Network Models for Sequence Evolution

Arndt von Haeseler[1]
Department of Zoology,
University of Munich,
Luisenstr. 14,
W-8000 Munich 2,
Germany.

Gary A. Churchill
Biometrics Unit
Cornell University
Ithaca, NY 14853
USA.

BU-1152-MA

August 1992

[1] To whom correspondence should be addressed

Abstract

We introduce a general class of models for sequence evolution that includes network phylogenies. Networks, a generalization of strictly tree–like phylogenies, are proposed to model situations where multiple lineages contribute to the observed sequences. An algorithm to compute the probability distribution of binary character state configurations is presented and statistical inference for this model is developed in a likelihood framework. A stepwise procedure based on likelihood ratios is used to explore the space of models. Starting with a star phylogeny new splits (non–trivial bipartitions of the sequence set) are successively added to the model until no significant change in the likelihood is observed. A novel feature of our approach is that the new splits are not necessarily constrained to be consistent with a tree–like mode of evolution. The fraction of invariable sites is estimated by maximum likelihood simultaneously with other model parameters and is essential to obtain a good fit to the data. The effect of finite sequence length on the inference methods is discussed. Finally, we provide an illustrative example using aligned VP1–genes from the Foot and Mouth Disease viruses (FMDV). The different serotypes of the FMDV exhibit a range of tree–like and network evolutionary relationships.

keywords convergent evolution — gene conversion — maximum likelihood — phylogeny — quasi–species — recombination — substitution rate — virus evolution

1 Introduction

The inference of phylogenetic relationships from molecular sequence data has received a great deal of attention in recent years and a variety of methods have been developed (*cf.* Swofford and Olson 1990, Felsenstein 1988). The methods presented here generalize the standard tree–like models of evolution by allowing for network relationships. Phylogenies among (closely) related sequences may be complicated by recombination, gene conversion or other horizontal tranfers of genetic information and thus can be viewed as a mixture of different tree–like histories. The phylogenetic relationships among RNA viruses, often described as quasi species (Steinhauer and Holland 1987, Domingo and Holland 1988), provide an example where tree–like phylogenies are not necessarily appropriate. Dopazo *et al.* (1990) used a non–statistical split decomposition method (Bandelt and Dress 1990) to display a network relationship among viral sequences. The present work provides a statistical basis for determining the significance of a split decomposition.

Cavender (1978) introduced a model of an evolutionary process on a two letter alphabet. We will adopt essentially this model with some extensions to be described below. Although nucleic acid sequences are formed from four nucleotides it has been helpful to study binary sequences to investigate the reliability of tree reconstruction methods (Churchill *et al.* 1991, Hendy and

Penny 1991). A generalization of the method presented here is possible, but for the sake of clarity we will present only the two letter case. This restriction is not too limiting when transition frequencies are substantially greater than transversion frequencies. In this case, the observed transition differences may contain very little phylogenetic information (Brown *et al.* 1982). On the other hand Ward *et al.* (1991) describe a collection of human mitochondrial DNA sequences in which only transition differences are observed.

The existence of invariable sites in amino–acid sequences was first proposed by Fitch and Margoliash (1967) and was later extended to include nucleic–acid sequences (Fitch 1986a,b). In a recent study, Shoemaker and Fitch (1989) found evidence of invariable nucleotide sites in nearly every data set they analyzed. We will extend the Cavender (1978) model by allowing a fraction of the sites to be invariable. The fraction of invariable sites is estimated by maximum likelihood in the context of the current model. Hasegawa *et al.* (1985) and Hasegawa and Kishino (1989) have proposed an alternative method of estimating the fraction of invariable sites and ignore this fraction in their tree building procedure. We demonstrate that the estimated fraction of invariable sites can be sensitive to the topology considered.

In principle, to determine which model (models) are most suitable for the given data, all possible models should be compared. Because the number of possible trees grows exponentially with the number of species, heuristic search algorithms are often used. The single tree with the highest likelihood among all trees considered is usually chosen to be the best tree and reported. We describe a sequential search method based on likelihood ratios that is analogous to stepwise procedures used in linear model building (Draper and Smith, 1981). The search may be restricted to tree–like topologies or may encompass the larger class of evolutionary network models.

Most tree reconstruction methods start with a fully resolved tree, *i.e.* a tree with vertices of degree one and three only, and search the space of alternative fully resolved trees by varying the topolcgy and comparing likelihoods maximized with respect to branch lengths. By looking only at fully resolved trees, one assumes a precision that may not be supported by the data. Methods for obtaining degenerate "consensus trees" using bootstrap resampling have been described (Felsenstein 1989, Swofford 1991). However the reported trees do not correspond to a single model which has been fit to the data and tested. Stepwise procedures that start with a degenerate tree have also been proposed (Saitou 1988). However these procedures are designed to stop only when a fully resolved tree is obtained. The forward stepwise procedure descibed here overcomes these problems by sequentially adding splits to the model until significant improvement is no longer possible. In many cases, the resulting trees may be less than fully resolved. The problem remains that there may be several models which provide a good fit to the data. These may be revealed by a broader search of the model space.

2 Methods

2.1 The Model

Suppose we have a set of K binary sequences each of length N. We will assume that the sequences are correctly aligned with homologous sites forming a column. The data are represented as a matrix with elements s_{ij} from a binary alphabet, coded as 0 and 1, where i is the species index and j is the site index. Each site j is represented by a K–dimensional vector of zeroes and ones, a binary character–state configuration. We will assume that the evolutionary process acts independently at each site and that the sites are identically distributed. We will also assume symmetry between zero to one and one to zero changes. Thus, two configurations are equivalent if the first is obtained from the second by reversing the labeling of zeroes and ones.

A split is a bipartition of the K species into two disjoint sets. A split can be represented by the set S, such that $1 \in S$, and the set $S^c = \{1, \ldots, K\} \setminus S$. Our definition includes the trivial split where $S = \{1, 2, \ldots, K\}$, and $S^c = \emptyset$.

Each of the 2^{K-1} non–equivalent configurations corresponds to a split on the species. The 2^{K-1} dimensional vector $\mathbf{X} = (x_0, x_1, \ldots, x_{2^{K-1}-1})$ counts the number of occurences of each non–equivalent configuration. A split is called non–zero split if $x_i > 0$, i.e. there is at least one site in the data that supports the split.

