Transition rate matrices determined by families of alignments give information about evolution

C. Landès-Devauchelle, J.-L. Risler, A. Hénaut

Génome et Informatique, Université de Versailles-Saint Quentin, Bat. Buffon, 45 Avenue des Etats Unis, 78035 Versailles Cedex, France

A. Grossmann, M. Holschneider, B. Torrésani

Centre de Physique Théorique, CNRS-Luminy, case 907, 13288 Marseille Cedex 09, France

Abstract— **The procedure described in this paper takes as input a set of alignments (here the set of all pairwise alignments between members of a family of homologous proteins) which need not have a high percentage of identities. It returns simultaneous estimates for (i) One or several matrices of transition rates between amino-acids, which may be thought of as driving the evolution within the family, and (ii) the relative times of divergence between the sequences of the family.**

The method was applied to the family of proteins encoded by the mitochondrial genome of metazoa. The relative divergence times of pairs of species were in agreement with expectations. The examination of rate matrices led to the detection and correction of an inconsistency in the mitochondrial genetic code of nematodes.

Keywords— **Dayhoff model, Protein, Alignment, Mitochondrial genome of metazoa, Mitochondrial genome of nematodes, Molecular evolution, Evolutionary model, Markov chain**

I. Introduction

THE concept of "molecular clock" was introduced to help explain the constancy in time of the evolution rate of proteins. The underlying assumptions are: (i) The evolution from an ancestral to the present-day sequence occurs by repeated mutation/selection events; (ii) At every stage the sequence is modified according to conditional probabilities given by an "elementary transition matrix" M which does not depend on time ; (iii) After a time t, the initial sequence is modified with probabilities given by a matrix power M^t.

The existence of one such matrix for a given family of proteins is not *a priori* obvious, and it is desirable to have it confirmed or infirmed by data originating within the family only. The data consist typically of statistics obtained from alignments between the proteins in the family. If the existence of an elementary transition matrix (rate matrix) associated with a family is compatible with the data, then this matrix should be computed, again without the use of any input extraneous to the family. The method should also point out situations where a description involving several matrices is preferable.

C. Landès-Devauchelle's email: landes@genetique.uvsq.fr
J.L. Risler's email: risler@genetique.uvsq.fr
A. Hénaut's email: henaut@genetique.uvsq.fr
A. Grossmann's email: grossman@cpt.univ-mrs.fr
M. Holschneider's: email hols@cpt.univ-mrs.fr
B. Torrésani's email: torresan@cpt.univ-mrs.fr

It is not immediately clear how to implement such a program, since the alignment data depend both on the matrix M and on the divergence times between the proteins in the family. Neither M nor the times are known, and the latter can vary widely within a given family. It is necessary to transform the data into a form which allows the "disentangling" of the influence of M from that of the times of divergence [8].

A partial solution is to consider only alignments for which the percentage of non-identity is small. This procedure has been successfully used in the construction of "universal" score matrices for alignment programs [4], [5], [9], [10], [6]. It is however not satisfactory for the purpose at hand, as it requires large families and hides possible departures from an assumed model.

In this paper we present and apply a method for the simultaneous estimation of a rate matrix and of divergence ages for a family of proteins. The method is based on the following simple idea, which will be described in more detail in Section 3: The relevant information from any alignment is contained in a 20 by 20 transition matrix, which can be considered as a point in a space of $20\times 20 = 400$ dimensions. A family of alignments can then be imagined as a cloud of points in that space. In order to estimate whether the alignments result from the action of a single rate matrix, we perform a coordinate transformation which brings powers of any matrix to a straight one-dimensional line in the 400-dimensional space. This involves taking the matrix-valued logarithm of a matrix (not to be confused with the log-odds matrix). Such a logarithm exists if all the eigenvalues of the transition matrix associated to the alignment are positive. This is an "acceptability condition" on the alignment; it extends far beyond the "one-transition" regime, allowing over 70% of non-identity.

After this transformation, we can use standard methods of multivariate analysis for estimating and visualizing the quantities of interest [11]. For example, this allows us to decide whether the data under consideration are compatible with the assumption of a unique rate matrix, of if several rate matrices have to be introduced.

This procedure has been applied to the family of proteins coded by the mitochondrial genome of metazoa. In our opinion, it has performed well, supporting a reasonable model with four matrices of transition rates.

