, one may ask to what extent the alignment data can be represented by *any* continuous reversible single-generator Markov models, i.e. whether there exists such a model that describes the data to some given degree of approximation. This is the main topic of this paper. We develop a method which yields estimators for the parameters of continuous reversible single generator Markov models when such models provide a reasonable description of alignment data, and gives indications on the departure from such models in the opposite case. Our method involves a transformation of the data into a form in which an underlying continuous reversible single-generator Markov model is immediately apparent. While some of the machinery may be a bit involved, the basic idea is not more complicated than the use of a logarithmic scale to help determining half-lives in radioactive decay measurements.

Starting with a multiple alignment, we consider all the pairwise alignments deduced from it. For each alignment of two sequences x and y, (denoted generically by "(x, y)"), we estimate a stochastic matrix $\mathbf{P}^{(x,y)}$ using maximum likelihood methods (which reduce in that case to simple countings). In the case of a reversible model, all such matrices $\mathbf{P}^{(x,y)}$ provide estimates for powers $\mathbf{P}^{\tau_{(x,y)}}$ of some "universal matrix" $\mathbf{P}$. And their matrix logarithms (when they exist) $\log \mathbf{P}^{(x,y)}$ provide estimates for multiples $\tau_{(x,y)}\mathbf{Q}$ of a unique matrix $\mathbf{Q} = \log \mathbf{P}$, called the *rate matrix* of the model. A linear regression (which in this case is performed via a principal components analysis) yields estimates for the parameters $\tau_{(x,y)}$ and the rate matrix $\mathbf{Q}$.

Because of our desire to study the adequacy of single-generator models, we do not systematically use symmetric counting procedures: the alignments (x, y) and (y, x) may thus yield significantly different matrices $\mathbf{P}^{(x,y)}$ and $\mathbf{P}^{(y,x)}$, which is a sign of departure from the single generator situation.

Not every matrix has a logarithm, and the "logarithmability" condition will play an important role in this paper as a restriction on the alignments that can be handled by our method. This restriction will mean, in practice, that the observed Markov matrices associated to pairwise alignments should be not too far from the identity. Such a restriction is apparently similar to the "one mutation" requirement made for the construction of PAM matrices (Dayhoff et al., 1983). At this point we should stress that the exact "logarithmability" condition, to be introduced below, is much less restrictive than the "one-mutation" requirement. In fact the transition to logarithms incorporates the possibility of arbitrarily many mutations at a given site, as long as they do not overwhelm the overall picture. In our experience, the breakdown of "logarithmability" is not far from the point where the alignments themselves become questionable.

Thanks to the fact that we only consider pairwise alignments, we do not have to dig into the problem of tree estimation. This is a drastic simplification. The parameters $\tau_{(x,y)}$ may be interpreted as "distances" between the sequences, very much in the spirit of the *LogDet* distances (Lockhart et al., 1994; Steel, 1995; Lake, 1994). Such distances may in turn be used for estimating a tree. However, such a tree need not be completely consistent with the underlying model.

As a byproduct, our method also yields graphical representations for the alignments (namely, the projection of the matrices from the 400 dimensional space onto the planes corresponding to the top eigenvectors in the principal components analysis), which helps in testing the homogeneity of the family of sequences under consideration.

2 Methods

2.1 From count matrices to logarithms of stochastic matrices: the non symetric case

Our starting point is a multiple alignment of "sufficiently related" sequences (the criterion for "sufficient relatedness" is introduced below). However, we limit our investigations to the analysis of all pairs of sequences in the multiple alignment, i.e. pairwise alignments, which will be the main object under consideration in this paper. In contrast to maximum likelihood methods (Adachi and Hasegawa, 1992; Felsenstein, 1981), we do not take into account multiple alignment information contained in columns of the multiple alignment.

As usual, we model a protein sequence as a sequence of letters in a finite alphabet of size m ($m = 20$ in our case). All the sites in the sequence are considered independent and identically distributed, and are therefore treated in the same way. Notice that we do not impose restrictions on the sites of the sequences to be considered. Given an ordered pair of sequences (x, y) in the multiple alignment under consideration, the sites containing an *indel* are removed from the pairwise alignment (but not from the multiple alignment).

Given an ordered pair (x, y) of aligned sequences, we first consider the count matrix, denoted by $\mathbf{C}^{(x,y)} = \{C_{ij}^{(x,y)}, i, j = 1, \ldots m\}$. Its elements are the numbers of pairs of amino acids in the alignment:

$$C_{ij}^{(x,y)} = \text{ number of sites } k \text{ where amino-acid } x(k) = i \text{ and } y(k) = j \ , \tag{1}$$

From $C^{(x,y)}$ we obtain the vectors of occurences $\mathbf{C}^{(x)}$:

$$C_i^{(x)} = \text{ number of sites } k \text{ where } x(k) = i \ . \tag{2}$$

The entries of $\mathbf{C}^{(x,y)}$ and $\mathbf{C}^{(x)}$ are non-negative integers. From these quantities, we obtain the vector of frequencies $\pi^{(x)} = (\pi_1^{(x)}, \ldots \pi_m^{(x)})$, defined by $\pi_i^{(x)} = C_i^{(x)}/c^{(x,y)}$, where $c^{(x,y)} = \sum_{i,j=1}^m C_{ij}^{(x,y)}$ is the length of the pairwise alignment (x, y). Notice that because of possible *insertions-deletions*, the vector of frequencies $\pi^{(x)}$ of the sequence x also depends on the sequence y (for the sake of simplicity we do not introduce a specific notation for that).

Finally, we also consider the matrices $\mathbf{F}^{(x,y)} = \{F_{ij}^{(x,y)}, i, j = 1, \ldots m\}$, defined by

$$\mathbf{F}^{(x,y)} = \frac{1}{c^{(x,y)}} \mathbf{C}^{(x,y)} \ , \tag{3}$$

and the diagonal matrices of frequencies

$$\Pi^{(x)} = diag(\pi^{(x)}) \tag{4}$$

By construction, $\mathbf{F}^{(x,y)}$ satisfies $\sum_{i,j=1}^m F_{ij}^{(x,y)} = 1$.

We shall limit ourselves here to sequences in which all amino acids are significantly represented. Therefore, all the frequencies $\pi_i^{(x)}$ may be assumed to be nonzero, i.e. the matrices $\Pi^{(x)}$ are non singular. We may then consider the matrix $\mathbf{P}^{(x,y)} = {\Pi^{(x)}}^{-1} \mathbf{F}^{(x,y)}$ defined by its matrix elements

$$P_{ij}^{(x,y)} = \frac{F_{ij}^{(x,y)}}{\pi_i^{(x)}} \ . \tag{5}$$

REMARK 1 $\mathbf{P}^{(x,y)}$ is clearly a stochastic matrix, *i.e.* its elements satisfy $0 \leq P_{ij}^{(x,y)} \leq 1$ for all i, j, and its rows sum to unity: $\sum_{j=1}^m P_{ij}^{(x,y)} = 1$ for all $i = 1, \ldots m$.

It is well known that if the multiple alignments were generated using a continuous reversible Markov model on a binary tree, and if the lengths of the sequences are large enough, as described in (Felsenstein, 1981), all the matrices $\mathbf{P}^{(x,y)}$ should be close to the powers $\mathbf{P}_0^{\tau(x,y)}$ of a unique stochastic matrix $\mathbf{P}_0 = e^{\mathbf{Q}_0}$, where $\mathbf{Q}_0$ is a rate matrix (see Appendix A for a definition). Then the question arises whether the family of matrices $\mathbf{P}^{(x,y)}$ under consideration are compatible with such a model, up to fluctuations.
A stochastic matrix $\mathbf{M}$ which may be written in the form $\mathbf{M} = e^{\mathbf{Q}}$ where $\mathbf{Q}$ is a rate matrix is said to be embeddable (see Appendix A for more details). The problem we address here is in some sense more complex, as we want to study whether a *family* of stochastic matrices is *jointly embeddable*, in the sense that it belongs to a one-parameter continuous family $e^{\tau\mathbf{Q}}$ of stochastic matrices. However, since the matrices under consideration are estimated from data, embeddabillity is not required in a strict sense, as statistical fluctuations have to be taken into account.
We propose to consider the matrix logarithms of the matrices $\mathbf{P}^{(x,y)}$. The logarithm of a matrix is formally defined *via* an infinite power series expansion

$$\log \mathbf{P} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (\mathbf{P} - 1)^k \ . \tag{6}$$

For the sake of the present discussion, it is enough to know that the expansion converges[1] and may be computed numerically as soon as the matrix $\mathbf{P}$ is "close enough" to the identity matrix, in the following sense: there exists a positive integer n such that

$$\|(\mathbf{P} - 1)^n\|_2 < 1 \ , \tag{7}$$

where 1 is the identity matrix, and the norm $\|\mathbf{M}\|_2$ of a matrix $\mathbf{M}$ is the square root of the sum of squares of matrix elements of $\mathbf{M}$: $\|\mathbf{M}\|_2 = \sqrt{\sum_{i,j=1}^{m} M_{ij}^2}$.
Notice that the matrix logarithm does not satisfy all the usual properties of the numerical logarithm: in general, $\log \mathbf{P}_1\mathbf{P}_2 \neq \log \mathbf{P}_1 + \log \mathbf{P}_2$. However, the relation

$$\log \mathbf{P}^{\tau} = \tau \log \mathbf{P}$$

is preserved.
This suggests the following definition.

