J Mol Evol (1993) 37:221–232

Journal of Molecular Evolution

Reconstructing Evolution from Eukaryotic Small-Ribosomal-Subunit RNA Sequences: Calibration of the Molecular Clock

Yves Van de Peer[1], Jean-Marc Neefs[1], Peter De Rijk[1], Rupert De Wachter[1]

[1] Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium

Abstract. The detailed descriptions now available for the secondary structure of small-ribosomal-subunit RNA, including areas of higly variable primary structure, facilitate the alignment of nucleotide sequences. However, for optimal exploitation of the information contained in the alignment, a method must be available that takes into account the local sequence variability in the computation of evolutionary distance. A quantitative definition for the variability of an alignment position is proposed in this study. It is a parameter in an equation which expresses the probability that the alignment position contains a different nucleotide in two sequences, as a function of the distance separating these sequences, i.e., the number of substitutions per nucleotide that occurred during their divergence. This parameter can be estimated from the distance matrix resulting from the conversion of pairwise sequence dissimilarities into pairwise distances. Alignment positions can then be subdivided into a number of sets of matching variability, and the average variability of each set can be derived. Next, the conversion of dissimilarity into distance can be recalculated for each set of alignment positions separately, using a modified version of the equation that corrects for multiple substitutions and changing for each set the parameter that reflects its average variability. The distances computed for each set are finally averaged, giving a more precise distance estimation.

Trees constructed by the algorithm based on variability calibration have a topology markedly different from that of trees constructed from the same alignments in the absence of calibration. This is illustrated by means of trees constructed from small-ribosomal-subunit RNA sequences of Metazoa. A reconstruction of vertebrate evolution based on calibrated alignments matches the consensus view of paleontologists, contrary to trees based on uncalibrated alignments. In trees derived from sequences covering several metazoan phyla, artefacts in topology that are probably due to a high clock rate in certain lineages are avoided.

Key words: Small-ribosomal-subunit RNA — Sequence variability — Substitution rate — Evolutionary distance — Vertebrate phylogeny — Metazoanphylogeny

Introduction

Ribosomal RNAs and the corresponding genes are probably the molecules most frequently used at present for the inference of evolutionary relationships among species on a molecular basis. The advantages of rRNAs as molecular clocks have been cited frequently (e.g., Woese 1987). They are, in essence, the functional constancy of the molecules and their universal occurrence in all forms of life that possess a protein-synthesizing system. The RNAs of the small and large ribosomal subunit (hereafter abbreviated as SSU rRNA and LSU rRNA; also called 16S-like and 23S-like rRNA)

Presented at the NATO Advanced Research Workshop on *Genome Organization and Evolution,* Spetsai, Greece, 16–22 September 1992

Abbreviations: SSU rRNA: small ribosomal subunit RNA, LSU rRNA: large ribosomal subunit RNA

Correspondence to: R. De Wachter

have the advantage over 5S rRNA of possessing a considerably larger chain length, which provides improved statistics of the measured sequence dissimilarity. At the moment, SSU rRNA presents the advantage that about seven times as many sequences are available as for LSU rRNA (De Rijk et al. 1992; Gutell et al. 1992) and that its secondary structure is known in more detail. The latter point is important because the boundaries of secondary-structure elements form a framework for the recognition of homologous nucleotides, which is helpful for the obtention of a meaningful sequence alignment, especially in areas of variable primary structure. Conversely, the addition of extra sequences to the alignment allows the detection of greater detail in higher-order structure, which reveals itself as compensating substitutions in corresponding columns of the alignment matrix. The elaboration of a sequence alignment and a secondary-structure model can thus be considered as a sort of cooperative process.

When selecting an informative macromolecule for the investigation of an evolutionary problem, one will look for a mutational rate adapted to the time scale of the study, i.e., a fast clock for examining recent divergence patterns, a slow clock for unravelling old relationships. A remarkable property of rRNAs is precisely that they contain an alternation of so-called conserved and variable sequence areas, the former functioning as slow clocks, the latter as fast ones. In the case of eukaryotic SSU rRNAs, about 38 of the 62 helices of the secondary-structure model (De Rijk et al. 1992) consist of more conserved sequences, while the remaining 24 helices are distributed over eight variable areas. The presence in rRNAs of both conserved and variable areas can be considered as an advantage since it allows the use of these molecules as clocks on an expanded evolutionary time scale. However, as explained in detail below, this is only possible if the variable areas can be clearly distinguished from the conserved ones and if the relative rate of fixation of substitutions is known quantitatively for the areas of different variability. In actual fact, the occurrence of variable areas is often considered a drawback rather than an advantage. This is because in the latter areas, not only substitutions but also deletion and insertion events occur more frequently, making sequence alignment and the derivation of the local secondary structure more difficult. For this reason, many investigators disregard the most variable areas in their computations, considering that their alignment is not dependable. Recently, however, the availability of an increasing number of SSU rRNA sequences has allowed the selection of plausible secondary-structure models for the variable areas (Neefs and De Wachter 1990; De Rijk et al. 1992) and the concomitant improvement of the sequence alignment in these areas.

Methods for tree construction accounting to different extents for the variability of nucleotide or amino acid sequences have been published previously. Golding (1983) and Olsen (1987) studied the effect of site-to-site differences in substitution rate on the estimation of evolutionary distance between nucleotide sequences. Manske and Chapman (1987), who used a matrix method to derive trees from 5S rRNA sequences, considered the "relative nucleotide variability" of each alignment position. When computing the evolutionary distance between two sequences, they used a weighting factor for positions where this quantity exceeds a certain threshold. A short criticism of this principle is formulated below. Williams and Fitch (1990) introduced weighted parsimony wherein weighting not only applies to alignment positions but also to the different types of substitutions possible at each position. A drawback of the latter two methods seems to be that different weighting methods are possible and that the choice influences the obtained tree topology to some extent.

In the present paper, we use a matrix method to compute trees. The merits of parsimony vs matrix methods will not be discussed here. Suffice it to say that the latter allows the construction of trees from much larger numbers of sequences, which is an important advantage in the case of SSU rRNA in view of the size of the available data set. In the method that we propose, alignment positions are subdivided into sets of similar variability on the basis of estimates of the substitution rate of each individual position. This allows an independent conversion of sequence dissimilarity into evolutionary distance for each set, achieved by changing a parameter in the equation that performs the conversion.

Nucleotide Sequences, Definitions, Algorithms, and Computer Programs

A database on SSU rRNA structure has been kept up to date in our laboratory since 1984 and its contents have been published yearly and made available upon request (De Rijk et al. 1992). In July 1992, this database comprised about 1,500 complete or nearly complete sequences, aligned on the basis of similarity in primary and secondary structure. Approximately 250 of these SSU rRNA sequences are from eukaryotes. Duplicate sequences belonging to the same species and multiple sequences belonging to closely related species of the same genus (e.g., *Tetrahymena*) were eliminated from this set. Table 1 shows how the 205 remaining sequences used in the present study are distributed over a number of large eukaryotic taxa.