An evolutionary model $M = (\mathcal{S}, \theta)$ is defined by a set of non–trivial splits and a parameter vector $\theta = \{\theta_S | S \in \mathcal{S}\}$, where θ_S is the probability that, at a given site, an odd number of substitutions separate the sequences in S from those in S^c. Assuming that the substitutions occur as a Markov process in time, the valid range of θ_S is from 0 to 1/2. The parameter vector θ is analogous to Hendy and Penny's (1989) $\mathbf{p}$ vector, the probability of character changes along an edge of a tree (Hendy 1989, Hendy and Penny 1989)

Tree phylogenies are a special case of our general evolutionary model. A model is representable as a tree if and only if all splits in $\mathcal{S}$ are pairwise compatible. Splits S_i and S_j are pairwise compatible if there exists a $B_i \in \{S_i, S_i^c\}$ and a $B_j \in \{S_j, S_j^c\}$ with $B_i \cap B_j = \emptyset$ (Buneman 1971). If the set of splits is not pairwise compatible the evolutionary model is called a network. For example, consider the model M_1, with $\mathcal{S} = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \{1\}, S_2 = \{1,3,4\}, S_3 = \{1,2,4\}, S_4 = \{1,2,3\}$. This is the familiar four species star phylogeny (figure 1a). A new model M_2 is created by adding the split $S_5 = \{1, 2\}$. This is the four species tree that pairs species 1 and 2 versus species 3 and 4 (figure 1b). If we now obtain a third model M_3 by adding the split $S_6 = \{1, 3\}$ the result is a network (figure 1c). The splits S_5 and S_6 are not compatible.

2.2 Computing the Spectrum of M

The observed character state configuration at a site is the result of substitutions that have occurred along one or more of the splits in the model. For example, if no substitution occurs along any split in model M_2 above, the resulting character state configuration will be constant, 0000 or 1111. If substitutions occur along the splits S_1 and S_5, the resulting character state configuration will be 0100 or the equivalent configuration 1011. For a given model M, the probability vector $\mathbf{P} \equiv \mathbf{P}(M) = (p_0(\theta), \ldots, p_{2^{K-1}-1}(\theta))$ defines the character state probability distribution. Following Hendy and Penny (1991) $\mathbf{P}$ is called spectrum of M. In this section we describe an algorithm to compute the spectrum.

For a model $M = (\mathcal{S}, \theta)$ let $S_1, S_2, \ldots, S_m$ be the sets in $\mathcal{S}$ and let $\theta_1, \ldots, \theta_m$ be the corresponding probabilities of substitution across a split. Consider one site in the sequence set. Since for each split in the model a substitution may occur or not, there are 2^m combinations of possible substitutions. Each of these combinations generates a character state configuration. Let $\mathbf{I} = (I_1, I_2, \ldots, I_m)$ be a vector of indicators such that $I_i = 1$ if an odd number of substitution events occur along split S_i and $I_i = 0$ otherwise. For example $\mathbf{I} = (0,0,0,0,0)$ indicates that at the site considered, no substitution occurred along any split in the model M_2. $\mathbf{I} = (1,0,0,0,1)$ corresponds to the example given above. For any given $\mathbf{I}$ the resulting character state configuration $\mathbf{c} = (c_1, c_2, \ldots, c_K)$, where $c_i \in \{0,1\}$, is calculated according to the following formula:

$$c_j = \sum_{\nu=1}^{m} I_\nu |S_\nu \cap \{j\}| \pmod 2 \tag{1}$$

This expression counts occurrences of a species j in the sets S_ν whose indicator value I_ν equals 1. If this number is odd the character state of species j equals 1, otherwise 0. For a proof of formula 1 see Appendix .

The probability of an indicator vector $\mathbf{I}$ is

$$\Pr(\mathbf{I}) = \prod_{\substack{\nu=1 \\ I_\nu=1}}^{m} \theta_\nu \cdot \prod_{\substack{\nu=1 \\ I_\nu=0}}^{m} (1-\theta_\nu). \tag{2}$$

Summing over all indicator vectors giving rise to $\mathbf{c}$ or its equivalent configuration $\mathbf{c}'$ yields the total probability of the corresponding probability p_i in the spectrum of the model M,

$$p_i = \sum_{\mathbf{I}} \Pr(\mathbf{c} \mid \mathbf{I}) \Pr(\mathbf{I}) + \Pr(\mathbf{c}' \mid \mathbf{I}) \Pr(\mathbf{I}). \tag{3}$$

Where $\Pr(\mathbf{c} \mid \mathbf{I})$ is 1 for the configuration generated by $\mathbf{I}$ and zero otherwise. The computation of this probability for each of the 2^{K-1} different configurations is easily implemented on a computer to define the spectrum for any given evolutionary model.

2.3 Invariable Sites

To include invariable sites in our model, we introduce a parameter θ_0 which is the fraction of invariable sites among all sites in the aligned sequences. This fraction is allowed to vary between 0 and 1. The 2^{K-1} dimensional spectral vector $\mathbf{Q}$ is defined as follows

$$\begin{aligned} q_0 &\equiv \theta_0 + (1-\theta_0)\cdot p_0 \qquad (4) \\ q_i &\equiv (1-\theta_0)\cdot p_i, \quad \text{where } i = 1,\ldots,2^{K-1}-1 \qquad (5) \end{aligned}$$

where p_0 is the probability of the constant configuration. The quantities $p_i(\theta)$ in the log–likelihood (equation 6, below) are replaced by $q_i(\theta)$. The invariable sites fraction is estimated by maximum likelihood simultaneously with the other model parameters.

2.4 Parameter Estimation

To estimate the parameter θ for a given model, we maximize the log–likelihood function

$$\ell(M,\theta) = \sum_{i=0}^{2^{K-1}-1} x_i \cdot \ln(p_i(\theta)). \qquad (6)$$

The multinomial form of the log–likelihood follows from the independent and identically distributed sites assumptions above. Once the spectrum of the model has been defined as a function of its parameters, the estimates $\hat{\theta}$ are computed by maximization of $\ell(M, \theta)$. This can be achieved by a standard optimization procedure such as the Newton–Raphson method. The required partial derivatives of the spectrum with respect to θ are readily obtained. Calculating $\hat{\theta}$ for tree–like evolutionary models in the context of a four letter alphabet has been proposed by various authors (Barry and Hartigan 1987, Felsenstein 1981).