Definition 1 *If two sequences x, y in a pairwise alignment are such that the corresponding matrices $\mathbf{P}^{(x,y)}$ and $\mathbf{P}^{(y,x)}$ admit a logarithm*

$$\mathbf{L}^{(x,y)} = \log \mathbf{P}^{(x,y)} \ , \qquad \mathbf{L}^{(y,x)} = \log \mathbf{P}^{(y,x)} \tag{8}$$

we will say that x and y are sufficiently related. *Then $\mathbf{L}^{(x,y)}$ and $\mathbf{L}^{(y,x)}$ satisfy*

$$\mathbf{P}^{(x,y)} = e^{\mathbf{L}^{(x,y)}} \ , \qquad \mathbf{P}^{(y,x)} = e^{\mathbf{L}^{(y,x)}} \ . \tag{9}$$

According to condition (7), a sufficient condition for x and y to be sufficiently related is that both $\mathbf{P}^{(x,y)}$ and $\mathbf{P}^{(y,x)}$ are "close enough to the identity matrix": there exists $n \in \mathbb{Z}^+$ such that

$$\|(\mathbf{P}^{(x,y)} - 1)^n\|_2 < 1 \ . \tag{10}$$

[1] The series (6) corresponds to well-known expansion for the logarithm of a number. It should be noticed, however, that the numerical series for $\log(x)$ cannot be absolutely convergent if $|x - 1| > 1$, while the matrix series (6) can very well converge if $\|\mathbf{P} - 1\|_2 > 1$. This is because $|xy| = |x||y|$ for numbers, while in the case of matrices, $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_2$, the equality being an exception.

In terms of the sequences x and y under consideration, this simply means that the two sequences have not diverged too much.

An important property is that whenever the matrix logarithms $\mathbf{L}^{(x,y)}$ are well defined, they are by construction pseudo rate matrices, in the sense of Appendix A. However, they need not be rate matrices, as their off-diagonal elements are not necessarily non-negative.

REMARK 2 Notice that the estimation procedure we use here is non-symetric: the matrices $\mathbf{C}^{(x,y)}$ and $\mathbf{C}^{(y,x)}$ constructed as in (1) need not be equal. This choice is justified by the fact that these two matrices are in some cases significantly different, which gives useful informations about the data under consideration. This is specially true in situations where the amino acid compositions of the sequences x and y are significantly different: in such cases, one could hardly justify symmetrization.
Let us stress that this does by no means suppose that one of the sequences is an ancestor of the other one. As we shall see later, a symmetrized counting procedure is fully justified in a framework of parameter estimation for *single generator reversible* Markov chain models, but this is no longer true as soon as one expects departures from such models (departure from reversibility, or from the "single generator" assumption, or both...).

The comparison of the matrices $\mathbf{L}^{(x,y)}$ and $\mathbf{L}^{(y,x)}$ for a given pair (x,y) is an interesting issue, as it may emphasize significant departure from simple models. However, when different pairs are to be compared, one ends up with four *a priori* different matrices. It is generally simpler in such situations to consider the average matrices $\overline{\mathbf{L}}^{(x,y)}$, defined by

$$\overline{\mathbf{L}}^{(x,y)} = \frac{1}{2}\left(\mathbf{L}^{(x,y)} + \mathbf{L}^{(y,x)}\right) \ . \tag{11}$$

This has the advantage of associating a single matrix to each alignment, and simplifying the analysis. This is the choice we have made in the numerical results presented in this paper.A more systematic analysis of quartets of matrices associated to pairs of alignments will be described elsewhere.

2.2 Symmetrized counts

It is a common practice (see for example (Dayhoff et al., 1972) or (Müller and Vingron, 2000)) to use symmetrical count matrices: the matrices $\mathbf{F}^{(x,y)}$ are then constrained to be symmetric. Such a property, which may be justified theoretically in the framework of *reversible* Markov chains models, and practically for some specific families of sequences, has also the advantage of simplifying considerably the numerical work. We briefly sketch here the main modifications needed for "symmetrization", and generically use matrices with "tilda" (i.e. $\tilde{M}$ instead of M) to distinguish the symmetrized versions.
The $\mathbf{F}^{(x,y)}$ matrices are replaced with $\tilde{\mathbf{F}}^{(x,y)}$, defined by

$$\tilde{F}_{ij}^{(x,y)} = \frac{\#\{k : x(k) = i \text{ and } y(k) = j, \text{ or } x(k) = j \text{ and } y(k) = i\}}{2c^{(x,y)}} \ , \tag{12}$$

where $c^{(x,y)}$ is again the length of the alignment (x,y), and the symbol $\#S$ stands for the cardinality of the set S. Symmetrized amino acid frequencies may also be introduced, by defining

$$\tilde{\pi}_i^{(x,y)} = \frac{\#\{k : x(k) = i \text{ or } y(k) = i\}}{2c^{(x,y)}} \ , \tag{13}$$

and introducing the matrix $\tilde{\Pi}^{(x,y)} = \mathrm{diag}(\tilde{\pi}_i^{(x,y)})$, i.e. the matrix whose non-diagonal elements vanish, and whose diagonal coincide with the frequencies $\tilde{\pi}_1^{(x,y)}, \ldots \tilde{\pi}_m^{(x,y)}$. The corresponding $\mathbf{P}$ and $\mathbf{L}$ matrices read

$$\tilde{\mathbf{P}}^{(x,y)} = (\tilde{\Pi}^{(x,y)})^{-1}\tilde{\mathbf{F}}^{(x,y)} \ , \qquad \tilde{\mathbf{L}}^{(x,y)} = \log \tilde{\mathbf{P}}^{(x,y)} \ . \tag{14}$$

As stressed before, such a symmetrized version is more natural if one thinks in terms of *reversible* Markov evolution. It is also fairly interesting from a more mathematical point of view, since it may be shown that the $\tilde{\mathbf{P}}$ matrices are symmetrizable, which makes most practical issues much simpler. Those aspects are described in some details in Appendix B.1.
We shall limit our analysis here to pairs of sequences which satisfy the following "closeness" condition:

DEFINITION 2 *Two sequences (x, y) are* sufficiently close *when the corresponding matrix $\tilde{\mathbf{F}}^{(x,y)}$ is positive definite, i.e. has positive eigenvalues (we recall that $\tilde{\mathbf{F}}^{(x,y)}$ is symmetric by construction).*

REMARK 3 Roughly speaking a positive definite matrix is a symmetric matrix whose diagonal elements are positive and "dominant". In particular, if $\tilde{\mathbf{F}}^{(x,y)}$ is positive definite, the matrix $\tilde{\Pi}^{(x,y)}$ is non-singular and $\tilde{\mathbf{P}}^{(x,y)}$ is well-defined.

It follows from the discussion in Appendix B.1 that given two sufficiently close sequences, the logarithms $\tilde{\mathbf{L}}^{(x,y)}$ and $\tilde{\mathbf{L}}^{(y,x)}$ of $\tilde{\mathbf{P}}^{(x,y)}$ and $\tilde{\mathbf{P}}^{(y,x)}$ are well defined thanks to equation (24), and then x and y are sufficiently related. Also, the corresponding $\tilde{\mathbf{L}}$ matrices are pseudo rate matrices, as defined in Appendix A.

Given two sufficiently close sequences, the corresponding matrix $\mathbf{P}^{(x,y)}$ is generally not embeddable. However, the following remarkable "weak embeddability" result states that given a matrix $\tilde{\mathbf{P}} = \tilde{\mathbf{P}}^{(x,y)}$, its powers $\tilde{\mathbf{P}}^{\tau}$ for τ large enough are stochastic matrices:

COROLLARY 1 *Let x, y be two sufficiently close sequences. There exists a real number τ_0 such that for all $\tau \geq \tau_0$, the matrix $(\tilde{\mathbf{P}}^{(x,y)})^{\tau}$ is a stochastic matrix.*

This result is an immediate consequence of PROPOSITION 1, which is proved in Appendix B.2. It has important practical implications when it comes to comparing different matrices $\tilde{\mathbf{P}}^{(x,y)}$, as we shall see in Section 2.4.

REMARK 4 This result is valid for the matrices $\tilde{\mathbf{P}}^{(x,y)}$ obtained by the symmetrized counting procedure. Our numerical results suggest a similar behavior in the case of the matrices $\mathbf{P}^{(x,y)}$. We do not have a proof in the latter situation.