Table 1. Distribution of SSU rRNA sequences used in this study to calibrate the variability of alignment positions

Taxon[a]	Number of sequences
Metazoa	39
Fungi[b]	68
Green plants	30
Green algae	22
Chromophytes	7
Red algae	5
Oomycetes	3
Ciliates	13
Other Protoctists[c]	18

[a] These taxa correspond to major clusters discernable in phylogenetic trees based on SSU rRNA sequence alignments (Sogin et al. 1989; Wolters 1991; Van de Peer et al. 1993)
[b] Included are chytridiomycetes, zygomycetes, ascomycetes, and basidiomycetes
[c] Included are dinoflagellates, Apicomplexa, kinetoplastids, slime moulds, microsporidians, and diplomonads

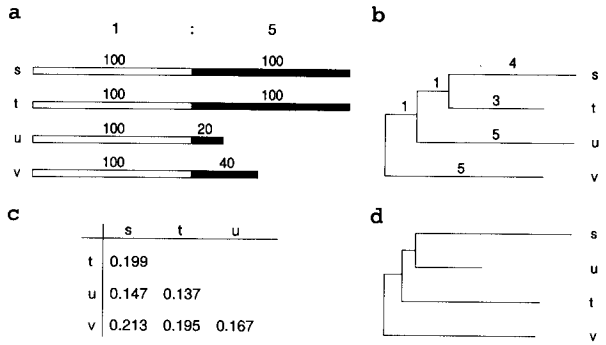

Fig. 1. Effect of sequence variability on tree construction. **a** Alignment of nucleotide sequences *s* to *v*. Conserved areas, drawn as *hollow bars*, are 100 nucleotides long. Variable areas, drawn as *filled bars*, have different lengths and the substitution rate during divergence is five times that in the conserved areas. **b** Actual divergence scheme that has given rise to sequences *s* to *v*. *Figures along the branches* give the number of substitutions per 100 nucleotides that have taken place in the conserved areas. **c** Distance matrix. Dissimilarity for each sequence pair was predicted on the basis of the known substitution rates for the conserved and variable areas according to equation (1b). (See text.) Conversion of dissimilarity into distance was according to Jukes and Cantor (1969). **d** Tree obtained by neighbor-joining from the matrix in **c**.

For the computation of dissimilarity between two sequences, only the alignment positions where both contain a nucleotide were compared; in other words, insertions and deletions in one sequence with respect to the other were disregarded in this computation. The dissimilarity between two aligned sequences is defined as the number of positions in which the sequences contain a different nucleotide divided by the total number of compared positions. The evolutionary distance between the two sequences is defined as the number of substitutions that have actually occurred in the compared positions since their divergence, divided by the total number of compared positions. In a first approximation, pairwise evolutionary distances were estimated from the observed pairwise dissimilarities by means of the equation of Jukes and Cantor (1969). A more sophisticated method, which accounts for the existence of areas of different variability, is explained in detail in the following paragraphs. Evolutionary trees were reconstructed from the resulting distance matrix by neighbor-joining (Saitou and Nei 1987) as implemented in the software package TREECON (Van de Peer and De Wachter 1993). Neighbor-joining appears to be one of the most effective tree construction methods available at the moment (Saitou and Nei 1987; Sourdis and Nei 1988; Saitou and Imanishi 1989).

Curve fitting by nonlinear regression was done on a VAX-Station 3100 (Digital), using the software module of Carmenes (1991).

Accounting for Sequence Variability in Tree Construction

Figure 1 demonstrates, by means of a simple example, how the presence of variable areas in sequences can distort not only the branch lengths of the tree obtained, but even its topology. Consider an alignment of four sequences, *s* to *v*, consisting of one conserved and one variable area (Fig. 1a). During evolution, the latter area accumulates substitutions five times faster than the former, and in addition it has acquired a different length in the four sequences because it also undergoes more deletions and insertions. Such a situation occurs, for instance, in variable area V4 of eukaryotic SSU rRNA (De Rijk et al. 1992), which ranges in length from three nucleotides in the microsporidian *Vairimorpha necatrix* to 546 nucleotides in the insect *Acyrthosiphon pisum*, while in the majority of species it covers about 230 nucleotides. Figure 1b shows the actual divergence scheme that has given rise to sequences *s* to *v*. Although *s* and *t* are the most recently diverged pair, a tree constructed by neighbor-joining (Fig. 1d) from a distance matrix (Fig. 1c) suggests that *s* is closer to *u* than to *t*. Other treeing methods such as UPGMA (unweighted pair group method; Sneath and Sokal 1973) would cluster *t* with *u*, as can be seen from the distance matrix, producing a wrong topology as well. If the variable area could have been distinguished from the conserved one, the relative substitution rates measured, and the conversion of dissimilarity to distance carried out by an appropriate method, then the correct topology would have been obtained. It is the purpose of the present study to elaborate methods that make these operations possible.

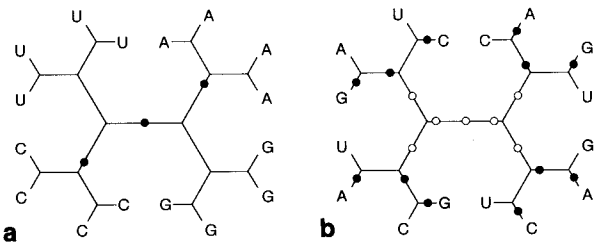

Fig. 2. Unrooted tree leading to 16 sequences, and the resulting nucleotide occupancy in two positions of the sequence alignment. The "relative nucleotide variability" (Manske and Chapman 1987) equals 2 for both positions. **a** More conserved position. The *black circles* represent the minimum number of substitutions required to obtain this occupancy. **b** More variable position. *Black circles* represent the minimum number of substitutions required. *Hollow circles* represent probable sites of additional substitutions assuming an approximately constant substitution rate during evolution.

It should be noted that errors in the distance computation would also occur with sequences of equal length containing areas of different substitution rate. The difference in length merely exacerbates the problem. Of course, the error could be eliminated by using only that part of the alignment common to all the sequences, but this would amount to throwing away 40% of the information when comparing s and t in the example of Fig. 1, and the sensitivity and accuracy of the distance measurement would be decreased. The availability of an estimate of the relative rate of substitution of the two areas would not only solve the problem illustrated in Fig. 1 but would also allow the simultaneous use in tree construction of partial and complete sequences without reduction of the alignment to the areas available for all the species to be compared.