An unforeseen, and encouraging, development was that the procedure pointed out a specific anomaly in the transition rates for nematodes. This anomaly was then traced to a discrepancy between the genetic code used in the data bank and the code proposed in [18].

While we believe the results of this paper to be new, the ingredients we use have been around for quite a while. For the background, the reader can consult Chapter 11 of [8], and Chapter 14 of [22]. More specifically, our estimates of alignment ages are closely related to "LogDet" and "paralinear distance" methods [8], [7], [12], [1], [2], [16], [13], [21] which have been in use for several years. What we stress here is that alignments, in addition to having an "age", can be grouped according to their "direction", given by a normalized rate matrix.

Acknowledgements

This work was supported by the GREG.

II. Data and methods

A. Sequences

The mitochondrial sequences were obtained from the 23 complete animal mitochondrial genomes in release 47 (june 1996) of the EMBL nucleotide sequence database : *Balaenoptera musculus* (blue whale) (X72204), *Balaenoptera physalus* (finback whale) (X61145), *Bos taurus* (cattle) (V00654; J01394), *Halichoerus grypus* (gray seal) (X72004), *Phoca vitulina* (harbor seal) (X63726; S37044), *Homo sapiens* (human) (J01415; M12548; M58503; M63932; M63933), *Mus musculus* (house mouse) (J01420), *Rattus norvegicus* (Norway rat) (X14848), *Didelphis virginiana* (North American opossum) (Z29573), *Xenopus laevis* (African clawed frog) (M10217; X01600; X01601; X02890), *Crossostoma lacustre* (pisces cypriniformes) (M91245), *Cyprinus carpio* (common carp) (X61010), *Paracentrotus lividus* (common urchin) (J04815), *Strongylocentrotus purpuratus* (purple urchin) (X12631), *Albinaria coerulea* (land snail) (X83390), *Cepaea nemoralis* (banded wood snail) (U23045), *Katharina tunicata* (black chiton) (U09810), *Artemia franciscana* (brine shrimp) (X69067), *Apis mellifera ligustica* (honeybee) (L06178), *Drosophila yakuba* (fruit fly) (X03240; J01400; J01402; J01403; J01406; J01408; V01521; X00563), *Lumbricus terrestris* (annelid worm) (U24570), *Ascaris suum* (pig roundworm) (X54253), *Caenorhabditis elegans* (nematode) (X54252; S93745).

B. Alignments

Homologous protein sequences (extracted from SwissProt according to references cross-linked with EMBL entries) were aligned with the program CLUSTAL W [20]. The aligments are robust and almost independent of the matrix and of the penalties (PAM series, gap open penalty 15, and gap extend penalty 0.15). Twelve of the proteins are common to all species, namely ATPase subunit 6, cytochrome B, the subunits I, II et III of cytochrome C oxidase, and the subunits 1, 2, 3, 4, 4L, 5 and 6 of NADH dehydrogenase. The gene of ATPase subunit 8 has not been localized in the mitochondrial genome of the two nematodes (*Ascaris suum* and *Caenorhabditis elegans*). The total length of alignments of the thirteen proteins is 4027 amino acids (3958 without the ATPase subunit 8).

The percentage of identity in the alignments varies from 31% to 97%.

C. Transition matrix associated to an alignment

A *transition* (or Markov) matrix is a square matrix with non-negative entries, such that every one of its rows sums up to 1.

A *rate* matrix is a square matrix such that every one of its rows sums up to 0, an that all of its off-diagonal entries are non-negative.

Given an alignment of a sequence "A" and a sequence "B", in a m-letter alphabet (for proteins, $m = 20$), one constructs traditionally [4] an associated transition matrix as follows:

1. Start with a m by m matrix containing only zeros, with rows and columns indexed by the letters of the alphabet.
2. Read pairs of letters in correspondence through the alignment, ignoring pairs that contain "-". If the two letters are different, say LI, add 1 to the matrix entry corresponding to L, I, and to the entry at the position I, L. If the two letters are identical, say LL, add 2 to the diagonal entry L, L. The matrix so obtained is symmetric.
3. Divide all entries in a row by the sum of the all entries in that row, obtaining a transition matrix, which is in general not symmetric.

