A new efficient method for assessing missing nucleotides in DNA sequences in the framework of a generic evolutionary model

Abdoulaye Baniré Diallo[1], Vladimir Makarenkov[2], Mathieu Blanchette[1], and François-Joseph Lapointe[3]

[1] McGill Centre for Bioinformatics and School of Computer Science,
McGill University 3775 University Street, Montreal, Quebec, H3A 2A7, Canada
[2] Département d'informatique, Université du Québec à Montréal,
C.P. 8888, Succ. Centre-Ville, Montréal (Québec), H3C 3P8, Canada
[3] Département de sciences biologiques, Université de Montréal,
C.P. 6128, Succ. Centre-Ville, Montréal (Québec), H3C 3J7, Canada

Abstract. The problem of phylogenetic inference from datasets including incomplete characters is among the most relevant issues in systematic biology. In this paper, we propose a new probabilistic method for estimating unknown nucleotides before computing evolutionary distances. It is developed in the framework of the Tamura-Nei evolutionary model (Tamura and Nei (1993)). The proposed strategy is compared, through simulations, to existing methods "Ignoring Missing Sites" (IMS) and "Proportional Distribution of Missing and Ambiguous Bases" (PDMAB) included in the PAUP package (Swofford (2001)).

1 Introduction

Incomplete datasets can arise in a variety of practical situations. For example, this is often the case in molecular biology, and more precisely in phylogenetics, where an additive tree (i.e. phylogenetic tree) represents an intuitive model of species evolution. The fear of missing data often deter systematists from including in the analysis the sites with missing characters (Sanderson et al. (1998), Wiens (1998)). Huelsenbeck (1991) and Makarenkov and Lapointe (2004) pointed out that the presence of taxa comprising big percentage of unknown nucleotides might considerably deteriorate the accuracy of the phylogenetic analysis. To avoid this, some authors proposed to exclude characters containing missing data (e.g. Hufford (1992) and Smith (1996)). In contrast, Wiens (1998) argued against excluding characters and showed a benefit of "filling the holes" in a data matrix as much as possible. The popular PAUP software (Swofford (2001)) includes two methods for computing evolutionary distances between species from incomplete sequence data. The first method, called IMS ("Ignoring missing sites"), is the most commonly used strategy. It proceeds by the elimination of incomplete sites while computing evolutionary distances. According to Wiens (2003), such an approach represents a viable solution only for long sequences because of the presence of a sufficient number of known nucleotides. The second method included in PAUP, called PDMAB ("Proportional distribution of missing and ambiguous bases"), computes evolutionary distances taking into account missing bases. In this paper we propose a new method, called PEMV ("Probabilistic estimation of missing values"), which estimates the identities of all missing bases prior to computing pairwise distances between taxa. To estimate a missing base, the new method proceeds by computing a similarity score between the sequence comprising the missing base and all other sequences. A probabilistic approach is used to determine the likelihood of an unknown base to be either A, C, G or T for DNA sequences. We show how this method can be applied in the framework of Tamura-Nei evolutionary model (Tamura and Nei (1993)). This model is considered as a further extension of the Jukes-Cantor (Jukes and Cantor (1969)), Kimura 2-parameter (Kimura, (1980)), HKY (Hasegawa et al. (1985)), and F84 (Felsenstein and Churchill (1996)) models. In the next section we introduce the new method for estimating missing entries in sequence data. Then, we discuss the results provided by the methods IMS, PDMAB and PEMV in a Monte Carlo simulation study carried out with DNA sequences of various lengths, containing different percentages of missing bases.

2 Probabilistic estimation of missing values

The new method for estimating unknown bases in nucleotide sequences, PEMV, is described here in the framework of the Tamura-Nei (Tamura and Nei (1993)) model of sequence evolution. This model assumes that the equilibrium frequencies of nucleotides (π_A, π_C, π_G and π_T) are unequal and substitutions are not equally likely. Furthermore, it allows for three types of nucleotide substitutions: from purine (A or G) to purine, from pyrimidine (C or T) to pyrimidine and from purine to pyrimidine (respectively, from pyrimidine to purine). To compute the evolutionary distance between a pair of sequences within this model, the following formula is used:

$$\begin{aligned} D = & -\frac{2\pi_A\pi_G}{\pi_R} ln\left(1 - \frac{\pi_R}{2\pi_A\pi_G}P_R - \frac{1}{2\pi_R}Q\right) \\ & -\frac{2\pi_C\pi_T}{\pi_Y} ln\left(1 - \frac{\pi_Y}{2\pi_C\pi_T}P_Y - \frac{1}{2\pi_Y}Q\right) \\ & -\left(\pi_R\pi_Y - \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \frac{\pi_C\pi_T\pi_R}{\pi_Y}\right) ln\left(1 - \frac{1}{2\pi_R\pi_Y}Q\right), \end{aligned} \tag{1}$$

where P_R,P_Y and Q are respectively the transitional difference between purines, the transitional difference between pyrimidines and the transversional difference involving pyrimidine and purine; π_R and π_Y are respectively the frequencies of purines ($\pi_A + \pi_G$) and pyrimidines ($\pi_C + \pi_T$).