Figure 2 illustrates why "relative nucleotide variability" (RNV; Manske and Chapman 1987) is not a very appropriate measure to correct for different substitution rates. It gives the pattern of nucleotide occupancy for two alignment positions in 16 sequences connected by an unrooted tree. Obviously, the pattern for position (a) results from a much lower substitution rate than the pattern for position (b). Nevertheless, an identical RNV of 2, which is the maximum value, would be assigned to both positions. This is because RNV only takes into account the distribution of nucleotides for a given alignment position, but ignores the average evolutionary distance necessary to achieve a substitution.

In the following paragraphs, the variability of an alignment position is defined as a quantity proportional to its substitution rate and a method is described for deriving this quantity for each position, relative to the average variability of all positions.

Calibration of Positional Variability in Eukaryotic SSU rRNA

Probability of Substitution as a Function of Evolutionary Distance

Consider two homologous nucleotide sequences that have diverged in the course of evolution. For the sake of simplicity, it will be assumed that the four nucleotides occur in approximately the same amounts, that divergence only involves substitutions, and that all substitutions are equally probable. Equation (1) gives the probability p that a position of the sequence alignment contains different nucleotides in the two sequences, as a function of the evolutionary distance d, accounting for the possibility of multiple substitutions per site.

$$p = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}d\right)\right] \tag{1a}$$

Consequently, the same equation gives the expected dissimilarity f between the two sequences, defined as the fraction of alignment positions containing different nucleotides in the two sequences:

$$f = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}d\right)\right] \tag{1b}$$

The inverse of function (1b) is the equation of Jukes and Cantor (1969), which allows one to compute the distance d between two sequences, corrected for multiple substitutions, from the observed dissimilarity f:

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}f\right) \tag{2}$$

The shape of functions (1) and (2) is illustrated in Fig. 3a and 3b.

Probability of Substitution for Positions of Different Variability

In fact, equation (1a) only applies if the probability of substitution is the same for all the positions of an alignment. In a sequence alignment of real molecules, this is not the case. The variability v_n of a position n, relative to the average variability of the entire alignment, can be defined as the ratio of the probability that a substitution occurs at this position to the average probability of substitution per site for the entire alignment:

$$v_n = \frac{s_n L}{\sum_{i=1}^{L} s_i}$$

where s_n and s_i are the probabilities of substitution, in a given time span, of the nucleotides at positions n and i, and L is the alignment length.

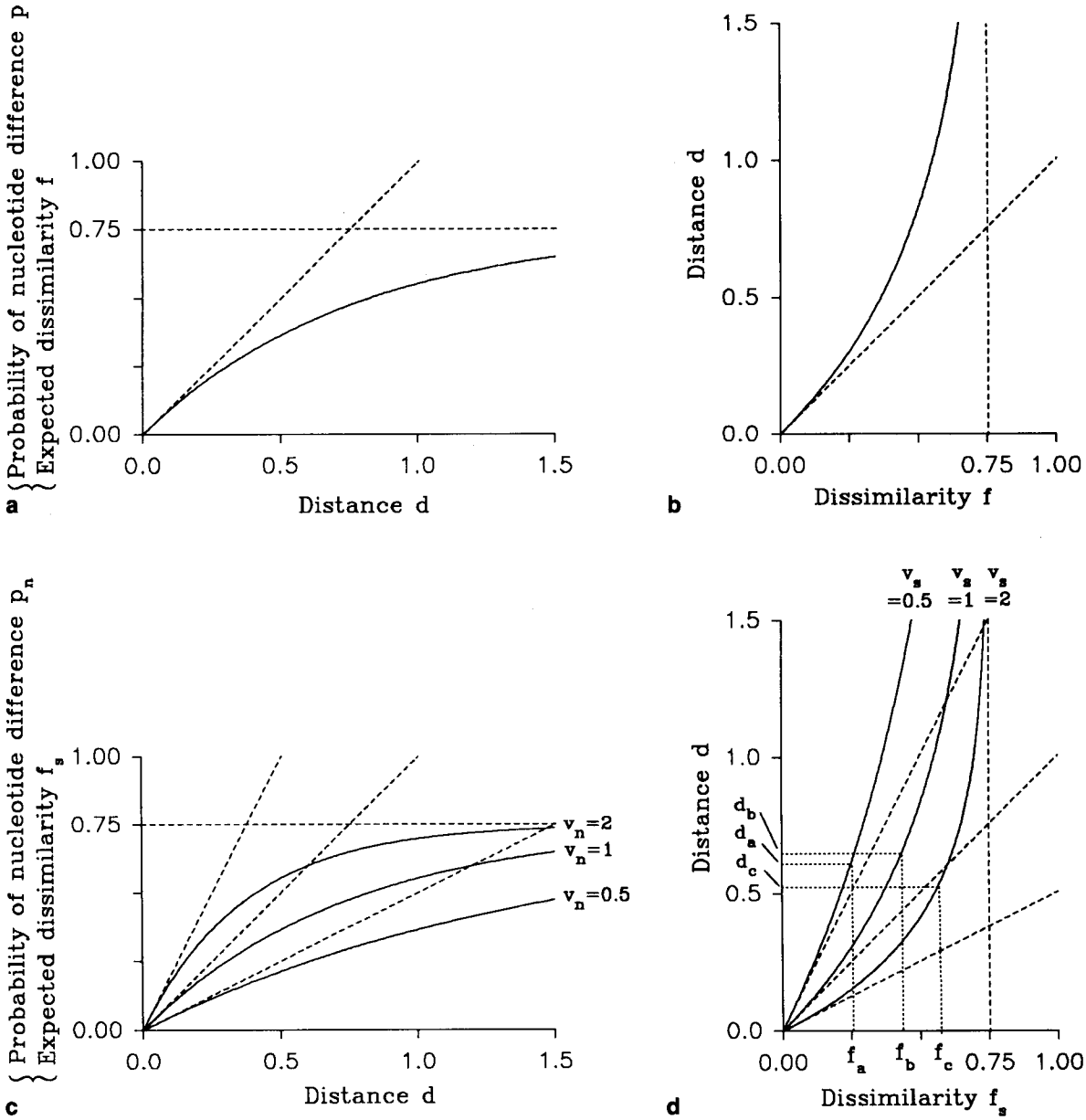

Fig. 3. Graphic representation of functions (1) to (4). **a** Equation (1): probability of observing a different nucleotide in homologous positions of a sequence pair (eq. 1a), or expected dissimilarity (eq. 1b), as a function of evolutionary distance (substitutions per nucleotide) separating two sequences. **b** Equation (2): inverse function of (1), allowing one to estimate evolutionary distance from the dissimilarity observed between a pair of sequences. **c** Equation (3): similar to equation (1), but expressing the probability of observing a different nucleotide in homologous positions of a sequence pair for a position of variability v_n (eq. 3a), or the expected dissimilarity (eq. 3b) for an area of variability v_s. The function is shown for three values of the parameter v_n (or v_s). **d** Equation (4): inverse function of (3), allowing one to estimate evolutionary distance from the dissimilarity observed between areas of calibrated variability in a sequence pair. The function is plotted for three areas of different variability. Ideally, dissimilarities f_a, f_b, and f_c observed for the three areas should give the same distance reading, but in practice the readings d_a, d_b, and d_c will differ slightly due to the stochastic nature of the substitution process.