A transition matrix obtained through such a symmetrized reading of an alignment is not arbitrary. One can show in particular that all of its eigenvalues are real.

D. Rate matrix associated to a family of alignments

We are concerned with the following problem: Given a family of alignments, obtain estimates for the rate matrix and alignment ages that "best" imitate the observed statistics. We shall give here an intuitive geometric description of the ideas. A fuller discussion will be given in an article in preparation [14].

A basic assumption of the Dayhoff model of proteins [4] [5] is that that their evolution can be reasonably well described by a model in which mutation events at individual sites are independent of each other. In such a model, all the information provided by an alignment A/B can be reduced to (i) Counts of occurences of amino acids in the sequences A and B and (ii) Counts of occurences of pairs of amino acids facing each other in the alignment. These counts assign a transition matrix $M(A/B)$ to any given alignment A/B, as described above.

Consequently, every alignment determines a point in a 400-dimensional space. The coordinate of that point along the axis numbered $(20(i-1)+j)$ is the number $M_{ij}(A/B)$, which is the probability that, if the amino acid i is found at a site in the alignment, the amino acid facing it will be j.

A family of alignments can then be represented as a cloud of points in 400-dimensional space. The questions are then

1. whether the action of a single transition rate matrix gives a reasonable statistical description of this cloud and
2. if the answer to the first question is 'yes', how to compute this rate matrix.

In order to understand the solution, it will be useful to consider first the case of very long alignments, where the statistical fluctuations can be neglected. Consider the action of a rate matrix Q. An alignment generated by this matrix at time t_α will give rise to a transition matrix

$$M_\alpha = e^{t_\alpha Q} \tag{1}$$

If one considers the times $t_1, \ldots, t_\alpha, \ldots, t_n$, the cloud of points M_α will have a complicated shape. However, if we introduce the points corresponding to the matrices

$$L_\alpha = \log(M_\alpha) = (M_\alpha - 1) - \frac{1}{2}(M_\alpha - 1)^2 + \ldots = t_\alpha Q \tag{2}$$

we see that these points all lie on the line consisting of all multiples of Q.

If we consider now the case of alignments of realistic length obtained by the action of a single rate matrix Q, the points $M_\alpha = M(A/B)$ will be affected by statistical fluctuations, but we can still expect them to form a cloud along the line of multiples of Q. The extent to which this is true can be tested by standard methods. One introduces the covariance matrix

$$V_{ij,kl} = \sum_\alpha (\log M_\alpha)_{ij} (\log M_\alpha))_{kl} \tag{3}$$

The matrix V is positive definite and 400 by 400 . The matrix elements of the estimated generator $\hat{Q}$ are the components of the eigenvector of V that corresponds to the largest eigenvalue of V. The divergence time of an alignment A/B is

$$t(A/B) = \sum_{i,j} (\log M(A/B))_{ij} \hat{Q}_{ij} \tag{4}$$

From (1) one can deduce that the logarithm of the determinant of M_α is equal to the product of t_α and of the sum of the diagonal elements of Q. This is a connection with the "LogDet" methods, mentioned in the Introduction. The present method provides estimates not only for ages of the alignments, but for one or several rate matrices that drive the evolution.

The calculations described in this section have been implemented as programs in C, and are available by anonymous ftp from ftp.infobiogen.fr.

III. Results

We consider here the proteins encoded by the mitochondrial genome of metazoa. As described above, the data consists of 23 genomes, which give rise to 253 pairwise alignments. For each pair of genomes, we have aligned the 13 homologous proteins (except for the nematodes which have only 12 of them) and cumulated the counts into a single replacement matrix. This procedure is justified by the monophyletic origin of the mitochondria and by the absence of genetic recombination. In addition, the 13 proteins are related in function, since they they are part of the oxidative phosphorylation chain. The mitochondrial genome may consequently be considered as a single polypeptide with 13 domains.

We have first applied the procedure described above to the set of all 253 alignments. Images a and b in Figure 1 show the projections onto the plane of axes 1,2 and 1,3, respectively. The direction of the first axis is the generator of evolution, and the first coordinate of any point is proportional to the age of the corresponding alignment. So "young" alignments are on the left, and the "old" ones on the right.