2.3 Quantities proportional to "divergence"

We now address the problem of comparing *sufficiently related sequences*, and therefore the corresponding matrices $\mathbf{P}^{(x,y)}$ and $\mathbf{L}^{(x,y)}$. We consider a set of p pairwise alignments (x, y) of sufficiently related sequences (in the sense of DEFINITION 1). To each pair (x, y) of sequences is associated a family of matrices $\mathbf{F}^{(x,y)}$, $\mathbf{P}^{(x,y)}$, $\mathbf{L}^{(x,y)}, \ldots$, which we would like to compare. We shall focus in particular on the $\mathbf{L}^{(x,y)}$ matrices. The simplest models suggest that the matrices $\mathbf{L}^{(x,y)}$ should be multiple of a unique matrix $\mathbf{Q}$ (up to fluctuations). $\mathbf{Q}$ is a pseudo-rate matrix, and need not be a rate matrix (the definitions of rate and pseudo rate matrices are given in Appendix A; sufficient conditions for the $\mathbf{L}^{(x,y)}$ matrices to be rate matrices are discussed in Appendix B.2).
The analysis of the family of matrices $\mathbf{L}^{(x,y)}$ provides a simple way to test such simple models. The matrices $\mathbf{L}^{(x,y)}$ may in fact be viewed as vectors in an $m^2 = 400$-dimensional space (actually a subspace

of smaller dimension if the properties of $\mathbf{L}^{(x,y)}$ are taken into account). The parameters, i.e. the family of "ages" $\tau_{(x,y)}$ of the alignments and the rate matrix $\mathbf{Q}$ may be estimated using linear regression (see also (Müller and Vingron, 2000), where a macimum likelihood estimation procedure is discussed). Let us denote globally by Θ the parameter set (the ages and the rate matrice). The problem

$$\min_{\Theta} \sum_{(x,y)} \|\mathbf{L}^{(x,y)} - \tau_{(x,y)}\mathbf{Q}\|_2^2$$

yields the equation

$$\mathbf{Q} = \frac{1}{\|\mathbf{Q}\|_2^2 \sum_{(x,y)} \tau_{(x,y)}^2} \sum_{(x,y)} \langle \mathbf{Q}, \mathbf{L}^{(x,y)} \rangle \mathbf{L}^{(x,y)} \ , \tag{15}$$

where the norm $\|.\|_2$ and the scalar product $\langle .,. \rangle$ in the space of matrices are defined by

$$\langle \mathbf{M}, \mathbf{M}' \rangle = \sum_{i,j} M_{ij} M'_{ij} \ , \quad \|\mathbf{M}\| = \sqrt{\sum_{i,j} M_{ij}^2} \ .$$

The solution is not unique (we recall that the ages $\tau_{(x,y)}$ and the rate matrix $\mathbf{Q}$ are defined up to a multiplicative constant). Equation (15) states that $\mathbf{Q}$ is an eigenvector of the linear mapping

$$\mathbf{M} \to \sum_{(x,y)} \langle \mathbf{M}, \mathbf{L}^{(x,y)} \rangle \mathbf{L}^{(x,y)} \ ,$$

with eigenvalue (the top eigenvalue in fact) $\|\mathbf{Q}\|^2 \sum_{(x,y)} \tau_{(x,y)}^2$. We shall come back to that linear mapping later on.

In the data that we have analyzed, the $\mathbf{L}^{(x,y)}$ matrices turn out to lie essentially within a subspace of much smaller dimension. The latter property is clearly seen from a principal component analysis of the set of matrices. We shall also see that the traces of the $\mathbf{L}^{(x,y)}$ matrices yield information close to the *LogDet* distance, i.e. information relative to divergence times of sequences.

2.3.1 Principal component analysis of a family of alignments.

We are interested in analyzing the position of the matrices $\mathbf{L}^{(x,y)}$ in the space they span. This may be achieved by means of a principal component analysis.

We consider a family of p alignments of related sequences, and we assume for the sake of simplicity that there are more pairwise alignments than matrix elements, i.e. $p \geq m^2$ (the case $p \leq m^2$ is handled analogously). Let K denote the "matrix of all matrices" (with p rows and m^2 columns), whose rows are the m^2 matrix coefficients of the matrices $\mathbf{L}^{(x,y)}$ of the alignments. The principal component decomposition of K reads

$$K = U \Sigma V_T \ , \tag{16}$$

where $\Sigma = diag(\sigma_1, \ldots \sigma_{m^2})$ is the diagonal $m^2 \times m^2$ matrix of the singular values of K, sorted in decreasing order. We assume that the singular values σ_β are multiplicity free. V is an orthogonal $m^2 \times m^2$ matrix, whose columns

$$\mathbf{v}_\beta = (V_{1\beta}, V_{2\beta} \ldots V_{m^2\beta})_T$$

form an orthonormal basis of the spaces of the $\mathbf{L}^{(x,y)}$ matrices. The vectors

$$\mathbf{u}_\beta = \frac{1}{\sigma_\beta} K \mathbf{v}_\beta \ , \beta = 1, \ldots m^2$$

form an orthonormal basis of an m^2-dimensional subspace of the p-dimensional space of alignments. We notice that the p rows $(U\Sigma)_{a\beta}$ of the matrix $U\Sigma$ represent the m^2 coordinates of the alignments $a = 1, \ldots p$ in the coordinate system provided by the vectors $\mathbf{v}_\beta$.

REMARK 5 The singular value decomposition is performed by diagonalizing the covariance matrix $\mathcal{C} = K_T K$. However, the latter is nothing but the matrix appearing in Equation (15), which may be written as

$$\mathbf{Q} = \frac{1}{\|\mathbf{Q}\|^2 \sum_{(x,y)} \tau^2_{(x,y)}} \mathcal{C}\mathbf{Q} \ .$$

When $p \leq m^2$ (which is the case in the example analyzed in section 3 below), the singular values decomposition of Σ may be performed as well, with slight modifications in the interpretation is the resulting matrices and vectors. In that case, the p matrices $\mathbf{L}^{(x,y)}$ span a p-dimensional subspace of the m^2-dimensional space of pseudo rate matrices. Only p singular values σ_β out of the m^2 are nonzero, and only the corresponding first p vectors $\mathbf{u}_\beta$ and $\mathbf{v}_\beta$ are of interest.
In most situations we have encountered the singular values σ_β have fast decay, so that only a few principal directions (i.e. a few vectors $\mathbf{v}_\beta$) are necessary to account for the family of matrices.
Owing to Remark 5, the first principal direction, i.e. the vector $\mathbf{v}_1$ associated with the top principal value σ_1 has a special status: in a reversible Markov model, the projection of an alignment (x, y) (a point in the p-dimensional space) onto this axis provides a measure of the divergence between the sequences in the alignment; in other words, it measures the "age" of the considered alignment. As such, it bears some similarities with the so-called *LogDet* distance, which we describe below. The projections onto the other principal directions measure the "dispersion".

2.3.2 LogDet distance.

The *LogDet* distance has been thoroughly studied and used in the literature as a measure of "evolutionary distance" between sequences. Such distances are often used as inputs for phylogenetic trees estimations. The *LogDet* distance is based upon the following simple remark: if A and B are matrices of the form $B = e^A$, then one has $\log(\det(B)) = tr(A)$, where $tr(A)$ stands for the trace of the matrix A. In particular, with the same notations as before,

$$\text{if} \quad \mathbf{P} = e^{\tau \mathbf{Q}} \ , \qquad \text{then} \quad \log(\det(\mathbf{P})) = \tau \ tr(\mathbf{Q}) = \tau \sum_{i=1}^{m} Q_{ii} \ . \tag{17}$$

The precise definition of the *LogDet* distance between two sequences x, y is slightly different, as follows (see (Lockhart et al., 1994), (Steel, 1995) and (Hillis et al., 1996) for details):

$$d_{(x,y)} = -\log\det(\mathbf{F}^{(x,y)}) = -\sum_{i=1}^{m} \log \pi_i^{(x)} - \log(\det(\mathbf{P}^{(x,y)})) \ . \tag{18}$$

Measuring the *LogDet* distance, or equivalently $\log\det \mathbf{P}^{(x,y)}$ for several pairwise alignments allows one to compare the respective "ages" of the alignments, provided they can all be described by same matrices $\mathbf{L}^{(x,y)}$ which are all multiples $\tau_{(x,y)}\mathbf{Q}$ of a single "generator" $\mathbf{Q}$, or at least by "close" rate matrices $\mathbf{Q}$: when $\mathbf{P}^{(x,y)} \approx e^{\tau_{(x,y)}\mathbf{Q}}$. In such a situation, the matrices $\mathbf{L}^{(x,y)}$ should be close to proportional to $\mathbf{Q}$, and the *LogDet* distance would provide an estimate for $\tau_{(x,y)}$.
Therefore, *LogDet* distances provide an information similar to the one carried by the projection of the matrices $\mathbf{L}^{(x,y)}$ onto the first principal axis in a principal component analysis.