The following equation expresses the probability of observing a substitution at position n in two molecules separated by a distance d:

$$p_n = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}v_n d\right)\right] \qquad (3a)$$

The shape of function (3a) is shown in Fig. 3c for positions of high, average, and low variability. The derivative of p_n as a function of d is:

$$\frac{dp_n}{dd} = v_n \exp\left(-\frac{4}{3}v_n d\right)$$

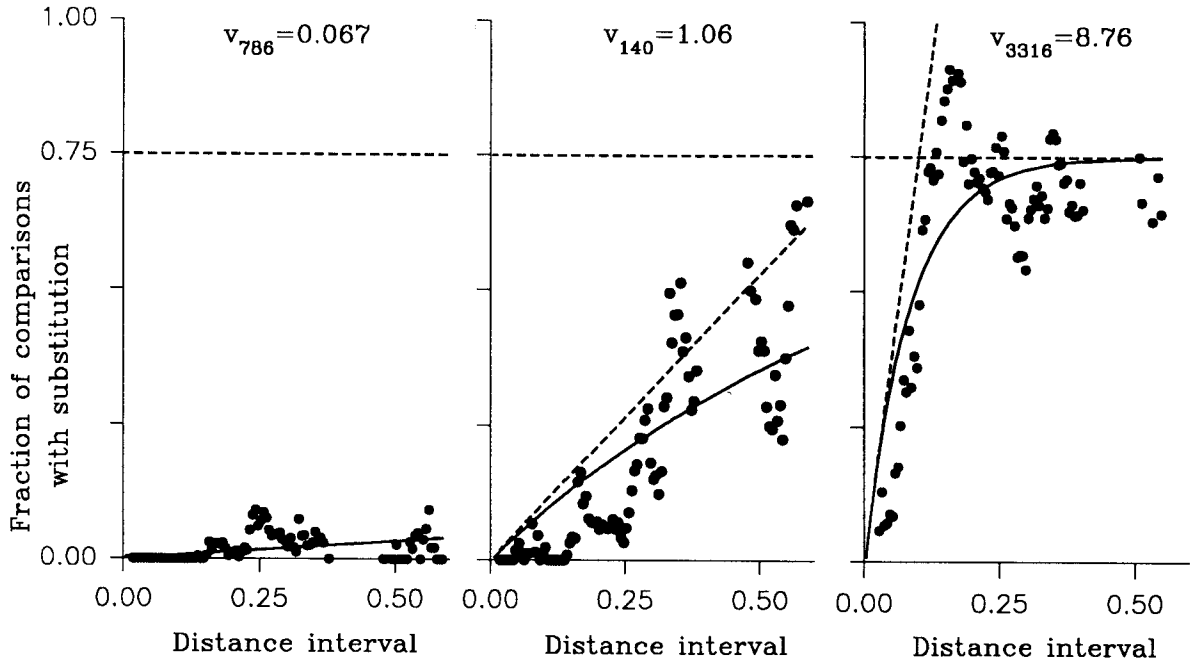

Fig. 4. Estimating the variability of alignment positions. The 20,910 values of the distance matrix for 205 eukaryotic SSU rRNA sequences, computed according to equation (2), were classified into intervals with a width of 0.005 distance units. For each interval, the fraction of sequence pairs showing a different nucleotide in the considered position was plotted against the distance. As an example, 493 sequence pairs are at a distance falling in the interval 0.280–0.285, and 111 of these pairs, or 22.5%, show a different nucleotide at position 140. The value 0.225 is therefore plotted against the distance 0.283, which is the mean of the distances falling in the interval, in the graph for position 140. The resulting graphs are shown for position 786 (low variability), 140 (medium variability), and 3318 (high variability) of the sequence alignment. A curve obeying equation (3a) was fitted to the points by nonlinear regression. The slope of this curve in the origin gives the variability v_n of position n.

and the value of this derivative for distance $d = 0$ is v_n. Hence the slopes of the curves in the origin (Fig. 3c) are equal to v_n, which means that position n undergoes substitution at a rate v_n times that of the average rate for the entire sequence.

If we now consider a subset of positions, s, with an average variability v_s different from the average variability of the complete sequence, then the same equation (3a) also predicts the fraction f_s of these positions that will contain different nucleotides if sequences separated by a distance d are compared:

$$f_s = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3} v_s d\right)\right] \qquad (3b)$$

Estimating the Variability of an Alignment Position

It is now possible to deduce the relative variability v_n of each position as follows. For an alignment of N sequences, a set of $N(N-1)/2$ pairwise distances d (the distance matrix) can be computed according to equation (2). All the pairwise distances are classified into a number of distance intervals, e.g., distances smaller than 0.005, distances from 0.005 to 0.010, and so on. For all the pairs falling within a given distance interval, the fraction accompanied by a nucleotide change in the considered position is computed. This fraction is plotted against the mean distance of the interval, as illustrated in Fig. 4 for three different positions of the SSU rRNA alignment. A curve obeying equation (3a) is fitted to the points by nonlinear regression. The slope of this curve in the origin yields parameter v_n for the position under consideration.

Sorting Alignment Positions Into Subsets of Similar Variability

After parameter v_n has been determined for all positions, they can be partitioned into a number of sets of similar variability—e.g., one set of variability less than 0.4, a second set of variability comprised between 0.4 and 0.8, etc. The distribution of v_n values obtained for the positions of the alignment of eukaryotic SSU rRNA sequences is shown in Fig. 5. In this case, 1,558 positions were partitioned into five sets of equal size, with the v_n values for each set comprised between the limits indicated on the figure. These five sets do not include those positions that contain a gap in more than 75% of the aligned sequences, nor do they include 276 invariable positions, which contain the same nucleotide in all sequences.