Assume that **C** is a matrix of aligned sequences, the base k,denoted as X, in the sequence i is missing and X is either A, C, G or T. To compute the distance between the sequence i and all other considered sequences, PEMV estimates, using Equation 2 below, the probabilities $P_{ik}(X)$, to have the nucleotide X at site k of the sequence i. The probability that an unknown base at site k of the sequence i is a specific nucleotide depends on the number of sequences having this nucleotide at this site as well as on the distance (computed ignoring the missing sites) between i and all other considered sequences having known nucleotides at site k. First, we calculate the similarity score δ between all observed sequences while ignoring missing data. For any pair of sequences, this score is equal to the number of matches between homologous nucleotides divided by the number of comparable sites.

$$P_{ik}(X) = \frac{1}{N_k}\left(\sum_{\{j|C_{jk}=X\}} \delta_{ij} + \frac{1}{3}\sum_{\{j|C_{jk}\neq X\}} (1-\delta_{ij})\right), \qquad (2)$$

where N_k is the number of known bases at site k (i.e. column k) of the considered aligned sequences, and δ_{ij} is the similarity score between the sequences i and j computed ignoring missing sites. The following theorem characterizing the probabilities $P_{ik}(A)$, $P_{ik}(C)$, $P_{ik}(G)$ and $P_{ik}(T)$, can be stated:

Theorem 1. *For any sequence i, and any site k of the matrix C, such that* C_{ik} *is a missing nucleotide, the following equality holds*: $P_{ik}(A) + P_{ik}(C) + P_{ik}(G) + P_{ik}(T) = 1$.

Due to space limitation the proof of this theorem is not presented here.

Once the different probabilities P_{ik} are obtained, we can compute for any pair of sequences i and j, the evolutionary distance using Equation 1. First, we have to calculate the nucleotide frequencies (Equation 3), the transitional differences P_R and P_Y (Equation 4), and the transversional difference Q (Equation 5). Let π_X be the new frequency of the nucleotide X:

$$\pi_X = \frac{\Lambda^i_X + \sum_{\{k|C_{ik}=?\}} P_{ik}\left(X\right) + \Lambda^j_X + \sum_{\{k|C_{jk}=?\}} P_{jk}\left(X\right)}{2L}, \qquad (3)$$

where X denotes the nucleotide A, C, G or T;Λ^i_X is the number of nucleotides X in the sequence i; symbol ? represents a missing nucleotide; L is the total number of sites compared.

$$P(i,j) = \frac{P'\left(i,j\right) + \sum_{\{k|(C_{ik}=?orC_{jk}=?)\}} P'\left(i,j,k\right)}{L}, \qquad (4)$$

$$Q(i,j) = \frac{Q'\left(i,j\right) + \sum_{\{k|(C_{ik}=?orC_{jk}=?)\}} Q'\left(i,j,k\right)}{L}, \qquad (5)$$

where $P'(i,j)$ is the number of transitions of the given type (either purine to purine P'_R, or pyrimidine to pyrimidine P'_Y) between the sequences i and j computed ignoring missing sites; $P'(i,j,k)$ is the probability of transition of the given type between the sequences i and j at site k when the nucleotide at site

k is missing either in i or in j (e.g. if the nucleotide at site k of the sequence i is A and the corresponding nucleotide in j is missing, the probability of transition between purines is the probability that the missing base of the sequence j is G, whereas the probability of transition between pyrimidines is 0); $Q'(i,j)$ is the number of transversions between i and j computed ignoring missing sites; $Q'(i,j,k)$ is the probability of transversion between i and j at site k when the nucleotide at site k is missing either in i or in j.