The cloud of points of proteins coded by mitochondrial genome is essentially localized near the first axis. More precisely, 35% of the variance is supported by the first axis, 5% by the second axis and 4% by the third axis. The direction of the first axis gives the average generator of the family, and the coordinate along this axis gives the divergence time of the pair of sequences -where *time* has to be understood as a *relative time* measuring a number of elementary evolutions (more precisely a power of a transition matrix),- and not as an astronomical time.

Each point represents the matrix L_α of a single alignment (see Equ. (1) and Equ. (2)), and involves two species. For the sake of simplicity we have chosen only one of the two species to characterize the alignment (see the legend of Figure 1 for details). The groups which may be distinguished along the the time axis, and the order in which they appear, confirm our present knowledge about the evolution of those species: vertebrates, urchins, other invertebrates, honey bee, snails and nematodes. The anomalous position of the honey bee on the time axis, in particular with respect to the fruit fly has already been reported in [3]. We will briefly come back to that point in the conclusion.

The sizes of the various groups which can be observed along the time axis shows how the dispersion increases with time. The vertebrates are easily aligned on a straight line going through the origin, while the cloud corresponding to *C. Elegans* is wider. However, it is impossible to argue that *Caenorhabditis* may be aligned on the same line as the vertebrates, even by invoking statistical fluctuations. Therefore, we are led to the conclusion that there exist several different generators in the family of proteins coded by mitochondrial genome.

The alignments involving a given species may be seen empirically to lie within a cone of aperture $\theta_a \approx 10°$. The value of the aperture angle θ_a seems to be the same for vertebrates as well as for all nematodes. For the family under consideration, the points may be aligned along 5 different lines (up to statistical fluctuations), corresponding to 5 different generators may be distinguished: one for vertebrates, one for the two snails, one for each nematode (surprisingly enough) and one for invertebrates (chiton, honey bee and other invertebrates in Figure 1.

In order to determine those generators more precisely, we have performed again the analysis by considering only the alignments corresponding to a unique generator (following the convention given in the legend of Figure 1). By doing so we observe a significant increase of the percentage of variance on the first axis: 52% for the vertebrates, 53% for the snails, 62% for Ascaris, 60% for Caenorhabditis and 45% for the other invertebrates.

The next question is to know where the differences between generators do lie: does a given evolutionary branch differ only by a few matrix coefficients, by all the coefficients corresponding to a given (group of) amino acid(s), or by a global change of all the matrix coefficients ? In order to compare the generators, we have normalized them in such a way that $M = e^Q$ corresponds to 1% of mutations. Then we have compared the transition matrices PAM120, i.e. $M^{120} = e^{120Q}$ (we have chosen a high power of the elementary matrix so as to emphasize the differences.)

The two matrices corresponding to nematodes are characterized by a higher mutability of arginine. A closer look at the amino acids replacing arginine preferentially shows that the prefered amino acid is serine (see Figure 2). Taking into accound the significant difference of the lateral chains of these two amino acids, such prefered replacements can hardly be explained on the basis of physico-chemical arguments. Therefore, we have looked for an origin in the genetic code. Indeed, arginine is coded by the codons CGN or AGR in the standard code, while codon AGY codes for serine. However in the case of mitochondria, the codons AGR may code for arginine (ciliata, yeast), serine (echinodermata,fruit fly) or even be absent (vertebrates). It is therefore reasonable to verify that the fairly large number of replacements R-S observed in the alignments involving nematodes can be attributed to wrong assignments to the codons AGR. Indeed, the proteins of nematodes extracted from SwissProt give arginine when the gene has a codon AGR, while the original article [18] and the CDS translated from EMBL entries indicate a serine.

We have verified that the strange location of the two clouds of nematodes *C. elegans* and *A. suum* in Figure 1 is due to this discrepancy in the genetic code. To do so, we have translated the mitochondrial genes of nematodes using the code indicated in [18]. The results (see Figure 3) show that *C. elegans* and *A. suum* are now within the cone of aperture angle $\theta_a = 10°$ alluded to above, and therefore are likely to follow the same evolution generator Q.

However, this generator is still different from that of vertebrates, snails or other invertebrates. The erroneous transition R-S was responsible for the misleading location of *Caenorhabditis* in Figure 1, which is now in Figure 3 at the same level as *Ascaris* (the location of the latter, which comprises very few codons AGR in its genes, is only weakly affected by the modification).