2.5 The Search Algorithm

The problem of selecting a suitable model remains. Our approach is different from most of the existing approaches in that instead of starting with a fully resolved tree structure (*e.g.* trees with vertex degree of three and one only) we start with a star phylogeny (*i.e.*, a phylogeny which includes all singleton splits, i.e. $S_1 = \{1\}$, and $S_i = \{1, \ldots, K\} \setminus \{i\}$, $i = 2, \ldots, K$). By including all of the singleton splits in the model, we ensure that all configurations have non–zero probability. Splits are then added successively to the model to resolve the phylogenetic relationships revealed by the data. If the current model is M_1

and the model M_2 is obtained by adding a split to M_1, the statistic

$$\Lambda(M_1, M_2) = 2(\ell(M_2, \hat{\theta}) - \ell(M_1, \tilde{\theta})) \tag{7}$$

will have (asymptotically) a χ_1^2 distribution with 1 degree of freedom and large values indicates a significant improvement in the fit of the model. In the stepwise forward search, a split is added to a current model only if the likelihood ratio statistic indicates a significant improvement in the goodness–of–fit. The procedure tests each possible split and stops adding to the model when there are no longer any splits that yield a significant improvement. If there is more than one alternative, the most significant split is used to extend the model. The search may be restricted to splits which are compatible with tree–like phylogeny or the set of all splits may be searched to find potential network structures. The search space can be further reduced by considering only non–zero splits. Although it is possible to include unsupported splits in the model, we have found that their estimated values are typically zero or very small.

The choice of an appropriate critical value for stopping is somewhat arbitrary and is complicated by the fact that the asymptotic χ_1^2 distribution is only an approximation to the true (small sample) distribution of $\Lambda(M_1, M_2)$. In the examples below we elected to use the critical value $\chi_{1,\alpha}^2 = 3.84$, where $\alpha = 0.05$. Figure 2 illustrates the stepwise forward procedure.

A backward step can be implemented using the same methods. Each split in the current model is tested one at a time by dropping it from the model. The split which makes the least significant change is removed from the current model. In a fully backward stepwise procedure, the starting point is a saturated model and splits are dropped until all remaining splits give a likelihood-ratio statistic which exceeds the critical value. A mixed strategy of forward and backward steps can be implemented to give a broader range to the search.

It is not possible for us to claim any optimality properties for these heuristic search procedures. However, similar methods have been widely used in linear and log–linear building (Draper and Smith 1981, Chap. 6). Although stepwise procedures can be informative regarding the relative importance of major features of the data, we recommend a more general search of the model space.

2.6 A Simulation Test

The χ^2 distribution of the likelihood–ratio statistics is an approximation which is valid when each of the configuration counts is large. The number of possible configurations grows exponentially with the number of species. Unless the sequence lengths are several times 2^{K-1} these counts will typically be sparse and the asymptotics of the likelihood ratio test may be unreliable. The following procedure is suggested to overcome this problem.

Random sequences can be generated according to the smaller model $M_1 = (\mathcal{S}, \theta)$, using the maximum likelihood parameter value $\hat{\theta}$, to obtain a simulated vector $\mathbf{X_r}$ of configuration counts. The likelihood ratio statistic in equation (7) can then be calculated for a large number of simulated data sets to obtain an empirical estimate of the sampling distribution. In particular the $100(1-\alpha)\%$ critical values of the distribution can be estimated.

3 The data

All sequences are VP-1 gene sequences from the Foot–and–Mouth Disease Virus (FMDV). The virus is divided in several subserotypes. The phylogenetic relationships with the subserotypes A, C, and O were investigated. The following aligned sequences were used:

A types: A10/61, A12/32, A27/76, A5Mor/83, A5Sp/86, and A5Ww/51.

C types: C3Ind/78, C3Ind/71, C3Arg/85, C3Res/55, C3Arg/84, CS10/79, and CS16/81 (Piccone *et al.* 1988, Sobrino *et al.* 1986).

O types: O2Norm/47, OMu/82, O1Bfs/68, O1Ca/58, Oth/81, OWupp/82, and OIsr/81 (Beck and Strohmaier 1987).

Sequences were translated into the two letter alphabet of purines and pyrimidines. Positions where a gap occured in at least one of the sequences were ignored.

4 Results

Figures 3, 4, and 5 show the final models obtained using the forward stepwise procedure with the FMDV sequence alignments. The results (almost) agree with the networks shown in Dopazo *et al.* (1990). nly the O–type sequences show a tree–like evolutionary relationship, whereas A– and C–type sequences are related by a network. However, there is a remarkable difference between A– and C–type sequences. The C–type sequences contain a considerably large amount of tree–likeness. It is instructive to look at the order in which splits were introduced in the stepwise forward model. The first three splits added to the star phylogeny model of C–type sequences are compatible with a tree structure. They account for most of the improvement in the goodness–of–fit. Only near the end of the stepwise forward procedure are non–compatible splits introduced. For the A–types, the second split added to the existing model is not compatible with the first split. Hence, the network relationship is a prominent feature of the A–type sequences.

When invariable sites are included in the model, we find that the complexity of the inferred evolutionary network is reduced. The inferred tree for O–type sequences remains unchanged. For the C–type (figure 4b) sequences, the network is replaced by a not fully resolved tree. Only three splits are added to the original star tree of the C–type. Whereas for the A–type sequences a simpler network is obtained (figure 3b).

The number of splits in a model influences the estimated fraction of invariable sites. Table 1 shows the estimates for the three FMDV serotypes. If the star phylogeny model is assumed, the proportion of invariable size is maximal and ranges from 84 % to 94 %. The introduction of splits monotonically decreases this fraction. When all non–zero splits are included in the model the proportion of invariable sites equals zero. This occurs because the model is saturated and places no constraints on the expected counts. The * in table 1 indicate the proportion of invariable sites for the models shown in figures 3–5. About 75 %, 85 %, and 45 % of the sites are invariable in the A–type, C–type, and O–type sequences respectively.

The introduction of invariable sites also affects the estimates $\hat{\theta}$, which now reflect the rates of change among the variable sites. The ratio of estimated $\theta_i, i \geq 1$ parameters for models with and without invariable sites indicates a 2 (O–type), 4 (A–type), and a 6.5 (C–type) fold increase in the estimated θ when invariable sites are allowed. The increase is approximately

$$\theta(\text{invariable sites}) \approx \frac{\theta(\text{no invariable sites})}{1-\theta_0}. \qquad (8)$$

This empirical result is consistent with the altered defintion of $\theta_i, i \geq 1$ when invariable sites are included in the model.