When both nucleotides at site k of i and j are missing, we use similar formulas as those described in Diallo et al. (2005). It is worth noting that PEMV method can be used to compute the evolutionary distance independently of the evolutionary model (Equation 6):

$$d^*_{ik} = \frac{N^c_{ij} - N^m_{ij} + \sum_{\{k|(C_{ik}=?orC_{jk}=?)\}}(1 - P^k_{ij})}{L}, \qquad (6)$$

where N^m_{ij} is the number of matches between homologous nucleotides in the sequences i and j; N^c_{ij} is the number of comparable pairs of nucleotides in i and j (i.e. when both nucleotides are known in the homologous sites of i and j); P^k_{ij} is the probability to have a pair of identical nucleotides at site k of i and j.

3 Simulation study

A Monte Carlo study has been conducted to test the ability of the new method to compute accurate distances matrices that can be used as input of distance-based methods of phylogenetic analysis. We examined how the new PEMV method performed, compared to the PAUP strategies, testing them on random phylogenetic data with different percentages of missing nucleotides. The results were obtained from simulations carried out with 1000 random binary phylogenetic trees with 16 and 24 leaves. In each case, a true tree topology, denoted T, was obtained using the random tree generation procedure proposed by Kuhner and Felsenstein (1994). The branch lengths of the true tree were computed using an exponential distribution. Following the approach of Guindon and Gascuel (2002), we added some noise to the branches of the true phylogeny to create a deviation from the molecular clock hypothesis. The source code of our tree generation program, written in C, is available at the following website: http://www.labunix.uqam.ca/˜makarenv/tree_generation.cpp.

The random trees were then submitted to the SeqGen program (Rambault and Grassly (1997)) to simulate sequence evolution along their branches. We used SeqGen to obtain the aligned sequences of the length l (with 250, 500, 750, and 1000 bases) generated according to the HKY evolutionary model (Hasegwa et al. (1985)) which is a submodel of Tamura-Nei. According to Takashi and Nei (2000), the following equilibrium nucleotide frequencies were chosen: $\pi_A = 0.15$, $\pi_C = 0.35$, $\pi_G = 0.35$, and $\pi_T = 0.15$. The transition/transversion rate was set to 4. To simulate missing data in the sequences, we used one of the two strategies described by Wiens (2003). This strategy consists of the random elimination of blocks of nucleotides of different sizes. This elimination is certainly more realistic from a biological point of view. Here, we generated data with 0 to 50% of missing bases. The obtained sequences were submitted to the three methods for computing evolutionary distances. For each distance matrix provided by IMS, PDMAB and PEMV, we inferred a phylogeny *T*' using the BioNJ algorithm (Gascuel (1997)). The phylogeny *T*' was then compared to the true phylogeny *T* using the Robinson and Foulds (1981) topological distance. The Robinson and Foulds distance between two phylogenies is the minimum number of operations, consisting of merging and splitting internal nodes, which are necessary to transform one