The family of proteins coded by mitochondrial genome we have studied does not follow an evolution given by a single generator; at least 4 different generators seem to be necessary to describe them with the formalism given in Section II: (i): vertebrates; (ii): snails; (iii): nematodes; and (iv): other invertebrates. The differences between those 4 generators do not involve a few particular transitions any more, but are scattered over a large number of matrix elements.

In the case of vertebrates, cysteine has a peculiar behavior: it is significantly more conserved than in the other groups (see Figure 4.) A closer look at under-represented transitions for vertebrates does not exhibit any specific feature, except for slightly smaller mutation probabilities to hydrophobic amino acids (F, I, L, M, V) and alanine.

IV. Discussion

The evolution model which underlies the present work is the Dayhoff's model [4], [5], which implies the following implicit assumptions: (i): independence of sites within a protein; (ii): Independence of alignments within a family; (iii): the rate matrix which does not vary with time; (iv): the rate matrix does not vary within the considered family. The first two assumptions are obviously incorrect, but present the advantage of simplifying considerably the estimation. The other two are at the very basis of molecular clock.

The method we just presented and the corresponding results show that we can estimate the matrix corresponding to a given family, and decide whether assumption (iv) is compatible with data. This was made possible by the disentangling of the estimations of the rate matrix and the time. Previous works were limited by the lack of large enough the amount of data, which forced to consider very large families for estimating transition matrix or to use a general matrix for estimating times.

The limitation of our approach is that we must have at our disposal a correct multiple alignment, and enough alignments such that the estimation of the corresponding rate matrix L_α is admissible (namely that the eigenvalues of M_α are all non-negative). However, numerical simulations (which will be presented in a companion paper in preparation [14]) show that the matrices L_α may be reliably estimated for alignments with percentages of differences ranging from 5% to 70%. This makes the study feasible for quite a large number of protein families. As an example, simulations show that a family of 20 sequences or average length 500 amino acids yields a good estimation of the corresponding rate matrix and times.

Let us stress that it is necessary to be extremely careful with statistical fluctuations when it comes to interpreting the results. In particular, the variance of the time is so large that it is hard to distinguish the alignments mammalians-amphibians from the alignments mammalians-fishs in the fourth cloud in the first axis of Figure 1 and 3.

Let us also stress that when there are several rate matrices within a family, the corresponding times are not necessarily comparable. Indeed, they are expressed in selection/mutation units, and those units are matrix-dependent.

V. Conclusions

We have presented and used a method for estimating simultaneously relative divergence times and transition matrices from a single multiple alignment, without any extra information. We have shown that such a technique allows one to test whether a single evolution generator or several ones have to be associated with a given family.

The methods presented here have many potential applications, and can in particular be useful as a first step in different problems. For instance, the divergence times may be used as input for constructing phylogenetic trees. It is also reasonable to expect that such trees may be improved by incorporating the extra information provided by the rate matrices, at least in situations where several such matrices are to be associated with the family of proteins under consideration.

This work has to be continued in several directions. A more detailed treatment of the theoretical issues will be given in [14], together with a systematic statistical study of the estimators we described. An important issue concerns the symmetry of the estimated replacement matrices $M(A/B)$ given in Section II-D. It may be shown that such symmetric matrices can yield unbiased estimators only when statistical equilibrium is attained, i.e. when the frequencies of amino acids do not depend on time. Such an implicit assumption, even though it is of standard practice, is far from innocuous, and is certainly incorrect in many situations. For example, it is likely that the anomalous behavior of mitochondrial sequences of honey bee we pointed out are connected to the fact that they have a significantly different amino acid composition, which is inconsistent with the equilibrium assumption. The methods of this paper have to be modified to account for possible changes of amino acid composition during evolution. This is work in progress.

References

[1] Adachi, J. & Hasagawa M. (1992). *MOLPHY : Programs for molecular phylogenetics I - PROTML : Maximum likelihood inference of protein phylogeny.* (Computer Science Monographs, **27**, Institute of Statistical Mathematics, Tokyo).

[2] Cao, Y., J. Adachi J. , Yano T. & Hasegawa M. (1994). *Mol. Biol. Evol.* **11**, 593-604.