2.4 Matrices associated with a continuous Markov chain model for sequence evolution

The methods described above are strongly inspired by Markov chain models. We briefly describe here the main ingredients of such models and some widely used procedures, namely maximum likelihood estimation, and *LogDet* correction. We then discuss the behavior of the matrices $\mathbf{P}^{(x,y)}$ and $\mathbf{L}^{(x,y)}$ in the framework of such models.

2.4.1 Maximum likelihood estimation; *LogDet* correction

In the simplest model, that protein sequences are assumed to consist of independent random variables, taking values in a finite state space (the $m = 20$ amino-acids), whose (identically distributed) evolutions are governed by a reversible, continuous time Markov chain. The latter is completely specified by a rate matrix $\mathbf{Q}$ (see Appendix A for definitions) and initial frequencies π_i, generally taken to be the equilibrium frequencies of the Markov chain: one writes $\mathbf{P} = e^{\mathbf{Q}}$, $\mathbf{P}^\tau = e^{\tau\mathbf{Q}}$, and the family $\mathbf{P}^\tau$ ($\tau \geq 0$) has a limit $\mathbf{M}$ as $\tau \to \infty$, such that $\mathbf{M}_{ij} = \pi_j$ for all i, j. When the Markov chain is reversible, the matrix $\mathbf{F}(\tau) = \Pi\mathbf{P}^\tau$ is symmetric for all τ.

Given a multiple alignment, it is customary to model its evolution using a binary tree, whose leaves are the present-day sequences, whose vertices (nodes) represent ancestors, and whose edges represent Markov evolution. The parameters of the model (namely, the transition matrices and the topology of the tree) are then estimated numerically using maximum likelihood (Felsenstein, 1981) or Bayesian methods (see e.g. (Durbin et al., 1998)). In the simplest situation, all transition matrices are supposed to be of the form $e^{\tau\mathbf{Q}}$, where τ is the edge length (divergence time) and where $\mathbf{Q}$ is a "universal" rate matrix (the generator of the model). The distribution of the estimators (the generator $\mathbf{Q}$ and the edge lengths τ) is estimated by bootstrap simulations, and yields indications on the significance of the results. It is important to realize that the computational burden grows exponentially with the number of sequences, and alternative strategies have to be used for very large families of sequences.

In a very similar context, the *LogDet* distance methode (Lake, 1994; Lockhart et al., 1994) provide estimates for the edge lengths. It may be shown (see (Tavaré, 1986)) that under the reversible Markov model assumptions, the *LogDet* distances provide unbiased estimators for the edge lengths, and are asymptotically normally distributed. We shall see below how distances similar to *LogDet* distances appear naturally in our context.

2.4.2 Behavior of the P and L matrices

If a reversible Markov chain is used to model a pairwise alignment (x, y) of protein sequences, one can easily show that the matrix $\mathbf{P}^{(x,y)}$ defined in (5) is a maximum likelihood estimator for the transition matrix $\mathbf{P}^\tau$ of the alignment (see for example (Lee et al., 1970) in the non reversible case; the modification for the reversible case is straightforward). If a Markov chain on a binary tree is used to describe a multiple alignment, one obtains similarly estimators for the corresponding transition matrices.

As an illustration, let us consider a simple tree associated with 3 sequences, say x, y and z and corresponding transition matrices $\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3$ and $\mathbf{P}_4$ as indicated in FIGURE 1, assumed to correspond to Markov chains. For the sake of simplicity, we also assume that the chains are reversible. It is easily seen that the transition matrix $\mathbf{P}^{(x,y)}$ corresponding to the alignment (x, y) yields an estimate for the matrix $\mathbf{P}_1\mathbf{P}_4\mathbf{P}_2$. Similarly, the transition matrix $\mathbf{P}^{(x,z)}$ corresponding to the alignment (x, z) yields an estimate

for the matrix $\mathbf{P}_1\mathbf{P}_4\mathbf{P}_3$, and the transition matrix $\mathbf{P}^{(y,z)}$ corresponding to the alignment (y,z) yields an estimate for the matrix $\mathbf{P}_2\mathbf{P}_3$.
If we assume in addition that the matrices $\mathbf{P}_i$ are of the form $\mathbf{P}_i = e^{\tau_i \mathbf{Q}}$, then $\mathbf{P}^{(x,y)}$ yields an estimate for $e^{(\tau_1+\tau_2+\tau_4)\mathbf{Q}}$, $\mathbf{P}^{(x,z)}$ yields an estimate for $e^{(\tau_1+\tau_3+\tau_4)\mathbf{Q}}$, and $\mathbf{P}^{(y,z)}$ yields an estimate for $e^{(\tau_2+\tau_3)\mathbf{Q}}$. In other words, $\mathbf{L}^{(x,y)}$ yields an estimate for $(\tau_1+\tau_2+\tau_4)\mathbf{Q}$ and so forth.
Therefore, using *Log-Det* distances on those matrices provides estimates for the times $\tau_1, \ldots \tau_4$. In addition, a more systematic (principal components) analysis of the matrices $\mathbf{L}^{(x,y)}$ provides an estimate for the generator $\mathbf{Q}$ (see below). The numerical simulations presented in the next section provide an example of parameter estimation using our approach.

In more general situations, a principal components analysis of the set of matrices $\mathbf{L}^{(x,y)}$ will also provide information on the different transition matrices $\mathbf{P}_1, \ldots \mathbf{P}_4$, but the interpretation becomes cumbersome. Sticking to our example of FIGURE 1, suppose that $\mathbf{P}_2 = e^{\mathbf{Q}_2}$ and $\mathbf{P}_3 = e^{\mathbf{Q}_3}$, with $\mathbf{Q}_2 \neq \mathbf{Q}_3$. Then $\mathbf{P}^{(y,z)}$ yields an estimate for $e^{\mathbf{Q}_2}e^{\mathbf{Q}_3}$, and $\mathbf{P}^{(z,y)}$ yields an estimate for $e^{\mathbf{Q}_3}e^{\mathbf{Q}_2}$. In general, there is no reason to expect that the matrices $\mathbf{Q}_2$ and $\mathbf{Q}_3$ commute (i.e. that $\mathbf{Q}_2\mathbf{Q}_3 = \mathbf{Q}_3\mathbf{Q}_2$), so that $\mathbf{P}^{(y,z)}$ and $\mathbf{P}^{(z,y)}$ are different. The Baker-Campbell-Hausdorff formula yields an infinite series expression for the corresponding $\mathbf{L}^{(x,y)}$ matrices:

$$\log\left(e^{\mathbf{Q}_1}e^{\mathbf{Q}_2}\right) = \mathbf{Q}_1 + \mathbf{Q}_2 + \frac{1}{2}\left(\mathbf{Q}_1\mathbf{Q}_2 - \mathbf{Q}_2\mathbf{Q}_1\right) + \ldots ,$$

but such an expression is difficult to exploit practically.[2] Nevertheless, we shall see that such an approach is sufficient to test the adequacy of the reversible Markov model, and provides useful information on the origins of the departure from the model.

2.4.3 Further Comments

It is clear that the kind of models described above can hardly be considered realistic, and most of their characteristics may be criticized. For example, site independence (i.e. the fact that all amino acids of a protein evolve independently of each other) is clearly false, as is site homogeneity (the fact that all sites have identically distributed evolution).
The simplest Markov chain models (Felsenstein, 1981) assume that the Markov evolution is governed by a unique generator (rate matrix). The SVD analysis we use can show departures from such single-generator models; in such a case, it provides qualitative indications on the number of different generators needed to describe the data (Section 3 provides a clear illustration of this fact). In any case, it provides an economic description of the alignment data
In fact, one often observes significant inhomogeneities in the amino acid compositions of sequences within a given family. Such inhomogeneities are clearly not compatible with reversible Markov chain models. The fact that our approach considers only pairwise alignments reduces the effect of composition inhomogeneity. However, when a symmetric replacement matrix is associated with any pair of sequences, the interpretation of such a matrix becomes questionable when the two sequences under consideration have significantly different amino acid compositions.

[2] Nevertheless, notice that the average rate matrices $\overline{L}$ allow one to get rid of the second order terms of the Baker-Campbell-Hausdorff formula, so that one may expect it to be less sensitive to the lack of commutation of $\mathbf{Q}_1$ and $\mathbf{Q}_2$.

3 Results

3.1 Simulations

Before discussing results obtained on real data, let us describe numerical simulations we have performed to validate the proposed approach. As was said before, it may be seen as a method for parameter estimation in the framework of Markov chain on a binary tree models. To validate this particular aspect, we have simulated 100 realizations of a family of sequences of length 5000 corresponding to the mitochondrial genome of vertebrates, using parameters (a phylogenetic tree, and a rate matrix $\mathbf{Q}$) estimated from real data (see below for details on the data). This choice was motivated by the fact that the "single rate matrix reversible" model seems fairly adequate for the sequences under consideration (see (Adachi and Hasegawa, 1992), (Adachi and Hasegawa, 1996) and the results below). For each realization, a reconstructed rate matrix was estimated with the help of the methods described above. In FIGURE 2 we display the deviations $\| Q_0 - Q_i \|_2$ where the norm $\| \ldots \|_2$ is defined in Section 2, Q_0 is the input rate matrix, and Q_i is the rate matrix coming from the i-th simulation. Notice that $\| Q_0 \|_2 = 1$ by construction. Notice that only six of the deviations are larger than 0.1.