Each set of positions selected as described above can now be considered as a subset of the complete

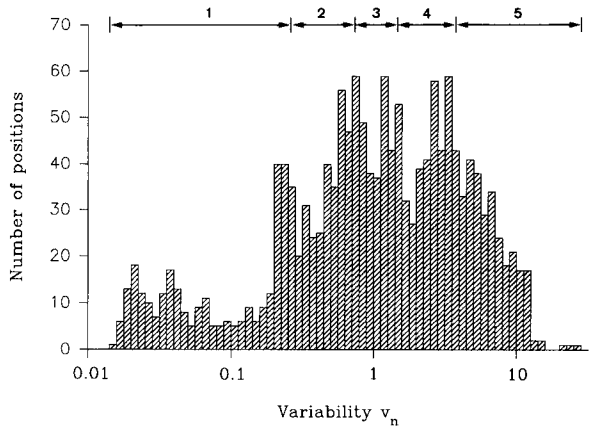

Fig. 5. Distribution of positional variabilities. Variabilities were estimated as shown in Fig. 4 for each of the 1,558 alignment positions that are not absolutely conserved and contain a nucleotide in at least 25% of the aligned sequences. The lowest variability found was 0.0147, the highest one 26.62, relative to the average variability of all positions. The *abscissa* of the histogram is on a logarithmic scale, with a ratio of 1.122 between the highest and the lowest variability in the same interval. Positions were divided into five sets of equal size as marked on top of the graph. The variability limits separating the sets are 0.2737, 0.7384, 1.559, and 3.474.

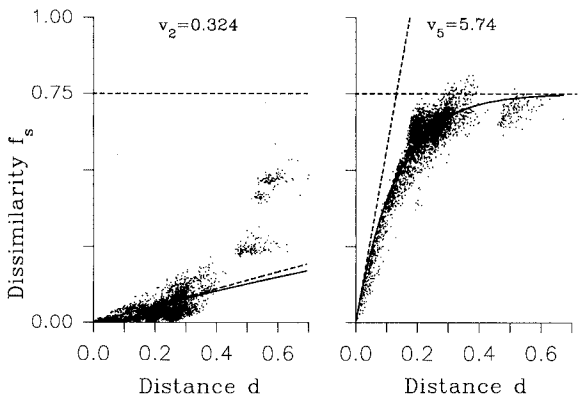

Fig. 6. Estimating the variability of a set of positions. For each of the 20,910 values of the distance matrix for 205 eukaryotic SSU rRNA sequences, dissimilarity was computed for each of the five subsets of positions defined as shown in Fig. 5. In each set, the 20,910 dissimilarity values were plotted against the distance between the complete sequences. The resulting graph is shown for set 2 and set 5 as defined in Fig. 5. A curve obeying equation (3b) was fitted to the graphs. The slope of this curve in the origin is the variability found for the subset of positions. The variabilities for the five subsets are $v_1 = 0.0918$, $v_2 = 0.324$, $v_3 = 0.977$, $v_4 = 2.38$, and $v_5 = 5.74$.

alignment, containing nucleotides that differ much less in relative variability than the complete set of nucleotides. For such a subset, the fraction of substituted nucleotides, f_s, is computed for each of the $N(N-1)/2$ pairwise comparisons that can be made. For each comparison, f_s is plotted against the distance between the complete sequences, d, obtained from equation (2). The curve obeying equation (3b) that best fits the plotted values is computed by nonlinear regression. The slope of this curve in the origin yields the average variability, v_s, for the set of positions. The curve fitting is shown in Fig. 6 for two of the five subsets of positions selected from the SSU rRNA alignment.

Variability Map of Eukaryotic SSU rRNA

Figure 7 shows a secondary structure model for eukaryotic SSU rRNA, where the positions belonging to the five sets of different average variability are indicated by dots of five different sizes. Invariant positions are indicated by a square. Since this model is based on the *Saccharomyces cerevisiae* 18S rRNA structure, which contains few insertions with respect to the majority of other sequences, nearly all positions are either invariant or assignable to one of the five sets. The map gives a much more detailed description of positional variability than the crude distinction between "conserved" and "variable" areas that is often made (e.g., De Rijk et al. 1992). Moreover, the description is not intuitive, but based on a quantitative estimate of relative substitution rates.

Computing Evolutionary Distance from an Alignment Calibrated for Variability

Since parameter v_s in equation (3b) is now known for each subset of positions, it is possible to use each subset independently to compute the evolutionary distance between two complete sequences. To this end the inverse function of equation (3b) is used:

$$d_s = -\frac{3}{4}\frac{1}{v_s}\ln\left(1 - \frac{4}{3}f_s\right) \qquad (4)$$

where v_s is the average variability of the positions of subset s and f_s is the fraction of the positions of subset s that are dissimilar in the compared sequences. Ideally, the derived distance d_s should be the same, regardless of the subset of positions considered. In practice, due to the stochastic nature of the mutational process, the values will be slightly different, as illustrated in Fig. 3d. The smaller the subset of positions, the larger the fluctuations to be expected. The average of the d_s values derived for each set is therefore computed, weighting each value proportionally to the number of positions in the subset. As an example, let us assume that the positions of an alignment have been partitioned into three subsets a, b, and c of different average variability (low, medium, and high). After computing the distances d_a, d_b, and d_c, according to equation (4), the average of the three values is obtained as:

$$d = \frac{L_a}{L}d_a + \frac{L_b}{L}d_b + \frac{L_c}{L}d_c \qquad (5)$$

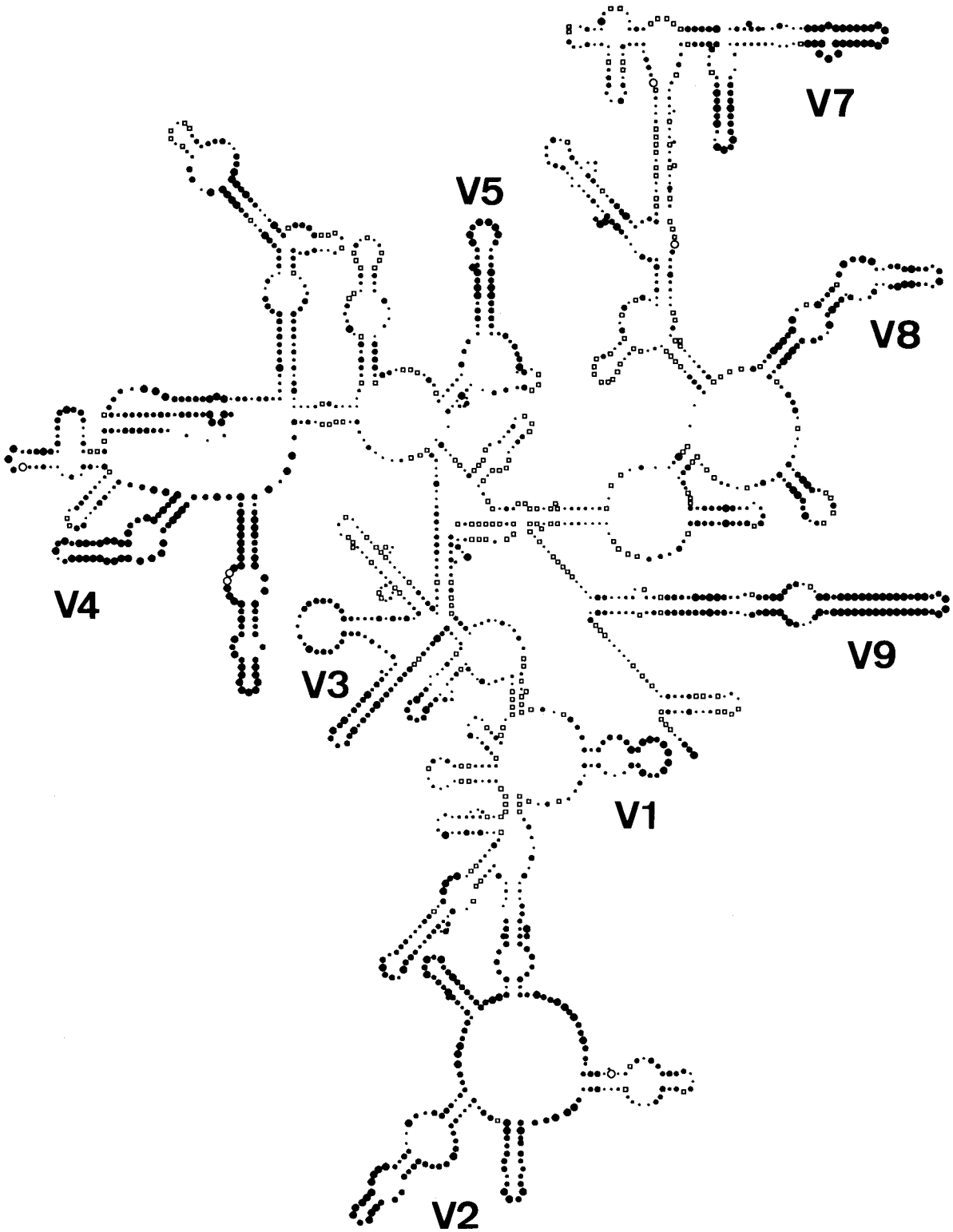

Fig. 7. Variability map of eukaryotic SSU rRNA. Each *dot* represents a nucleotide of *Saccharomyces cerevisiae* 18S rRNA drawn in the secondary structure model adopted in De Rijk et al. (1992). Positions belonging to five different variability sets (see Fig. 5) are shown as *filled circles* of a size commensurate with their variability. *Hollow squares* are invariable positions. *Hollow circles* are nucleotides that are deleted in most other eukaryotic SSU rRNAs. *V1* to *V5* and *V7* to *V9* indicate areas previously distinguished (e.g., De Rijk et al. 1992) on an intuitive basis.

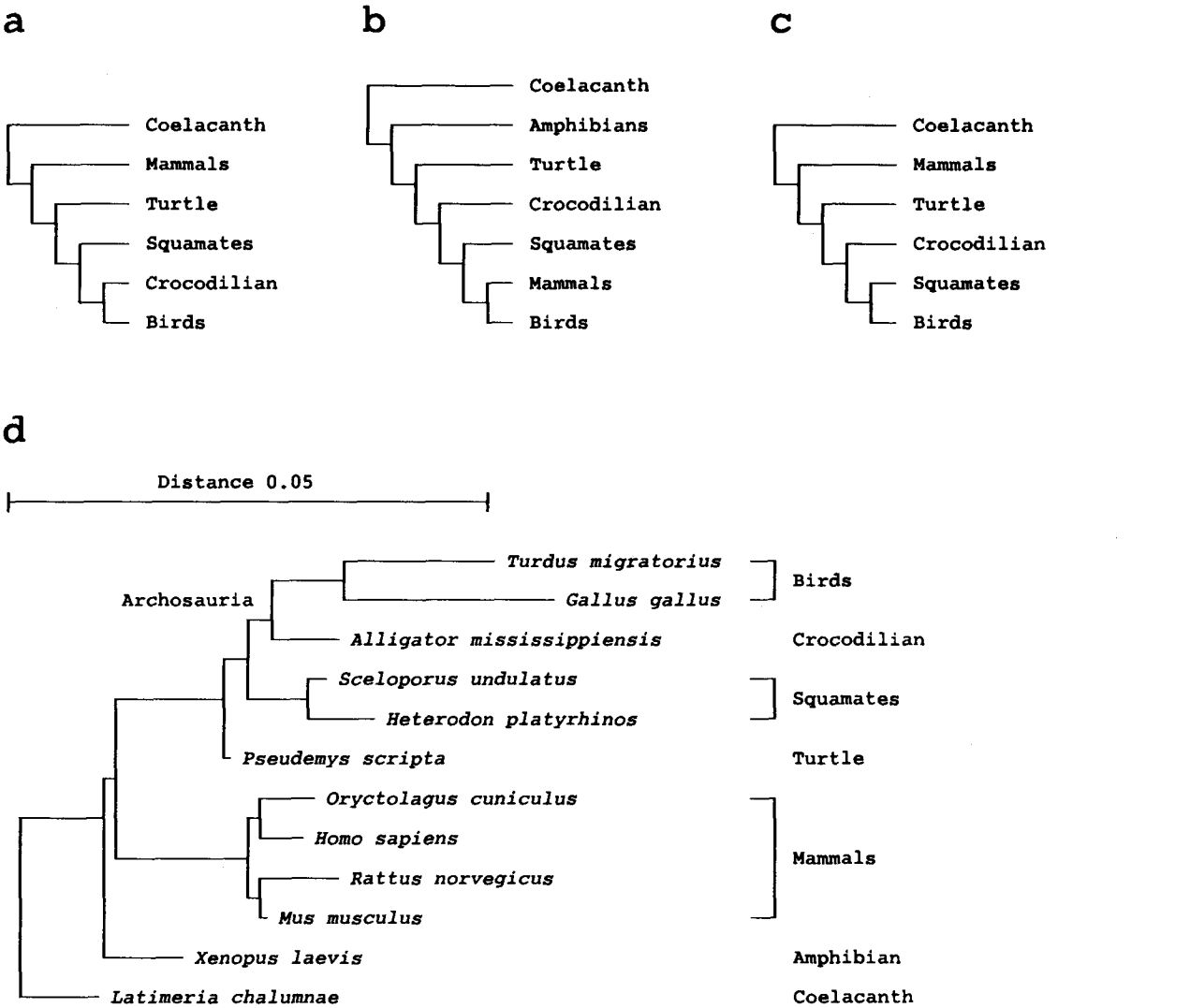

Fig. 8. Alternative phylogenies of major amniote and vertebrate classes. **a** Paleontological phylogeny (Donoghue et al. 1989). **b** Inferred phylogeny based on (unweighted) SSU rRNA analysis (Hedges et al. 1990). **c** Amniote phylogeny based on dynamically weighted parsimony analysis of SSU rRNA as applied by Marshall (1992). **d** Detailed evolutionary tree of vertebrates as found by calibration of alignment positions. Evolutionary distances were calculated as described in the text. The distance between two organisms is obtained by summing the lengths of the connecting branches along the horizontal axis, using the *scale* on top.

where L is the total number of nucleotides compared and L_a, L_b, and L_c are the numbers of nucleotides in subsets a, b, and c. In the case of the SSU rRNA alignment, where the positions were divided into five subsets of the same size, the distance d practically equals the arithmetic average of the distances derived for each of the five sets. A new matrix of distances, computed according to equation (5), is thus obtained and can serve as input for a tree construction method.