In order to check the effects of the asymptotic approximation to the likelihood ratio test on our results, we reanalyzed the data using the simulation test described in the method section. As illustrated in figure 6 the results agree in their major features. The simulation based procedure is slightly more conservative and tends to introduce fewer splits. The O– and C–type sequences conform to tree topologies, whereas the A–type sequences still form a network.

5 Discussion

The methods described here belong to a large class of likelihood based procedures used to compute phylogenetic relationships among sequences (e.g. Felsenstein 1981, Barry and Hartigan 1987). This approach generalizes other methods by allowing the addition of non–compatible splits and thus is not restricted to tree–like phylogenies. For viral sequences, the quasi–species model of evolution provides an explanation for the network (Steinhauer and Holland 1987, Domingo and Holland 1988). For sequences which should display a tree–like relationship, a network phylogeny may indicate recombination events between different lineages, convergent evolution or horizontal transfer of genetic information. These situations violate the usual assumption that evolutionary events along different lineages are independent (see Navidi et al. 1991 assumption 3, also Barry and Hartigan 1987 and Felsenstein 1988). Thus, network models provide a diagnostic to detect violations of this assumption. It may be worthwhile to reexamine the sequences looking for contingous subsequences which could have arisen from recombination events (see Sawyer 1989).

Another possible cause for an inferred network is variation in substitution rates between sites. When rate heterogeneity is present, estimated rates assuming a homogeneous model will be an average of the actual rates. Sites with a high rate of change are likely to show an apparent excess of parallel changes in independent lineages. It is often instructive to look at the development of an evolutionary model in the stepwise procedure. If the true relationship is a tree but rate heterogeneity is present, the first splits added to the star phylogeny are usually pairwise compatible. Non-compatible splits are added only at the later stages to adjust for the parallel changes. The introduction of invariable sites is one step towards a realistic evolutionary model with rate variation. However, further work on heterogeneity of rates is needed to obtain fully satisfactory solution.

An important feature of the sequential approach to model building is that it allows the inferred structure to be less than fully resolved. The stepwise forward procedure, although intuitive and easily implemented in an automated system, may not give a complete picture of the range of plausible alternative models and should not be strictly applied. A more comprehensive exploration of the model space is recommended.

The methods described here are computationally expensive. The size of the spectrum grows exponentially in K. It may be possible to reduce this computation by focusing only on components of the spectrum that correspond to non–zero splits. The Newton–Raphson optimization of the loglikelihood is also slow. With the current implementation, network models for up to 8 or 10 sequences can be explored in reasonable amounts of time. As additional sequences are added, the asymptotic approximations for the likelihood ratio test become less reliable. If the simulation based test must be used to assess significance, the computational expense will increase substantially.

6 Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft, the National Science Foundation, the National Institutes of Health, and the University of Southern California Faculty and Innovation Fund. We would like to thank William Navidi, Simon Tavaré, and Michael Waterman for helpful suggestions and interesting discussions. We also thank Joaquin Dopazo for providing us with the FMDV sequences. Parts of this work were completed while one of us (A.v.H.) was staying at the Mathematics Department of the University of Southern California at Los Angeles.

7 Appendix: Proof of formula (1)

Fix an evolutionary model M with m splits. Given an indicator vector $\mathbf{I}$ we need to prove that the resulting character state configuration $\mathbf{c}$ is given by formula (1).

Let $\mu(\mathbf{I}) = \sum_{\nu=1}^{m} I_\nu$ be the number of ones of a given indicator vector. If $\mu(\mathbf{I}) = 1$, then $c_j = 1$ for each $j \in S_\nu$, since there is only one split $S_\nu, (1 \leq \nu \leq m)$ with indicator value equal to one. This proves formula (1) for the case $\mu(\mathbf{I}) = 1$. Assume that $\mathbf{c}$ has been calculated for every $\mathbf{I}$ with $\mu(\mathbf{I}) = k, 1 \leq k \leq m - 1$. For each $\mathbf{I}'$ with $\mu(\mathbf{I}) = k + 1$ exists an $\mathbf{I}$ with $\mu(\mathbf{I}) = k$ such that

$\mathbf{I}'$ has exactly one component *e.g.* $I_{\nu'}$ which is one in the indicator vector of $\mathbf{I}'$ but not in $\mathbf{I}$. Hence,

$$c_j(\mathbf{I}') = \sum_{\nu=1}^{m} I'_\nu |S_\nu \cap \{j\}| \pmod 2 \tag{9}$$

$$= (|S_{\nu'} \cap \{j\}| + \sum_{\substack{\nu=1 \\ I_\nu=1}}^{m} |S_\nu \cap \{j\}|) \pmod 2 \tag{10}$$

$$= c_j(\mathbf{I}) + |S_{\nu'} \cap \{j\}| \pmod 2 \tag{11}$$

We define

$$\sigma(\mathbf{I}) \equiv \{j | c_j(\mathbf{I}) = c_1(\mathbf{I})\} \text{ and } \overline{\sigma(\mathbf{I})} \equiv \{j | c_j(\mathbf{I}) \neq c_1(\mathbf{I})\} \tag{12}$$

$\sigma(\mathbf{I})$ is the set of all species having the same character state like species 1, given the indicator vector $\mathbf{I}$. In $\overline{\sigma(\mathbf{I})}$ the remaining sequences are collected. W.l.o.g we assume that all species in the set $S_{\nu'}$ change from their current character state to the second state. We can decompose $\sigma(\mathbf{I})$ and $\overline{\sigma(\mathbf{I})}$ as follows

$$\sigma(\mathbf{I}) = A_1 \cup B_1, \overline{\sigma(\mathbf{I})} = A_2 \cup B_2, \tag{13}$$

where

$$A_1 \equiv \sigma(\mathbf{I}) \cap S_{\nu'} \text{ and } B_1 \equiv \sigma(\mathbf{I}) - A_1, \tag{14}$$

and

$$A_2 \equiv \overline{\sigma(\mathbf{I})} \cap S_{\nu'} \text{ and } B_2 \equiv \overline{\sigma(\mathbf{I})} - A_2. \tag{15}$$

The species in A_1 and A_2 have different character states. Since they are also elements of $S_{\nu'}$ they change their character states concomitantly. Whereas the character states of the species in the B–sets remain unchanged. The new character state configuration is now given by

$$\sigma(\mathbf{I}') = A_2 \cup B_1 \text{ and } \overline{\sigma(\mathbf{I})} = A_1 \cup B_2. \tag{16}$$

If j is an element of $S_{\nu'}$ and has character state $c_j(\mathbf{I})$ then the new state is

$$c_j(\mathbf{I}') = c_j(\mathbf{I}) + 1 \pmod 2 = c_j(\mathbf{I}) + |S_{\nu'} \cap \{j\}| \pmod 2. \tag{17}$$

If j is not in $S_{\nu'}$, then its character state does not change. Hence,

$$c_j(\mathbf{I}') = c_j(\mathbf{I}) + 0 = c_j(\mathbf{I}) + |S_{\nu'} \cap \{j\}| \pmod 2. \tag{18}$$

This proves formula 1.