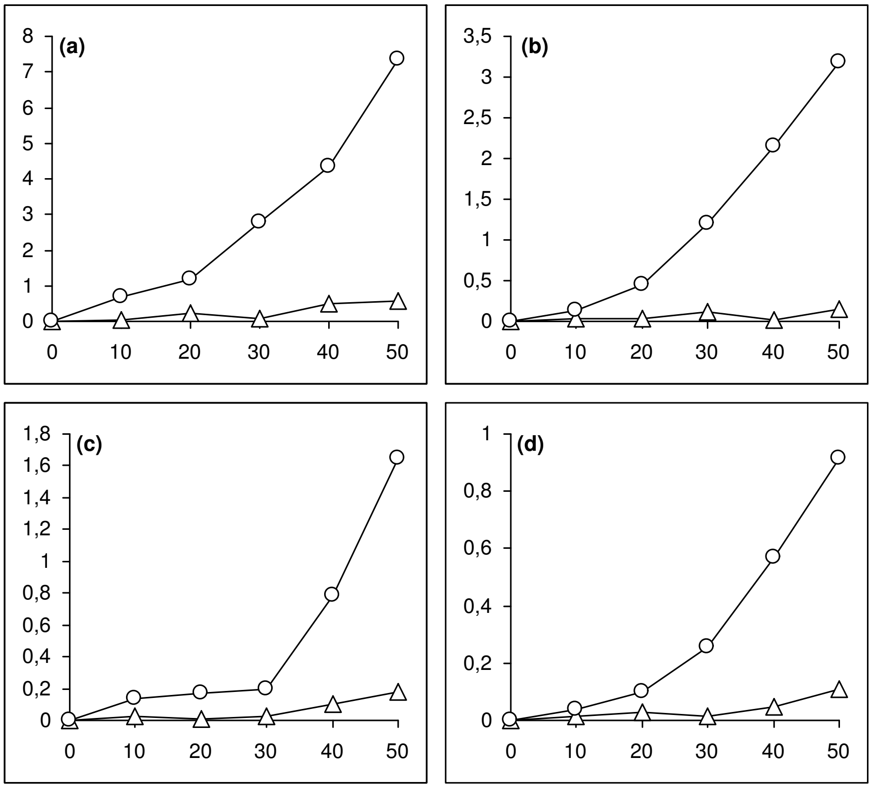

Fig. 1. Improvement in topological recovery obtained for random phylogenetic trees with 16 species. The percentage of missing bases varies from 0 to 50% (abscissa axis). The curves represent the gain (in %) against the less accurate method of PAUP. The difference was measured as the variation of the Robinson and Foulds topological distance between the less accurate method of PAUP and the most accurate method of PAUP (△) and PEMV (◯).The sequences with (a) 250 bases, (b) 500 bases, (c) 750 bases, and (d) 1000 bases are represented.

tree into another. This distance is reported as percentage of its maximum value (2n-6 for a phylogeny with n leaves). The lower this value is, the closer the obtained tree T' to the true tree T.

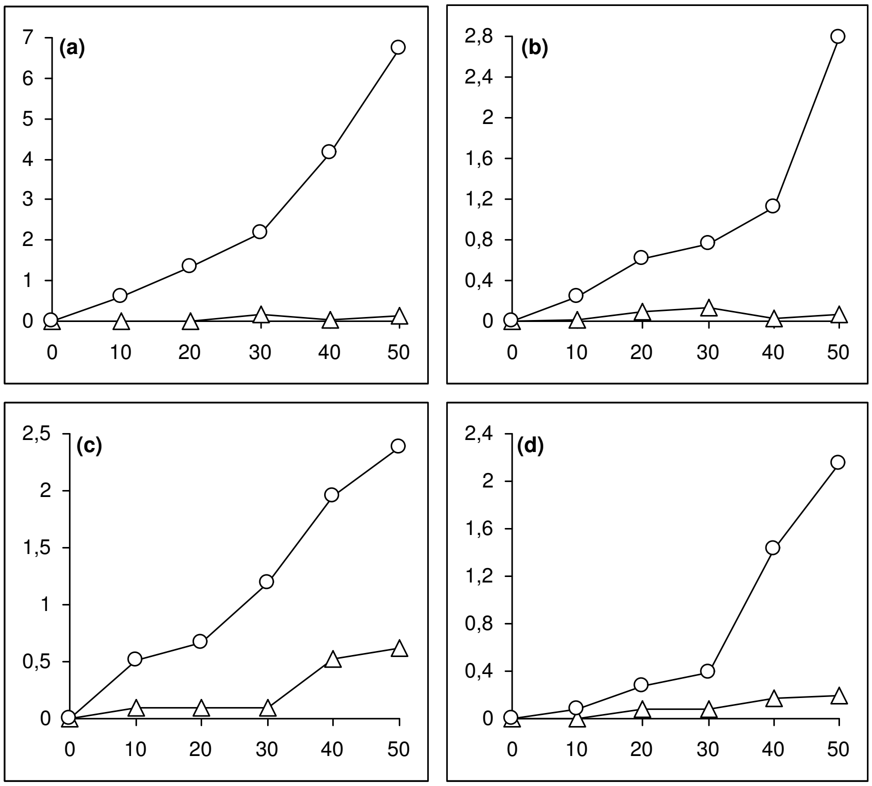

Fig. 2. Improvement in topological recovery obtained for random phylogenetic trees with 24 species. The percentage of missing bases varies from 0 to 50% (abscissa axis). The curves represent the gain (in %) against the less accurate method of PAUP. The difference was measured as the variation of the Robinson and Foulds topological distance between the less accurate method of PAUP and the most accurate method of PAUP (△) and PEMV (○). The sequences with (a) 250 bases, (b) 500 bases, (c) 750 bases, and (d) 1000 bases are represented.

For each dataset, we tested the performance of the three methods depending on the sequence length. Figures 1 and 2 present the results given by the three competing methods for the phylogenies with 16 and 24 leaves. First, for the phylogenies of both sizes PEMV clearly outperformed the PAUP methods (IMS and PDMAB) when the percentage of missing data was large (30% to 50%). Second, the results obtained with IMS were very similar to those given by PDMAB. Third, the gain obtained by our method was decreasing while the sequences length was increasing. At the same time, the following trend can be observed: the impact of missing data decreases when sequence length increases. Note that the same tendency has been pointed out by Wiens (2003).