Application to Tree Construction

Two examples, both taken from the field of animal evolution, are given below in order to demonstrate the effect of variability calibration on the topology of the trees obtained.

Recently, the controversy over amniote relationships was revived by a study based on 18S rRNA sequences. As illustrated in Fig. 8a, paleontological data suggest that birds and crocodiles are sister groups (Donoghue et al. 1989). However, analysis of rRNA sequences, both by a distance and a parsimony method (Hedges et al. 1990), points to a closer relationship of birds to mammals than to crocodiles (Fig. 8b). The latter finding is supported by studies on two other genes, viz. hemoglobin and myoglobin, while histone H2B and pancreatic polypeptide identify the crocodiles as the sister group of birds. Still other studies based on alpha cristallin A,

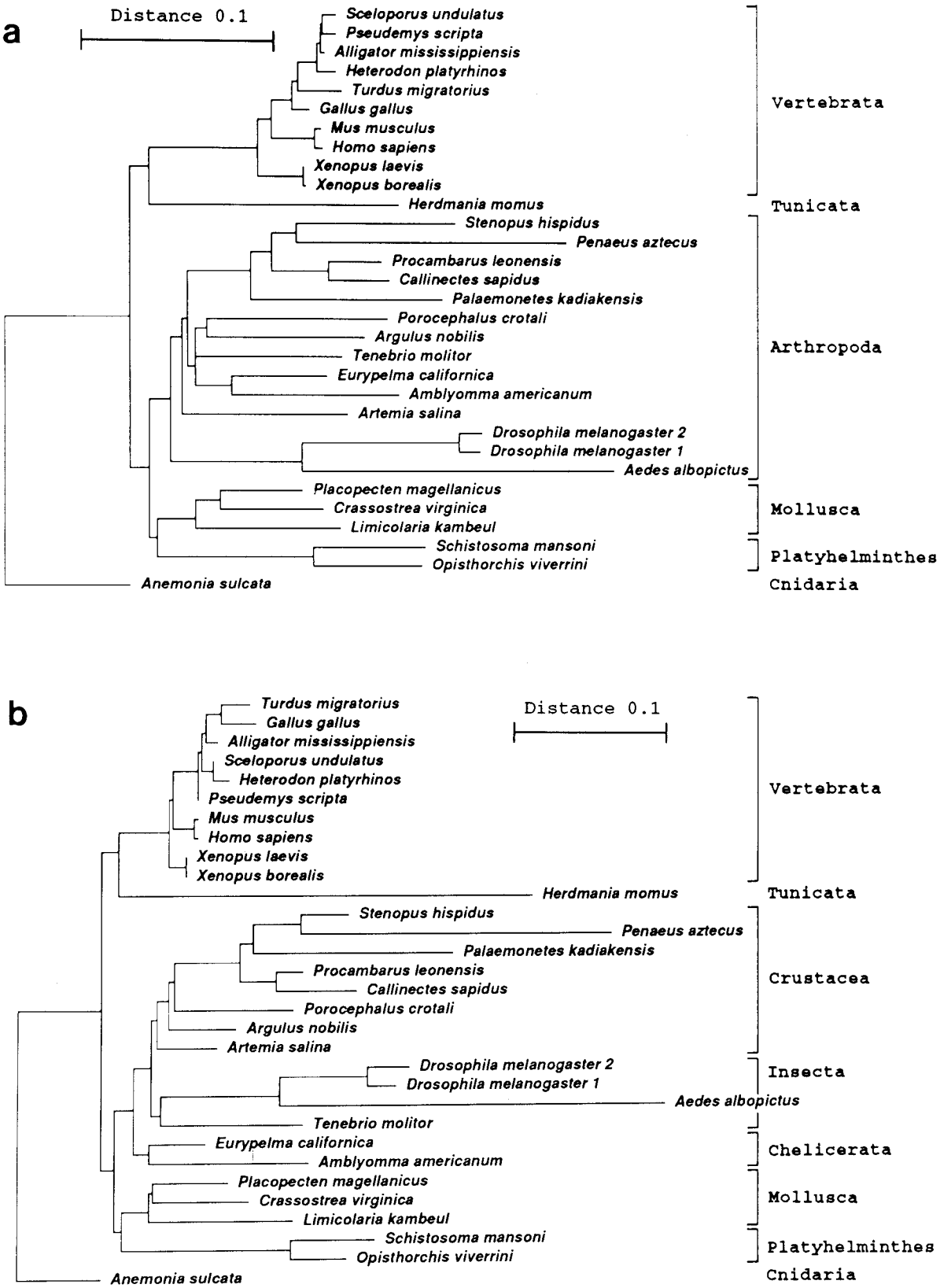

Fig. 9. Phylogeny of the Metazoa as derived from SSU rRNA sequences. Taxon designations are placed to the right of the corresponding cluster. **a** Phylogenetic tree inferred on the basis of a noncalibrated alignment. **b** Phylogenetic tree of the same set of organisms as in **a** but inferred on the basis of a calibrated alignment.

alpha hemoglobin, insulin, and large-ribosomal-subunit RNA support various relationships among birds, reptiles, and mammals depending on the type of analysis (Hedges et al. 1990; Hedges and Maxson 1991). Marshall (1992) argued that the bird–mammal relationship found by using the 18S rRNA is probably an artefact caused by substitution biases. In applying the dynamically weighted parsimony method of William and Fitch (1990), he indeed found a different tree topology, with the birds more closely related to the crocodiles (Fig. 8c) but still separated from them by the Squamata. When applying the calibration method described in this paper to the amniote sequences, we find a tree topology that is exactly the same as the paleontological one (Fig. 8d). In this tree, the turtle is the first reptile that branches off, followed by the squamates (represented by the lizard *Sceleporus undulatus* and the snake *Heterodon platyrhinos*), while the crocodile and the birds form a monophyletic assemblage, which is traditionally named Archosauria.