References

[1] Bandelt HJ, Dress AWM (1990) A canonical decomposition theory for metrics on a finite set. Preprint 90–032, SFB 343, Universität Bielefeld.

[2] Barry D, Hartigan JA (1987) Statistical analysis of hominoid molecular evolution. Stat. Sci. 2:191–210.

[3] Beck E, Strohmaier K (1987) Subtyping of european FMDV outbreak strains by nucleotide sequence determination. J. Virol. 61:1621–1629.

[4] Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18:225–239.

[5] Buneman P (1971) The recovery of trees from measures of dissimilarity. In: Hodson FR, Kendall DG, Tantu P (eds) Mathematics in the archaeological and historical science. Proc. of the Anglo–Romanian–Conference 1970. Univ. Press Edinburgh. pp 387–395.

[6] Cavender J (1978) Taxonomy with confidence. Math. Biosc. 40:271–280.

[7] Churchill GA, von Haeseler A, Navidi WC (1992) Sample size for a phylogenetic inference. Mol. Biol. Evol. 9:753–769.

[8] Dopazo J, Dress A, von Haeseler A (1990) Split decomposition: a new technique to analyze viral evolution. Preprint 90–037, Sonderforschungsbereich 343 Diskrete Strukturen in der Mathematik. Universität Bielefeld.

[9] Domingo E, Holland JJ (1988) High error rates, population equilibrium and evolution of RNA replication systems. In: Domingo E, Holland JJ, Ahlquist P (eds). RNA genetics, vol III. CRC Press, Boca Raton, Florida, pp 3–36.

[10] Draper NR, Smith H (1981) *Applied Regression Analysis* 2nd ed. New York: John Wiley.

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

[12] Felsenstein J (1988) Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22:521–565.

[13] Felsenstein J (1989) Phylip manual, Version 3.2 University Herbarium of the University of California at Berkeley.

[14] Fitch WM (1986a) The estimate of total nucleotide substitutions from pairwise difference is biased. Philos. Trans. R. Soc. Lond.[B] 312:317–324.

[15] Fitch WM (1986b) An estimation of the number of invariable sites is necessary for the accurate estimation of nucleotide substitutions since a common ancestor. In: Gershowitz H (ed) Evolutionary perspectives and the new Genetics. Alan R. Liss, New York, pp 149–159.

[16] Fitch WM, Margoliash E (1967) A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. Biochem. Genet. 1:65–71.

[17] Hasegawa M, Kishino H (1989) Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial DNA sequences. Evolution 43:672–677.

[18] Hasegawa M, Kishino H, Yano K (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol 22:160–174.

[19] Hendy MD (1989) The relationship between simple evolutionary tree models and observable sequence data. Syst. Zool. 38:310–321.

[20] Hendy MD, Penny D (1991) Spectral analysis of phylogenetic data. (preprint)

[21] Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

[22] Navidi WC, Churchill GA, von Haeseler A (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. 8:128–143.

[23] Piccone ME, Kaplan G, Giavedoni L, Domingo E, Palma EL (1988) VP1 of serotype C foot–and–mouth disease virus: long–term conservation of sequences. J. Virol. 62:1469–1473.

[24] Saitou N (1988) Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. 27:261–273.

[25] Sawyer S (1989) Statistical test for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

[26] Shoemaker JS, Fitch WM (1989) Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270–289.

[27] Sobrino F, Palma EL, Beck E, Davila M, de la Torre JC, Negro P, Villaneuva N, Ortin J, Domingo E (1986) Fixation of mutations in the viral genome during an outbreak of foot–and–mouth disease: heterogennity and rate variations. Gene 50:149–159.

[28] Steinhauer DA, Holland JJ (1987) Rapid evolution of RNA viruses. Ann. Rev. Microbiol. 41:409–433.

[29] Swofford DL, Olsen GJ (1990) Phylogeny reconstruction. In: Hillis DM and Moritz C (eds) Molecular Systematics. Sinauer Associates Inc., Sunderland MA. pp. 411–501.

[30] Swofford DL (1991) PAUP 3.0 user's manual (Draft 2.9.91). Illinois Natural History Survey, Champaign, 1991.

[31] Ward RH, Frazer BS, Dew K, Pääbo S (1991) A single north–american tribal group contains extensive mitochondrial diversity. Proc. Natl. Acad. Sci. USA 88:8720–8724.

Tables

Table 1: **Estimated proportion of invariable sites**

number of splits	A–type	C–type	O–type
0	87	94	84
1	86	94	80
2	84	93	74
3	75 *	85 *	56
4	56	79	45 *
5	0	64	26
6	—	0	0

*Percentage of invariable sites among the constant sites in three subserotype families of FMDV VP1–gene. * indicates the corresponding fraction for the best model, computed by the stepwise forward procedure.*

Figure Captions

Figure 1: Successive addition of splits to an evolutionary model. Starting with a star phylogeny (a) the addition of split $S_5 = \{1,2\}$ creates the well known binary four species tree (b). The introduction of $S_5 = \{1,3\}$ produces the network shown in (c).