4 Conclusion

The PEMV technique introduced in this article is a new efficient method that can be applied to infer phylogenies from nucleotide sequences comprising missing data. The simulations conducted in this study demonstrated the usefulness of PEMV in estimating missing bases prior to phylogenetic reconstruction. Tested in the framework of the Tamura-Nei model (Tamura and Nei (1993)), the PEMV method provided very promising results. The deletion of missing sites, as it is done in the IMS method, or their estimation using PDMAB (two methods available in PAUP) can remove important features of the data at hand. In this paper, we presented PEMV in the framework of the Tamura-Nei (Tamura and Nei (1993)) model which can be viewed as a generalization of the popular F84 (Felsenstein and Churchill (1996) and HKY85 (Hasegawa et al. (1985)) models. It would be interesting to extend and test this probabilistic approach within Maximum Likelihood and Maximum Parsimony models. It is also important to compare the results provided by BioNJ to those obtained using other distance-based methods of phylogenetic reconstruction, as for example, NJ (Saitou and Nei (1987)), FITCH (Felsenstein (1997)) or MW (Makarenkov and Leclerc (1999)).

References

DIALLO, Ab. B., DIALLO, Al. B. and MAKARENKOV, V. (2005): Une nouvelle mthode efficace pour l'estimation des données manquantes en vue de l'inférence phylogénétique. In: *Proceeding of the 12th meeting of Société Francophone de Classification.*Montréal, Canada, *121–125.*

FELSENSTEIN, J. and CHURCHILL, G.A. (1996): A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology Evolution, 13, 93–104.*

FELSENSTEIN, J. (1997): An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology, 46, 101–111.*

GASCUEL, O. (1997): An improved version of NJ algorithm based on a simple model of sequence Data. *Molecular Biology Evolution, 14, 685–695.*

GUINDON, S. and GASCUEL, O. (2002): Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Molecular Biology Evolution, 19, 534–543.*

HUELSENBECK, J. P. (1991): When are fossils better than existent taxa in phylogenetic analysis? *Systematic Zoology, 40, 458–469.*

HASEGAWA, M., KISHINO, H. and YANO, T.(1985): Dating the humanape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution, 22, 160–174.*

HUFFORD, L. (1992): Rosidaea and their relationships to other nonmagnoliid dicotyledons: A phylogenetic analysis using morphological and chemical data. *Annals of the Missouri Botanical Garden, 79, 218–248.*

JUKES, T. H. and CANTOR, C. (1969): Mammalian Protein Metabolism, chapter Evolution of protein molecules. *Academic Press, New York, 21–132.*

KIMURA, M. (1980): A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequence. *Journal of Molecular Evolution, 16, 111–120.*

KUHNER, M. and FELSENSTEIN. J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology Evolution, 11, 459–468.*

MAKARENKOV, V. and LECLERC, B. (1999): An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion. *Journal of Classification, 16, 3–26.*

MAKARENKOV, V. and LAPOINTE, F-J. (2004): A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics, 20, 2113–2121.*

RAMBAULT, A. and GRASSLY, N. (1997): SeqGen: An application for the Monte Carlo simulation of DNA sequences evolution along phylogenetic trees. *Bioinformatics, 13, 235–238.*

ROBINSON, D. and FOULDS, L. (1981): Comparison of phylogenetic trees. *Mathematical Biosciences, 53, 131–147.*

SAITOU, N. and NEI, M.(1987): The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology Evolution, 4, 406–425.*

SANDERSON, M.J., PURVIS, A. and HENZE, C. (1998): Phylogenetic supertrees: Assembing the tree of life. *Trends in Ecology and Evolution, 13, 105–109.*

SMITH, J.F.(1997): Tribal relationships within Gesneriaceae: A cladistic analysis of morphological data. *Systematic Botanic, 21, 497–513.*

SWOFFORD, D. L. (2001): PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. *Sinauer Associates*, Sunderland, Massachusetts.

TAKAHASHI, K. and NEI, M. (2000): Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution, 17, 1251–1258.*

TAMURA, N. and NEI, M. (1993): Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution, 10/3, 512–526.*

WIENS, J. J. (1998): Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology, 52, 528–538.*

WIENS, J. J. (2003): Does adding characters with missing data increase or decrease phylogenetic accuracy. *Systematic Biology, 47, 625–640.*