3.2 Results on real data

3.2.1 Sequences

We have applied the method just described to the study of molecular evolution of mitochondrial DNA. We analyzed a set of proteins encoded by the mtDNA of 26 species of Metazoa (see TABLE 1 for details). This sample is representative of completely sequenced mitochondrial genomes as of December 1998. It contains the set of vertebrates studied in (Russo et al., 1996), and a few representatives of more distant phyla. The sample thus contains 11 vertebrates, 6 arthropods, (5 insects and 1 crustacean), 3 echinoderms, one annelid, 3 molluscs and 2 nematodes.

3.2.2 Alignments

Using ClustalW (Thompson et al., 1994) we have aligned the 12 mtDNA-encoded proteins that are present in all the species considered here. They are: The subunits 1, 2, 3 of cytochrome c oxidase, the subunits 1, 2, 3, 4, 4L, 5 and 6 of the NADH-ubiquinone oxidoreductase, the cytochrome b, and the subunit 6 of ATP synthase. There exists a 13th mtDNA-encoded protein in animals (the subunit 8 of ATP synthase), but it is not present in nematodes and so does not appear in the alignments considered here.
The alignments are accessible at the address:

`http://chlora.lgi.infobiogen.fr:1234/landes`

3.2.3 Replacements

As we mentioned earlier, we will be only concerned with pairwise alignments, and from now on "alignment" will mean "alignment between two sequences". After summation over the 12 proteins of two species x and y, we obtain two matrices of counts, $\mathbf{C}^{(x,y)}$ and $\mathbf{C}^{(y,x)}$. All the sites are taken into account, except the ones facing an indel. If this occurs, the corresponding site is ignored, but only for the pairwise alignment under consideration. The column of the multiple alignment is kept. The effective number of sites goes from 3275 for the pair CE_CN to 3758 for the pair PL_SP (see TABLE 1). The total length of the alignments, including indels, is 3957.

3.2.4 The matrices P, L and $\overline{\mathbf{L}}$

For the set of alignments under consideration, the matrices $\mathbf{P}^{(x,y)}$ can be calculated for all the $650 = 26*25$ pairs of sequences. This means that the alignment data satisfy the conditions (5) and (7). The percentage of identity of the pairwise alignments varies between 32.2 and 97.1. The majority of alignments are outside the "one mutation" limit of Dayhoff (Dayhoff et al., 1983; Jones et al., 1992) which requires at least 85% identity. However all the matrices $\mathbf{P}^{(x,y)}$ are sufficiently close to the identity in the sense of (7), so that their logarithms can be calculated. As we saw before, the matrices $\mathbf{L}$ (or $\overline{\mathbf{L}}$) can be viewed as points in a space of 400 dimensions. However, this cloud of points is contained to a very good approximation in a space of much lower dimension. This can be seen by a principal component analysis, described in Sec. 2.3.1. In this paper, we choosed to focus on the average rate matrices. The first 40 singular values (in decreasing order) are plotted in FIGURE 3. The first one is significantly larger than the following ones. This shows that a Markov model for protein sequences is adequate, but not very good. The matrices $\overline{\mathbf{L}}$ are close to a half-line through the origin corresponding to a single generator. We shall now see, however, that there are systematic deviations from this global conclusion.

3.3 SVD analysis of the matrices $\overline{\mathbf{L}}$

FIGURE 4 gives the projections of the $325 = 26*25/2$ matrices $\overline{\mathbf{L}}$ into the $1-2$-plane (FIGURE 4a), and the $2-3$-plane (FIGURE 4b) of the PCA. Points corresponding to alignments within a taxon are identified by specific symbols, and alignments involving two different taxons are represented by dots. So in FIGURE 4a the symbol furthest to the left corresponds to an alignment between two chordata, and more specifically between two whales $\overline{\mathbf{L}}^{(BP-BM)}$. The solid circle furthest to the right describes the alignment between two molluscs $\overline{\mathbf{L}}^{(KT-CN)}$. The dots with x-coordinate larger than 5 correspond to alignments in which one of the participants is a nematode or a mollusc. The coordinates and labels are also available at

`http://chlora.lgi.infobiogen.fr:1234/landes`

The first axis (the abscissa on FIGURE 4a) points in the direction of maximal dispersion of the cloud of points. If the evolution of the sequences were given by a continuous reversible Markov process, all the matrices $\overline{\mathbf{L}}$ would be of the form $\overline{\mathbf{L}}^{(x,y)} = \tau_{xy}\mathbf{Q}$ where τ_{xy} is proportional to the divergence between the two sequences.

Because of the predominance of the first singular value, the component along the axis 1 is almost exactly proportional to the LogDet distance, as can be seen on FIGURE 5.

FIGURE 4a shows the dominance of axis 1 over axis 2, which means, as already mentioned, that a Markov model with one generator is approximately valid for the proteins in our set. The deduced transition matrix $\mathbf{P} = e^{\mathbf{Q}}$ is accessible by anonymous ftp. It corresponds to a PAM21 matrix (i.e.$\sum_{i=1}^{m} \pi_i q_{ii} = 0.21$)(for the definition of PAM normalisation see (Dayhoff et al., 1972)). When compared with other more recently published matrices (Jones et al., 1992; Jones et al., 1994; Adachi and Hasegawa, 1996), we find that our matrix $\mathbf{P}$ is closer to that of Adachi and Hasegawa calculated by maximum likelihood methods on a set of mtDNA encoded proteins from vertebrates.

However, a closer examination of the projection on the second axis shows biologically significant deviations (FIGURE 4a). The direction of points C_C (chordates with chordates) is different from the direction of other intra-taxa points. FIGURE 4b shows the projection of the matrices into the $2-3$-plane, which is orthogonal to the first axis. Consequently, the divergence (correponding to the first axis), has been eliminated in this projection, and we only see the directions corresponding to individual rate matrices. In this projection, points near the origin on the first axis are necessarily also near the origin on the plane

$2-3$, and so are not very informative. We limit our attention to intra-taxa points, i.e. to alignments within a taxon. Neglecting the points corresponding to closely related sequences (in gray on FIGURE 4b), we observe two groups and three isolated points. The first group contains only chordate-chordate points, and the second consists of all arthropod-arthropod, echinoderm-echinoderm and nematode-nematode alignments. The 3 isolated points correspond to pairwise alignments between the three molluscs.

To conclude, the FIGURE 4b shows that the evolution of mtDNA-encoded proteins of chordates seems to correspond to a rate matrix somewhat different from that of the other taxa of our set. FIGURE 4 also clearly shows that any simple Markov model does not apply to molluscs.

These conclusions are drawn here from a small number of species, and need a re-examination based on the set of all completely sequenced metazoan mitochondrial genomes.

4 Discussion

In this paper, we have introduced a method for estimating the agreement of alignment data with the predictions of an arbitrary reversible Markov model, which need not be known in advance. The two main ingredients of the method are:

- The computation of the matrix-valued logarithms of the stochastic matrices associated to pairwise alignments.
- An appropriate principal component analysis of these logarithms, each matrix being considered as a point in a space of 400 dimensions.

The existence of the logarithm of a matrix is not guaranteed in general. We have chosen this existence as an objective criterion for accepting or rejecting an alignment. In our experience, whenever this criterion is not satisfied, there are other reasons for questioning the quality of the pairwise alignment under consideration.
To the best of our knowledge, the above points have not been made in the existing literature.
We have chosen to work with the 20-letter alphabet of amino acids. This allows us to consider deeper branchings, and to analyse simultaneously the effects of evolution and of physico-chemical properties of amino acids on replacement rate matrices. In addition, the implementation of the method with an alphabet of this size is not very difficult on present-day computers.
The example described in this paper was chosen mainly as an illustration. The set of animal mtDNA-encoded proteins analyzed here is far from complete, but it includes sequences that are sufficiently divergent to test the limits of reversible Markov models.
The fitting of the data with such a model is a reasonable first approximation. However the deviations from this model can clearly be seen on the data, and they correlate well with known phylogenetic information. Many questions remain open. At this stage, we do not have a model with not too many parameters that would fit the data. Furthermore, the use of the methods of this paper on large sets of sequences about which not much is known will require statistical tools which we do not have at present.

5 Acknowledgments

We would like to thank P. Lockhart, D. Penny, M. Steel and M. Vingron for helpful comments and discussions. We would also like to thank D. Cellier, S. Robin, C. Duby and J.J. Daudin for their suggestions about this work.