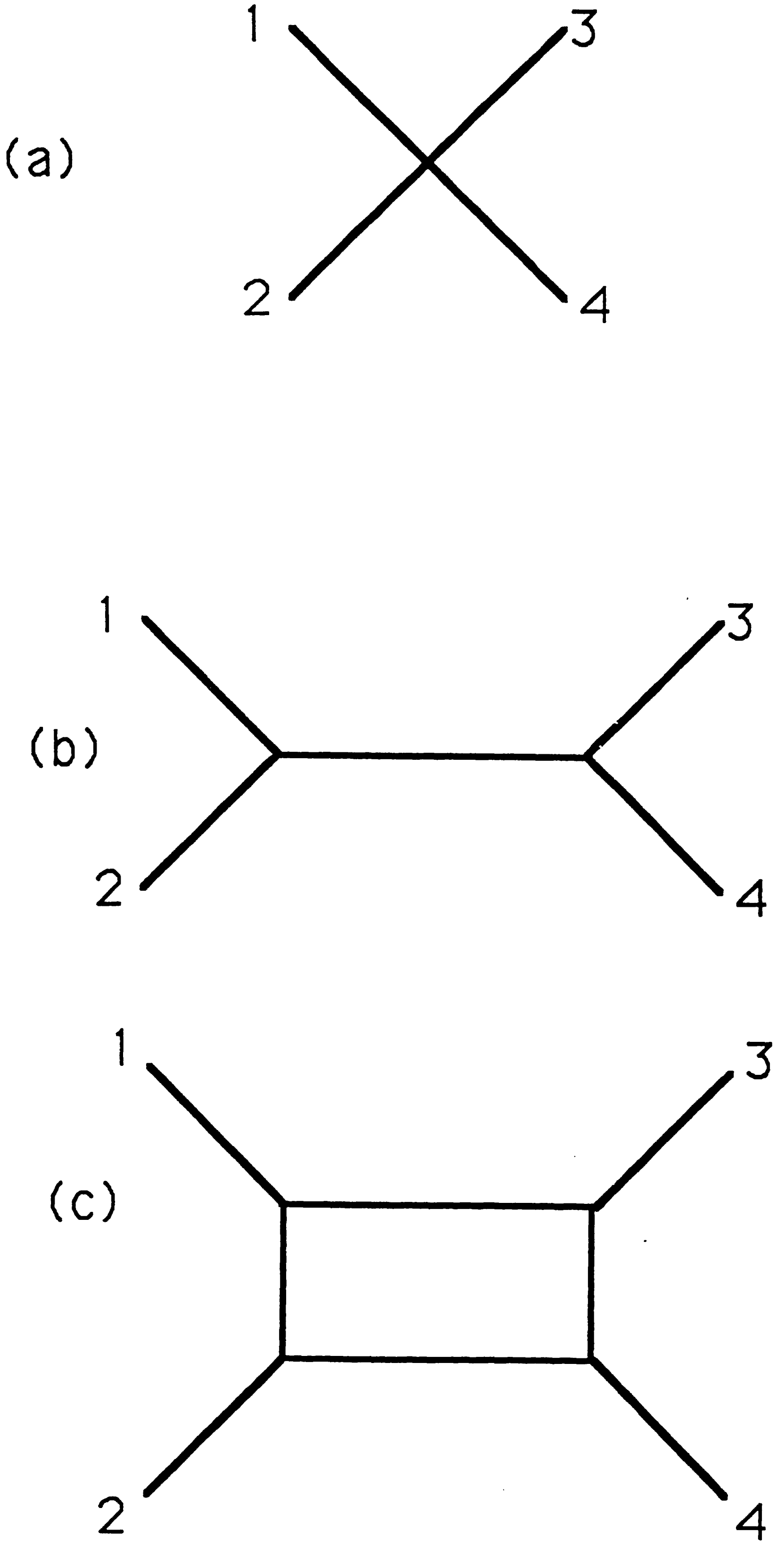
Figure 2: Successive addition of splits to an evolutionary model. Starting with a star phylogeny each non–zero split is tested by adding it to the current model. The split that produces the most significant improvement is added to establish a new model and the procedure is repeated. This example stops with a fully resolved tree for the O–serotype FMD virus sequences. Edge lengths are proportional to the substitution rates

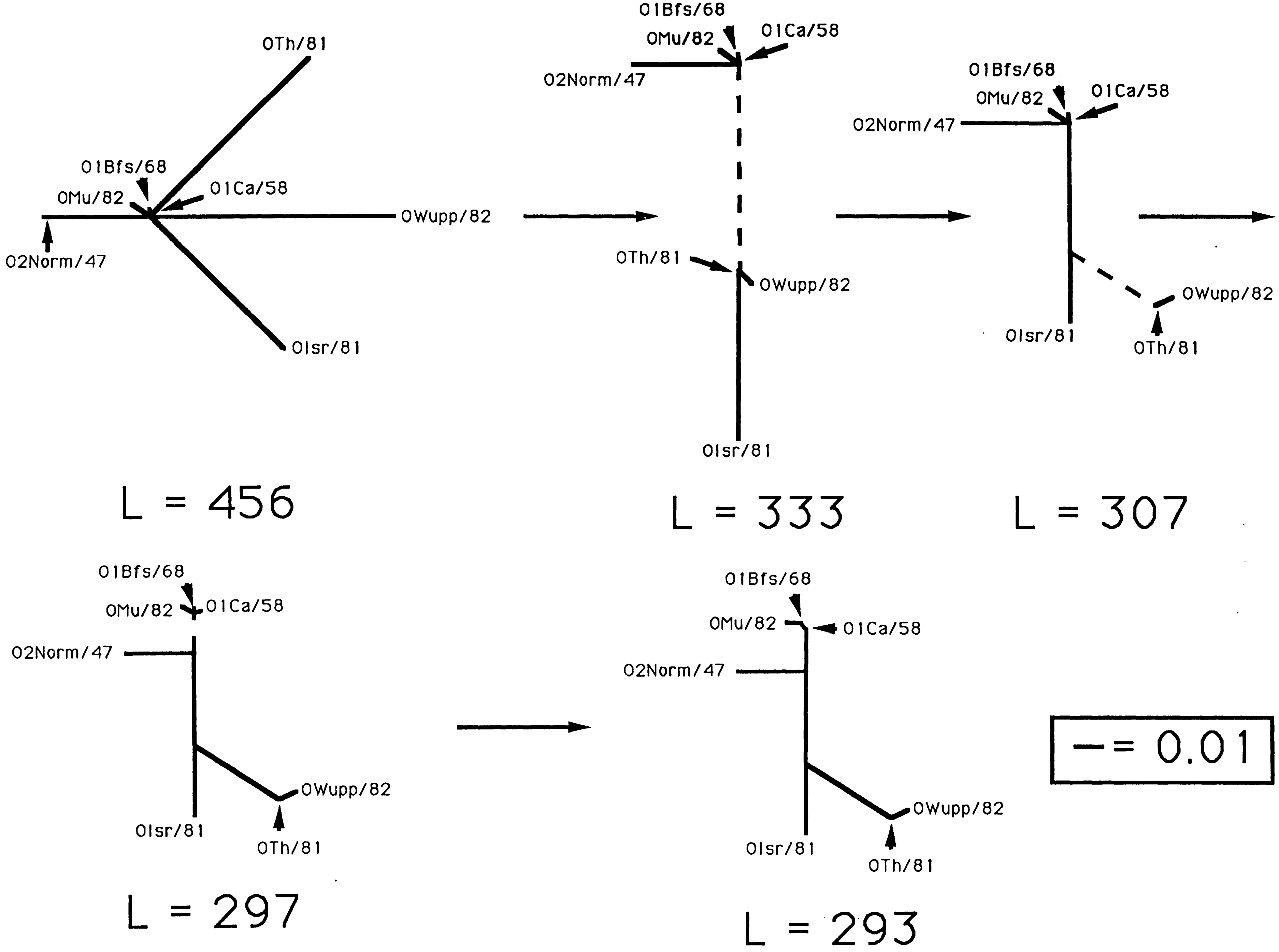
Figure 3: Evolutionary network of A–type FMDV VP1–genes. The complete model is the best fit to the data without the assumption of invariable sites. Edge lengths are proportional to estimated substitution rates. The additional estimation of invariable sites reduces the model to a simpler network (b). Splits are represented by parallel lines. The dotted lines for example represent the split A12/32, A10/61 versus the remaining sequences.