[3] Crozier, R.H., and Crozier, Y.C. (1993). *Genetics* **130**, 471-498.

[4] Dayhoff, M.O., Eck, R.V. & Park, C.M. (1972). *Atlas of Protein Sequence and Structure*, vol. 5, pp. 89-99, (National Biomedical Research Foundation, Washington, DC).

[5] Dayhoff, M.O., Barker, W.C. & Hunt, L.T. (1983). *Methods Enzymol.* **91**, 524-544.

[6] Gonnet, G.H., Cohen, M.A. & Benner, S.A. (1992). *Science* **256** 1443-1445.

[7] Hasegawa, M., Kishino, H. & Yano, T. (1985). *J. Mol. Evol.* **22**, 160-174.

[8] Hillis, D.M., Moritz, C.M. & Mable, B.K. (1996). *Molecular Systematics, Second Edition* (Sinauer Associates, Inc. Sunderland, Massachusetts U.S.A.).

[9] Jones, D.T., Taylor, W.R. & Thornton, J. M. (1992). *CABIOS* **8**, 275-282.

[10] Jones, D.T., Taylor, W.R. & Thornton, J.M. (1994). *FEBS Letters* **339**, 269-275.

[11] Kendall M. (1980). *A course in multivariate analysis* (Charles Griffin & Company Limited, London & High Wycombe)

[12] Kishino, H., Miyata T. & Hasegawa, M. (1990). *J. Mol. Evol.* **31**, 151-160.

[13] Lake, J.A. (1994). *Proc. Natl. Acad. Sci. USA* **91**, 1455-1459.

[14] Landès-Devauchelle, C., Hénaut, A., Risler, J.L., Grossmann, A., Holschneider, M. & Torrésani, B. (1997). In preparation.

[15] Lee, T.C., Judge, G.G. & Zellner, A. (1970). *Contribution to Economic Analysis series* (North Holland Publ. Comp.)

[16] Lockhart, P.J., Steel M.A., Hendy M.D. & Penny, D. (1994). *Mol. Biol. Evol.* **11**, 605-612.

[17] Okimoto R., Macfarlane, J.L. & Wolstenholme, D.R (1990). *Nucleic Acid Res.* **18**, 6113-6118.

[18] Okimoto R., Macfarlane, J.L. & Wolstenholme, D.R (1992). *Genetics* **130**, 471-498.

[19] Press W.H, Flannery B.P., Teukolsky, S.A.& Vetterling, W. T.(1994). *Numerical Recipes in C* (2d edition, Cambridge University Press, Cambridge).

[20] Thompson, J.D., Higgins, D.G & Gibson, T.J. (1994). *Nucleic Acids Res.*, **22**, 4673-4680.

[21] Waddell, P.J. (1995). *Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms, and maximum likelihood.* (Ph. D. dissertation, Massey University, New Zeland).

[22] Waterman, P.S. (1995). *Introduction to computational biology* (Chapman & Hall, London).

[23] Yang, Z. (1994). *J. Mol. Evol.* **39**, 105-111.

FIGURE CAPTIONS

Fig. 1. Projection on the 1-2 plane (a) and on the 1-3 plane (b) of the cloud of points of the alignments coded by the mitochondrial genome of metazoa. Each point corresponds to the matrix Q of an alignment and therefore represents two sequences. By convention, we have chosen the symbol of the rightmost species in the figure. For example, the alignment *Ascaris/Cattle* is labeled by Ascaris, since the alignments involving Ascaris are at the right of those involving vertebrates. The meaning of symbols is the following (see Section II-A for more details): ∘: vertebrates; □: urchins; ⋄: other invertebrates (shrimp, chiton, worm, fruit fly); △: bee; ×: land snail; +: wood snail; •: Ascaris; ■: Caenorhabditis.

Fig. 2. Probabilities of mutation from arginine to the 20 amino acids in the transition matrices M^{120}.

Fig. 3. Cloud of points of the alignments of proteins coded by mitochondrial genome after translating the genes of nematodes following the code given by Okimoto et al (1992).Projection on the 1-2 plane (a) and on the 1-3 plane (b). The conventions for symbols are as in Figure 1.

Fig. 4. Probabilities of mutation from cysteine to the 20 amino acids in the transition matrices M^{120}.