A Stochastic and rate matrices, the embedding problem

A *rate matrix* is a square $m \times m$ matrix $\mathbf{Q}$ such that $\sum_{j=1}^{m} Q_{ij} = 0$ for all i, $Q_{ii} \leq 0$ for all i, and $Q_{ij} \geq 0$ for all $i \neq j$. A square matrix which satisfies the first two properties above and fails to satisfy the third one is called a *pseudo-rate matrix*. A *stochastic matrix* (or *transition matrix*, or *Markov matrix*) is an $m \times m$ matrix $\mathbf{P}$ such that $0 \leq P_{ij} \leq 1$ for all i, j and $\sum_{j=1}^{m} P_{ij} = 1$ for all i. In general, the eigenvalues of a transition matrix are complex numbers, of modulus smaller than or equal to 1.
A Markov semigroup is a family of stochastic matrices $t \in \mathbb{R}^+ \to \mathbf{P}(t)$ satisfying the Chapman-Kolmogorov equation $\mathbf{P}(t)\mathbf{P}(t') = \mathbf{P}(t+t')$, and such that for all i, j, $P(0)_{ij} = \delta_{ij}$ and $\lim_{t\to 0} P(0)_{ii} = 1$. Given a Markov semigroup, there exists a matrix $\mathbf{Q} = \mathbf{P}'(0)$ such that $\mathbf{P}(t) = e^{t\mathbf{Q}}$, and $\mathbf{Q}$ is a rate matrix. It is well known (see e.g. (Freeman, 1967)) that if $\mathbf{Q}$ is a rate matrix, then the exponentials $\exp\{t\mathbf{Q}\}$, where $t \in \mathbb{R}^+$, form a Markov semigroup. The opposite question is interesting too: given a stochastic matrix $\mathbf{P}$, does there exist a corresponding Markov semigroup $t \to \mathbf{P}(t)$ such that $\mathbf{P} = \mathbf{P}(1)$? or equivalently, does there exist a rate matrix $\mathbf{Q}$ such that $\mathbf{P} = e^{\mathbf{Q}}$? when this is so, the matrix $\mathbf{P}$ is said to be embeddable into a Markov semigroup, or simply embeddable. Characterizations of embeddability have been given in the case of 2×2 and 3×3 stochastic matrices (see (Carette, 1995) and references therein). For example, a 2×2 stochastic matrix $\mathbf{P}$ is embeddable if and only if its determinant is positive. To our knowledge, the problem in arbitrary dimension is open.

Let $\mathbf{P}$ be a general $m \times m$ stochastic matrix, and assume that the matrix logarithm $\mathbf{L} = \log \mathbf{P}$ of $\mathbf{P}$ defined in (6) exists. Then one may write $\mathbf{P} = e^{\mathbf{L}}$, but $\mathbf{L}$ need not be a rate matrix. Letting $v = (1, 1, \ldots 1)_T \in \mathbb{R}^m$, we have $\mathbf{P}v = v$, so that $\mathbf{L}v = 0$. In other words, $\sum_{j=1}^{m} L_{ij} = 0$, and $\mathbf{L}$ is a pseudo rate matrix. However, we do not necessarily have $L_{ij} \geq 0$ for all $i \neq j$.
The analysis of the latter condition is simpler in cases where the matrix $\mathbf{P}$ corresponds to a reversible Markov chain. This is discussed below.

B Symmetrization and consequences

B.1 Some consequences of symmetrization

A simple symmetrization procedure may yield significant simplifications. We describe here the mathematical aspects of symmetrization, limiting our analysis to sequences which are sufficiently close as defined in Definition 2.
Let us then consider a pair of close sequences (x, y); for the sake of simplicity, we drop the superscript "$^{(x,y)}$" in what follows. Since $\tilde{\Pi}$ is non-singular, the matrix $\mathbf{S}$ defined by

$$\mathbf{S} = \tilde{\Pi}^{-1/2}\tilde{\mathbf{F}}\tilde{\Pi}^{-1/2} \tag{19}$$

is a symmetric matrix. In addition, we have that

$$\tilde{\mathbf{P}} = \tilde{\Pi}^{-1}\tilde{\mathbf{F}} = \tilde{\Pi}^{-1/2}\mathbf{S}\tilde{\Pi}^{1/2} \ .$$

Therefore, $\tilde{\mathbf{P}}$ is similar to the symmetric matrix $\mathbf{S}$, and has the same (real) eigenvalues. Let us denote by $\lambda_1, \lambda_2, \ldots \lambda_m$ those eigenvalues, sorted by decreasing order. Since $\tilde{\mathbf{P}}$ is the transition matrix of a reversible Markov chain, it follows that $\lambda_i \in (0, 1]$ for all i, and $\lambda_1 = 1$.
Assume that all the eigenvalues λ_i are distinct from each other. Then there exists a unique (up to a sign) orthogonal matrix $\mathbf{R}$ such that

$$\mathbf{S} = \mathbf{R}_T\Lambda\mathbf{R} \ = \mathbf{R}^{-1}\Lambda\mathbf{R}, \tag{20}$$

where $\Lambda = \text{diag}\ (\lambda_1, \dots \lambda_m)$ is the diagonal form of $\mathbf{S}$, and the subscript T denotes matrix transposition. We then obtain immediately

$$\tilde{\mathbf{P}} = \tilde{\Pi}^{-1/2} \mathbf{R}_T \Lambda \mathbf{R} \tilde{\Pi}^{1/2} \ . \tag{21}$$

The rows of the matrix $\mathbf{R}\tilde{\Pi}^{1/2}$ are the left eigenvectors of the matrix $\tilde{\mathbf{P}}$, and in particular, the first row of $\mathbf{R}\tilde{\Pi}^{1/2}$ is the vector of frequencies $(\pi_1, \dots \pi_m)$.
Once the matrices $\mathbf{R}$ and Λ are known, matrix-valued functions of $\tilde{\mathbf{P}}$ may be computed easily. For example, the square $\tilde{\mathbf{P}}^2$ of $\tilde{\mathbf{P}}$ reads

$$\tilde{\mathbf{P}}^2 = \tilde{\Pi}^{-1/2} \mathbf{R}_T \Lambda^2 \mathbf{R} \tilde{\Pi}^{1/2} \ ,$$

and more generally, for any positive real number τ, one may compute the τ-th power $\tilde{\mathbf{P}}^\tau$ of $\tilde{\mathbf{P}}$:

$$\tilde{\mathbf{P}}^\tau = \tilde{\Pi}^{-1/2} \mathbf{R}_T \Lambda^\tau \mathbf{R} \tilde{\Pi}^{1/2} \ , \tag{22}$$

where $\Lambda^\tau = \text{diag}\ (\lambda_1^\tau, \dots \lambda_m^\tau)$ is the diagonal matrix of the τ-th powers of the eigenvalues of $\tilde{\mathbf{P}}$. In components notation, we obtain

$$(\tilde{P}^\tau)_{ij} = \sqrt{\frac{\pi_j}{\pi_i}} \sum_{k=1}^{m} R_{ki} R_{kj}\, \lambda_k^\tau \ . \tag{23}$$

Another important example in what follows is the matrix-valued logarithm $\tilde{\mathbf{L}} = \log(\tilde{\mathbf{P}})$. $\tilde{\mathbf{L}}$ may be computed easily using the decomposition (21):

$$\tilde{\mathbf{L}} = \log(\tilde{\mathbf{P}}) = \tilde{\Pi}^{-1/2} \log(\mathbf{S}) \tilde{\Pi}^{1/2} = \tilde{\Pi}^{-1/2} \mathbf{R}_T \log(\Lambda) \mathbf{R} \tilde{\Pi}^{1/2} \ , \tag{24}$$

where $\log(\Lambda) = \text{diag}(\log(\lambda_1), \dots \log(\lambda_m)$ is the diagonal matrix of logarithms of eigenvalues of $\tilde{\mathbf{P}}$. The matrix $\tilde{\mathbf{L}}$ so obtained is a *pseudo rate matrix*, as defined in Appendix A

REMARK 6 Let us stress that the above diagonalizations make all the matrix calculations (logarithms, powers,...) extremely efficient, as may be seen from (24) in the case of the logarithm for example.

B.2 Remarks on rate and pseudo rate matrices in the symmetrized case

As we have seen above, the logarithm $\tilde{\mathbf{L}} = \log \tilde{\mathbf{P}}$ of the transition matrix associated to a pair of close sequences is a pseudo rate matrix; however, $\tilde{\mathbf{L}}$ is *not* a rate matrix in general, so that it does not necessarily make sense to consider matrices $\tilde{\mathbf{P}}_\tau$ as transition matrices for arbitrary positive values of τ. We now discuss that point in some details.

PROPOSITION 1 *Let $\tilde{\mathbf{P}}$ be the transition matrix associated with a pair of close sequences, with the same notations as above. Let $\tau > 0$, and let $\mathbf{M} = \tilde{\mathbf{P}}^\tau$.*

1. *The matrix element M_{ij} is positive if and only if*

$$-\sum_{k=2}^{m} R_{ki} R_{kj}\, \lambda_k^\tau < \sqrt{\pi_i \pi_j} \ . \tag{25}$$

2. *If*

$$\sum_{k=2}^{m} \lambda_k^\tau < \sqrt{\pi_i \pi_j} \ , \tag{26}$$

then $M_{ij} > 0$.