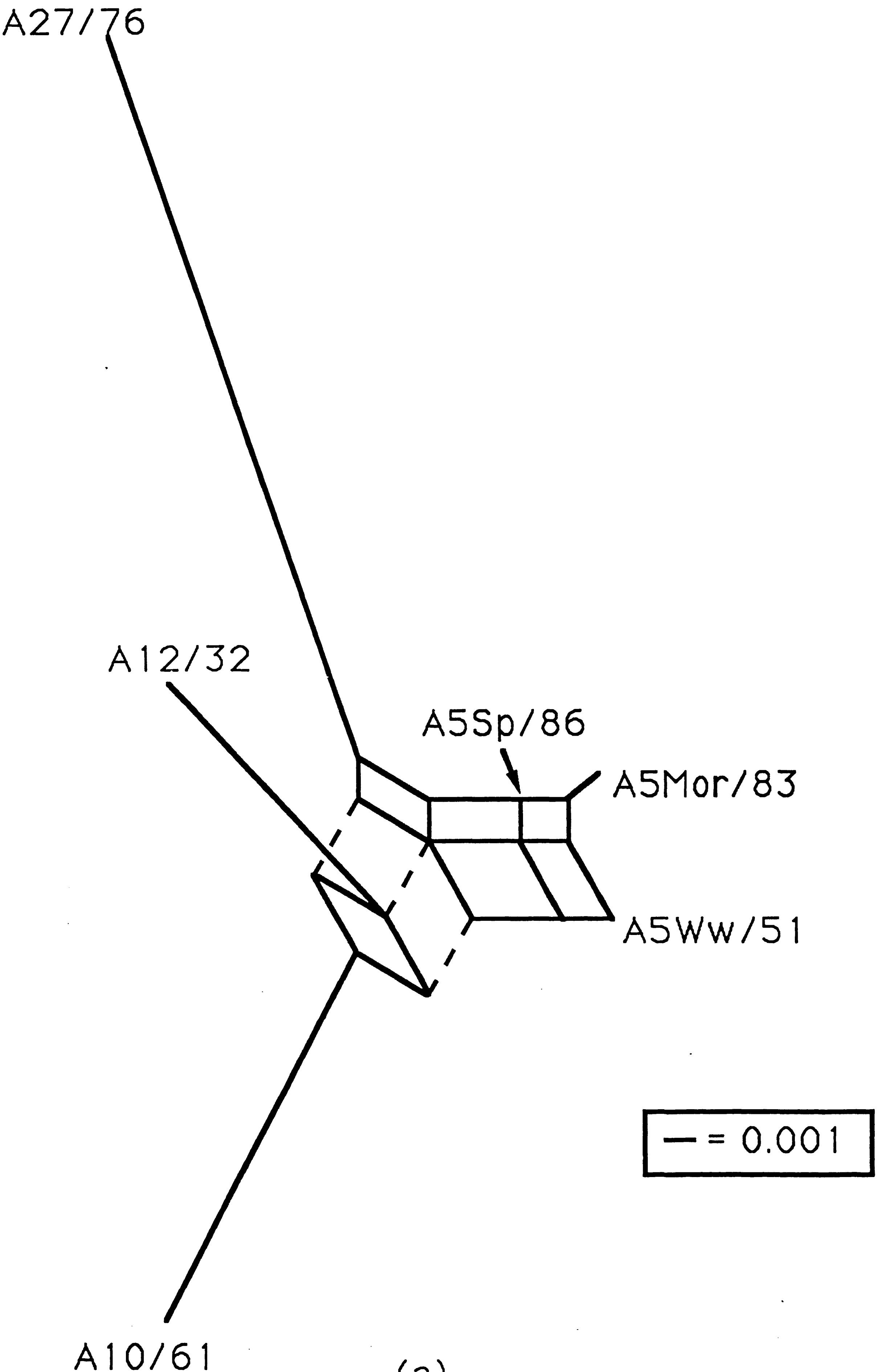
Figure 4: Evolutionary network of C–type FMDV VP1–genes (a). If invariable sites are also estimated, the relationship is tree–like (b).

Figure 5: Evolutionary network of 0–type FMDV VP1–genes. The additional estimation of invariable sites does not change the model.