A tree for a more diverse set of metazoan sequences, constructed on a noncalibrated alignment, is shown in Fig. 9a. Beside vertebrates, a tunicate, arthropods, molluscs, and Plathyhelminthes are also included. The relationship among amphibians, mammals, reptiles, and birds is different from the one observed in Fig. 8d. In addition, the insects do not form a single cluster. As is usually the case with SSU rRNA trees, the insects *Drosophila* and *Aedes* are found at the base of the arthropod cluster, while *Tenebrio* branches off later (Hendriks et al. 1990, 1991). This phenomenon is probably caused by the faster-evolved SSU rRNA sequences of *Drosophila* and *Aedes*. Higher evolutionary rates tend to pull organisms closer to the base of the tree (Olsen 1987; Woese 1991). In contrast, the tree based on a calibrated alignment (Fig. 9b) shows the same relationships within the vertebrate cluster as in Fig. 8d. In addition, the insects form a monophyletic grouping. Hence, the calibration solves both anomalies present in the tree of Fig. 9a.

Discussion

It has been noted by Olsen (1987) that the inference of reliable phylogenetic trees with a distance method depends on the accurate estimation of evolutionary distances from the sequence dissimilarity. The latter author showed that the existence of different substitution rates in different areas of a sequence alignment makes the equation of Jukes and Cantor (1969) less suitable for conversion of observed dissimilarity into estimated distance. He proposed a different algorithm, but this still converts a single dissimilarity into a single distance for each sequence pair. In the present paper, on the contrary, we dissect the sequences into a number of distinct subsets of nucleotides, each subset consisting of nucleotides that show a much narrower range of variability than the entire sequence. Dissimilarity is measured for each subset and converted into distance by means of an equation (4) in which an appropriate parameter v_s, after being quantitatively measured, can be introduced for each subset. Thus the information present in the sequence alignment is no longer reduced to a single number for each pairwise sequence comparison, as is usually the case in matrix methods, but is expressed in as many numbers as nucleotide subsets of similar variability that are considered. Due to the narrow range of variability, each measurement gives a better precision, and the different measurements are then averaged. In the present case we considered five subsets, but we intend to examine other sizes and numbers of sets in the future. Conceivably, the precision of the distance measurement may increase further as the number of sets, five in the present study, is increased. At some point this effect will be counterbalanced by the fact that the subsets become too small and the statistical accuracy on each measurement becomes too low.

A drawback of the method is that the calibration of nucleotide variability is rather time-consuming and that new calibration should be considered whenever the alignment is improved or extended with a considerable number of new sequences, since this may result in slight rearrangements of the subsets, or in a more accurate estimation of the variability parameters of equation (3b). However, as long as an alignment remains valid, the parameters can be used for all trees constructed from it. Conversion of distance into dissimilarity for five sets of nucleotides rather than for a single set does not noticeably increase the computer time needed for tree construction.

Variability calibration has been hitherto carried out for eukaryotic SSU rRNA sequence alignments and applied to the study of Metazoan evolution because comparison is possible with morphological and paleontological hypotheses. Since the results are promising and demonstrate the importance of accounting for differences in substitution rate, application to bacterial SSU rRNAs is planned in the near future. Application to large-subunit rRNA sequence alignments will be envisaged as soon as sufficiently detailed secondary-structure models, and hence dependable alignments, are available for the most variable areas. The possibility of the application to protein sequences, where a strong difference in substitution rate exists between replacement sites and silent sites, also deserves investigation. Silent sites may not have to be discarded from the

analysis, as is now the case in many evolution studies based on proteins, but may be included in the calculation of evolutionary distances provided that their high substitution rate is adequately accounted for.

Acknowledgments. This work was supported by the Incentive Programme for Fundamental Research in the Life Sciences (contract BIO/03) and the Programme on Interuniversity Poles of Attraction (contract 23) of the Office for Science Policy Programming of the Belgian State, and by the Fund for Medical Scientific Research. P. De Rijk is research assistant of the National Fund for Scientific Research.

References

Carmenes RS (1991) LSTSQ: a module for reliable constrained and unconstrained nonlinear regression. Comput Applic Biosci 7:373–378

De Rijk P, Neefs J-M, Van de Peer Y, De Wachter R (1992) Compilation of small ribosomal subunit RNA sequences. Nucleic Acids Res 20:2075–2089

Donoghue MJ, Doyle JA, Gauthier J, Kluge AG, Rowe T (1989) The importance of fossils in phylogeny reconstruction. Annu Rev Ecol Syst 20:431–460

Golding GB (1983) Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol Biol Evol 1:125–142

Gutell RR, Schnare MN, Gray MW (1992) A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. Nucleic Acids Res 20:2095–2109

Hedges SB, Moberg KD, Maxson LR (1990) Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol Biol Evol 7:607–633

Hedges SB, Maxson LR (1991) Pancreatic polypeptide and the sister group of birds. Mol Biol Evol 8:888–891

Hendriks L, Van de Peer Y, Van Herck M, Neefs JM, De Wachter R (1990) The 18S ribosomal RNA sequence of the sea anemone *Anemonia sulcata* and its evolutionary position among other eukaryotes. FEBS Lett 269:445–449

Hendriks L, De Baere R, Van de Peer Y, Neefs JM, Goris A, De Wachter R (1991) The evolutionary position of the rhodophyte *Porphyra umbilicalis* and the basidiomycete *Leucosporidium scottii* among other eukaryotes as deduced from complete sequences of small ribosomal subunit RNA. J Mol Evol 32:167–177

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Manske CL, Chapman DJ (1987) Nonuniformity of nucleotide substitution rates in molecular evolution: computer simulation and analysis of 5S ribosomal RNA sequences. J Mol Evol 26:226–251

Marshall CR (1992) Substitution bias, weighted parsimony, and amniote phylogeny as inferred from 18S rRNA sequences. Mol Biol Evol 9:370–373

Neefs J-M, De Wachter R (1990) A proposal for the secondary structure of a variable area of eukaryotic small ribosomal subunit RNA involving the existence of a pseudoknot. Nucleic Acids Res 18:5695–5704

Olsen GJ (1987) Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol LII:825–837

Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Saitou N, Imanishi T (1989) Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbour-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol 6:514–525

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Sogin ML, Edman U, Elwood H (1989) A single kingdom of eukaryotes. In: Fernholm B, Bremer K, Jörnvall H (eds) The hierarchy of life. Elsevier Science Publishers, Amsterdam, pp 133–143

Sourdis J, Nei M (1988) Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol Biol Evol 5:298–311

Van de Peer Y, De Wachter R (1992) TREECON: a software package for the construction and drawing of evolutionary trees. Comput Applic Biosci 9:177–182

Van de Peer Y, Neefs JM, De Rijk P, De Wachter R (1993) Evolution of eukaryotes as deduced from small ribosomal subunit RNA sequences. Biochem Syst Ecol 21:43–55

Williams PL, Fitch WM (1990) Phylogeny determination using dynamically weighted parsimony method. Methods Enzymol 183:615–626

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR (1991) The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria. In: Selander RK, Clark AG, Whittam TS (eds) Evolution at the molecular level. Sinauer, Sunderland, MA, pp 1–24

Wolters J (1991) The troublesome parasites—molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-ciliate clade. BioSystems 25:75–83