The proof is a consequence of the expression of the matrix element M_{ij} (see Eq. (23). We know that the rows of $\mathbf{R}\Pi^{1/2}$ are the left eigenvectors of P, and in particular the first row $(\mathbf{R}\Pi^{1/2})_{i1}$ is proportional to the vector of frequencies $(\pi_1, \ldots \pi_m)_T$. Therefore, the first column of $\mathbf{R}$ equals $(\sqrt{\pi_1}, \ldots, \sqrt{\pi_m})_T$, and we obtain the following expression (recall that $\lambda_1 = 1$)

$$M_{ij} = \sqrt{\frac{\pi_j}{\pi_i}} \left(\sqrt{\pi_i \pi_j} + \sum_{k=2}^{m} R_{ki} R_{kj} \, \lambda_k^\tau \right) .$$

The first part of the proposition follows directly from that expression. For the second part, we simply observe that since the columns of the $\mathbf{R}$ matrix are orthonormal, $-1 \leq R_{ik} \leq 1$ for all i, k. Therefore, $-\sum_{k=2}^{m} R_{ki} R_{kj} \, \lambda_k^\tau \leq \sum_{k=2}^{m} \lambda_k^\tau$, which proves the result.

The important consequence is that, since $\lambda_k \leq 1$ for all $k \geq 2$ (with strict inequality if the top eigenvalue λ_1 is non degenerate), the function $\tau \to \sum_{k=2}^{m} \lambda_k^\tau$ is monotonically decreasing. Hence, for τ large enough, the sufficient condition above is automatically fulfilled.

References

Adachi, J. and Hasegawa, M. (1992). Amino acid substitution of protein coded in mitochondrial DNA during mammalian evolution. *Jpn. J. of Genetics*, 67:187–197.

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42:459–468.

Barry, D. and Hartigan, J. (1987). Statistical analysis of hominoid molecular evolution. *Stat. Sci.*, 2:191–210.

Carette, P. (1995). Characterization of embeddable 3×3 stochastic matrices with a negative eigenvalue. *New York J. Math.*, 1:120–129.

Dayhoff, M., Barker, W., and Hunt, L. (1983). Establishing homologies in protein sequences. *Methods Enzymol.*, 91:524–544.

Dayhoff, M., Eck, R., and Park, C. (1972). A model of evolutionary change in proteins. *Atlas of Proteins Sequence and Structure*, 5:89–99.

Durbin, R., Eddy, S., Krogh, and A., Mitchison, G. (1998). *Biological Sequence Analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.

Freeman, D. (1967). *Markov Chains.* Holden Day Publ. Comp.

Hillis, D., Moritz, C., and Mable, B. E. (1996). *Molecular systematics.* Sinauer Associates Inc. Publishers, Sunderland, MA.

Jones, D., Taylor, W., and Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 9:275–282.

Jones, D., Taylor, W., and Thornton, J. (1994). A mutation data matrix for transmembrane proteins. *FEBS Lett.*, 339:269–275.

Lake, J. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Nat. Acad. Sci. USA*, 91:1455–1459.

Lee, T., Judge, G., and Zellner, A. (1970). *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data.* North Holland Publ. Comp. Contribution to Economic Analysis series.

Lockhart, P., Steel, M. A., Hendy, M., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, 11:605–612.

Müller, T. and Vingron (2000). Modeling amino acid replacement. *J. Comp. Biol.*, 7(6):761–776.

Russo, C. A. M., Takezaki, N., and Nei, M. (1996). Efficiencies of different genes and different trees-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.*, 13:525–536.

Steel, M. A. (1995). Reconstructing evolutionary trees under a variety of markov style models. In Tavaré, S., editor, *Proceedings of Phylogeny workshop held at Princeton University*, page 51. DIMACS Technical Report 95-48.

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

Thompson, J., Higgins, D., and Gibson, T. (1994). Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680.

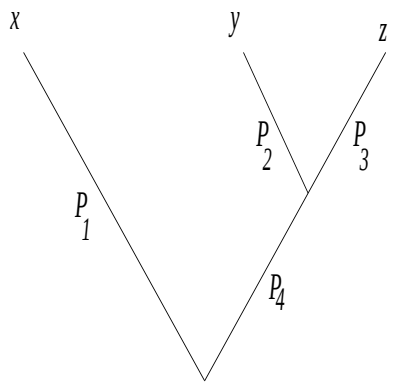

Figure 1: Example of Markov chain on a tree.

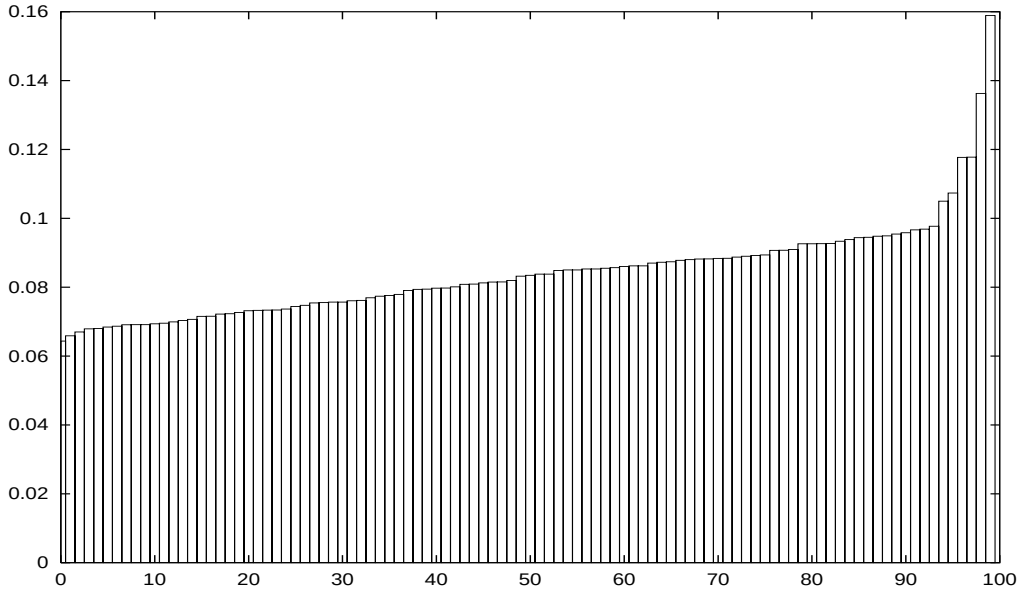

Figure 2: Distances between input rate matrix and reconstituted rate matrices

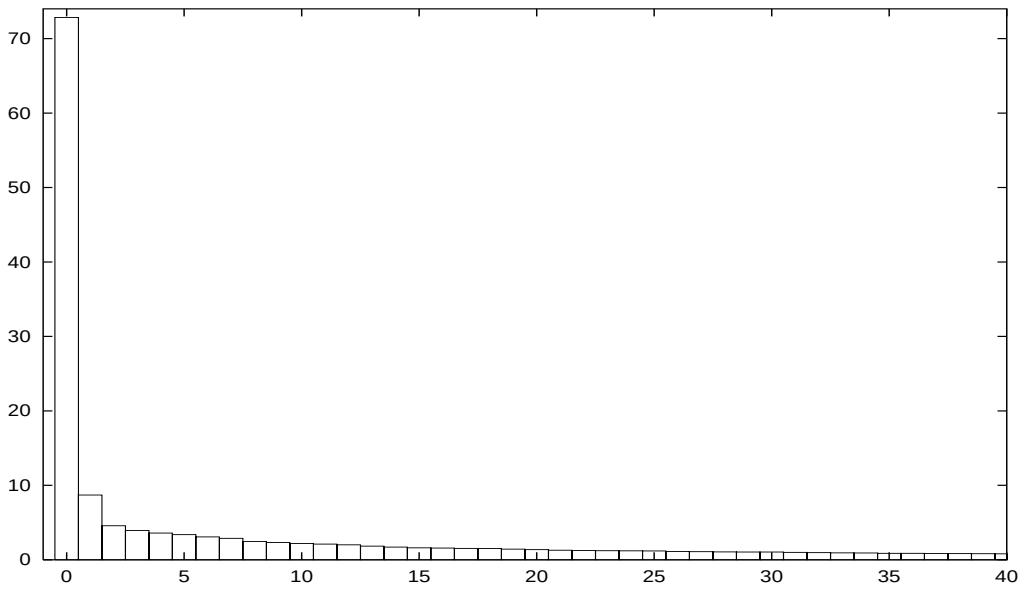

Figure 3: Singular values for the cloud of matrices $\overline{\mathbf{L}}$.

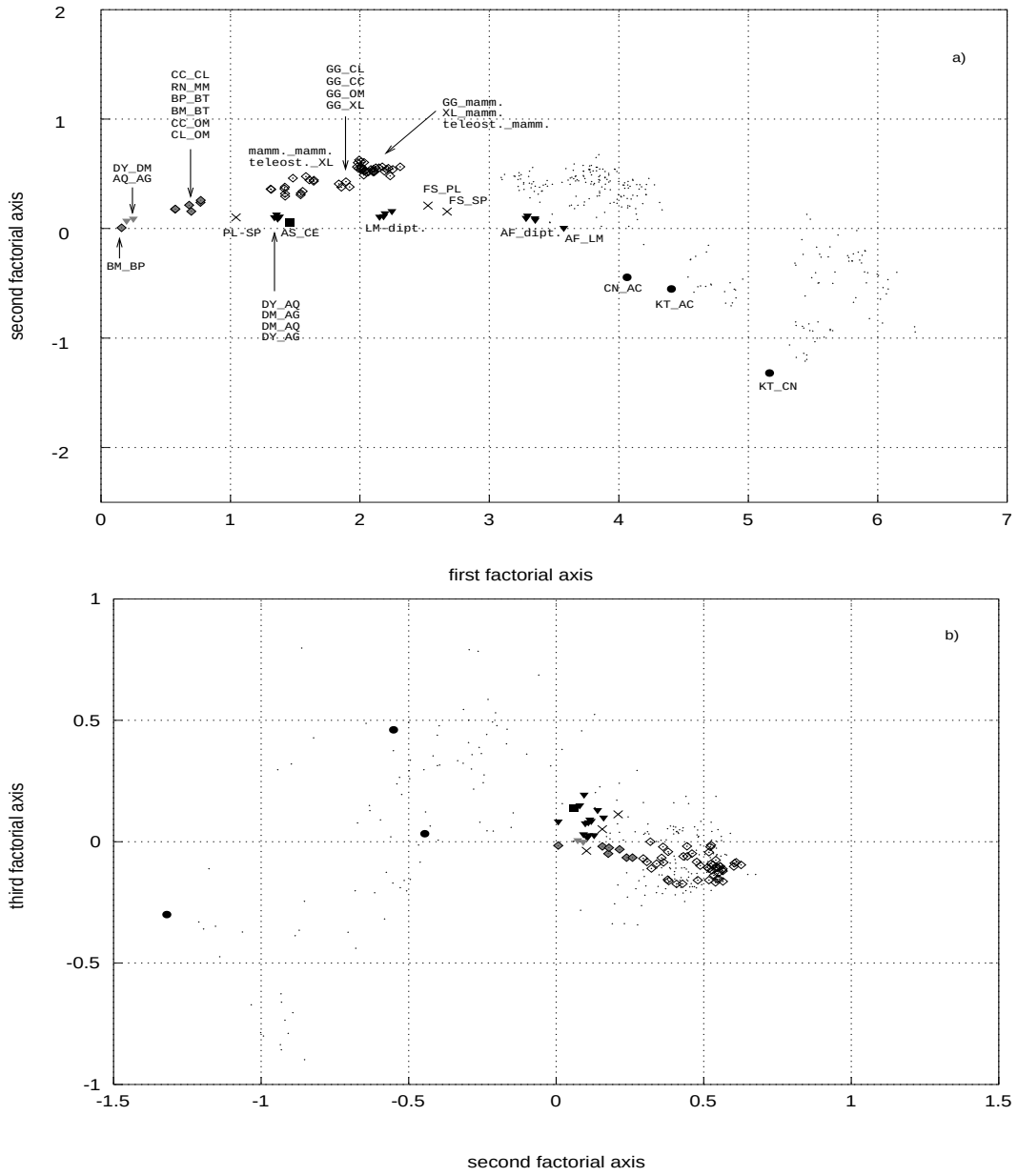

Figure 4: Projections of the cloud of matrices $\overline{\mathbf{L}}$ on the planes $1-2$ and $2-3$. Each point represents an alignment between two species. The alignments involving pairs from the same taxon (*cf.* Table 1) are represented as follows: diamond ($\diamond$) for pairs chordata-chordata, triangle (∇) for pairs arthropod-arthropod, square ($\square$) for the pair nematode-nematode and circle ($\bullet$) for the mollusc-mollusc pairs. The other points, involving alignments between two distinct taxa of Table 1 are denoted by dots. Fig. a: Projections of the matrices $\overline{L}$ on the $1-2$ plane. Abbreviations: mamm. for mammalia, dipt. for diptera, teleost. for teleostei. Fig. b: projections to the $2-3$ plane. The symbols are the same as in Fig.a; the gray ones correspond to "young" alignments (abscissa < 1.0) in Fig. a.

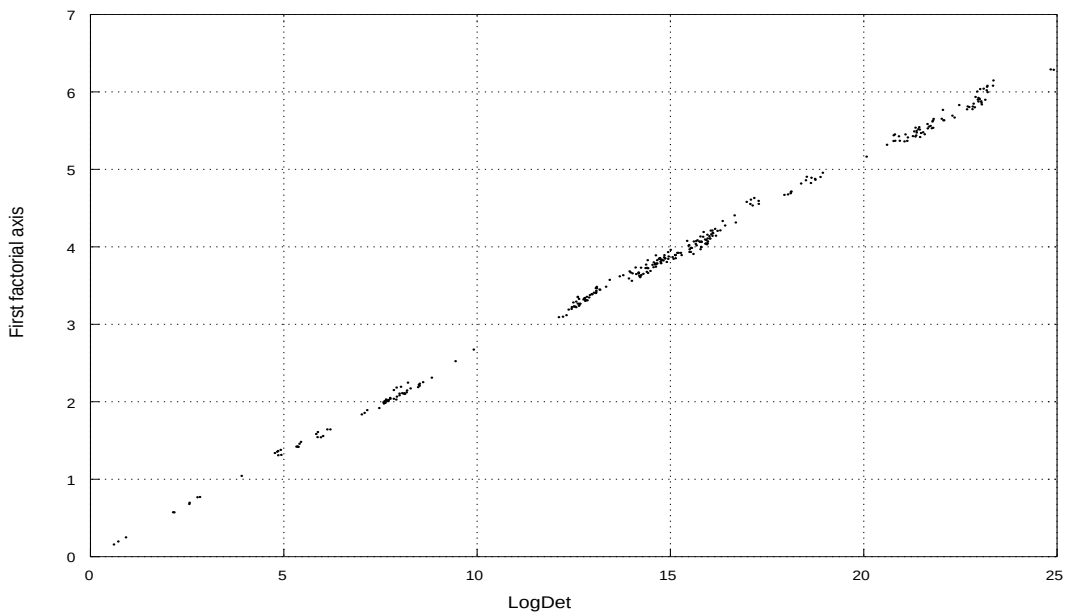

Figure 5: LogDet distance versus principal axis projection. The abscissa is $\log\det \mathbf{P} = -tr\overline{\mathbf{L}}$. The ordinate is the projection onto the first principal axis.

Abbr.[a]	Scientific name	Short name	EMBL accession number	Taxa
BT	*Bos taurus*	cow	V00654	Chordata (M)[b]
BM	*Balaenoptera musculus*	blue whale	X72204	Chordata (M)
BP	*Balaenoptera physalus*	fin whale	X61145	Chordata (M)
MM	*Mus musculus*	house mouse	J01420	Chordata (M)
RM	*Rattus norvegicus*	Norway rat	X14848	Chordata (M)
DV	*Didelphis virginiana*	North American opossum	Z29573	Chordata (M)
GG	*Gallus gallus*	chicken	X52392	Chordata
XL	*Xenopus laevis*	African clawed frog	M10217	Chordata
CL	*Crossostoma lacustre*	oriental steam loach	M91245	Chordata (T)[c]
CC	*Cyprinus carpio*	common carp	X61010	Chordata (T)
OM	*Onchorhynchus mykiss*	rainbow trout	L29771	Chordata (T)
DL	*Drosophila melanogaster*	fruit fly	U37541	Arthropoda (D)[d]
DY	*Drosophila yakuba*		X03240	Arthropoda (D)
AG	*Anopheles gambiae*	African malaria mosquito	L20934	Arthropoda (D)
AQ	*Anopheles quadrimaculatus*		L04272	Arthropoda (D)
LM	*Locusta migratoria*	migratory locust	X80245	Arthropoda
AF	*Artemia franciscana*	brine shrimps	X69067	Arthropoda
FS	*Florometra sarratissima*	crinoid florometra	AF049132	Echinodermata
SP	*Strongylocentrotus purpuratus*	purple sea urchin	X12631	Echinodermata
PL	*Paracentrotus lividus*	common urchin	J04815	Echinodermata
LT	*Lumbricus terrestris*	common earthworm	U24570	Annelida
KT	*Katharina tunicata*	black chiton	U09810	Mollusca
CN	*Cepaea nemoralis*	banded wood snail	U23045	Mollusca
AC	*Albinaria caerulea*	land snail	X83390	Mollusca
CE	*Caenorhabditis elegans*	nematode	X54252	Nematoda
AS	*Ascaris suum*	pig roundworm	X54253	Nematoda

Table 1: Set of complete mitochondrial genomes analysed.

[a] Abbr: abbreviation
[b] M: mammalia
[c] T: teleostei
[d] D: diptera