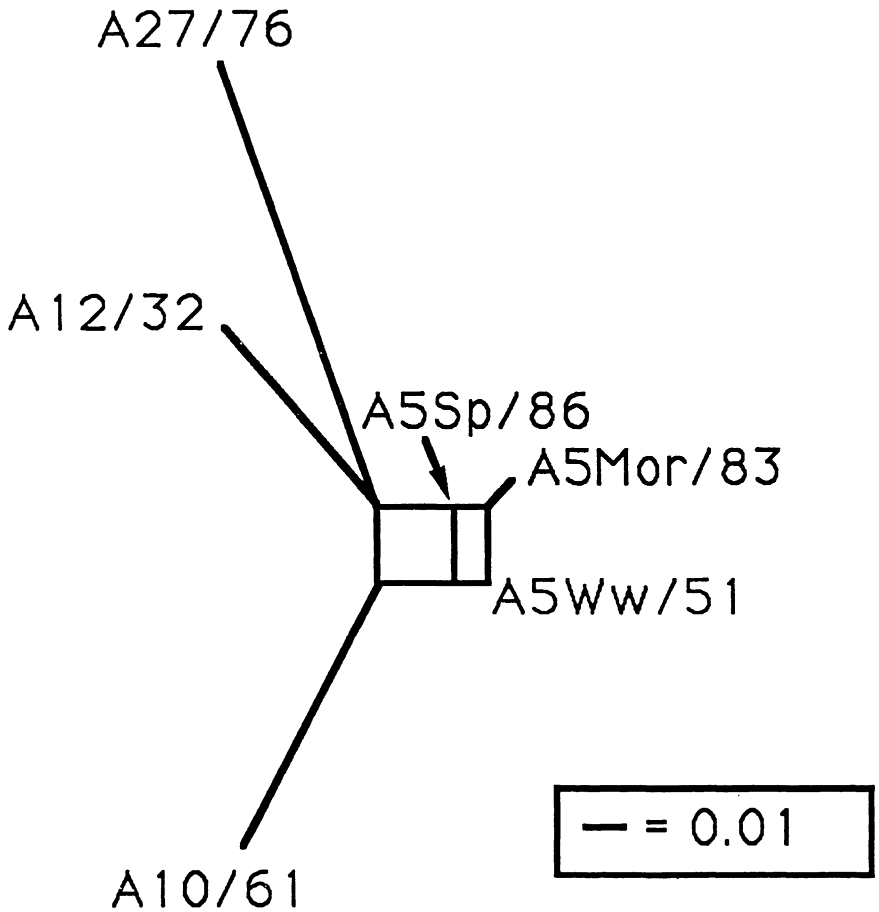
Figure 6: Evolutionary models computed with the simulation based on the simulation test. Edge lengths are proportional to the expected number of substitutions. In all instances calculations were done without assuming invariable sites in the sequences. While the A–type sequences form a network (a), the C– and O–types do not (b and c, respectively).


1
3
(a)
2
4
1
3
(b)
2
4
1
3
(c)
2
4

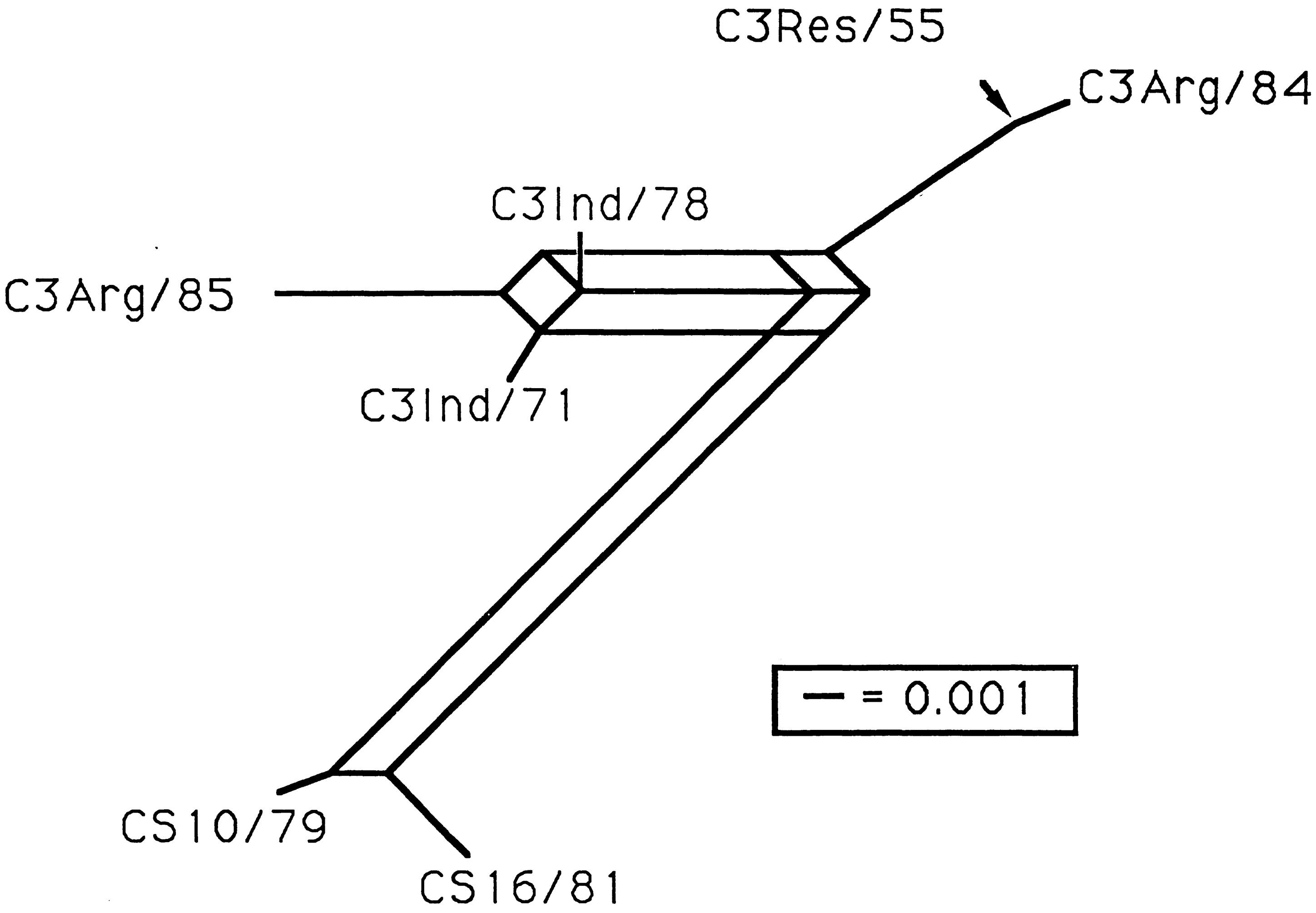

OTh/81
O1Bfs/68
OMu/82
O1Ca/58
OWupp/82
O2Norm/47
OIsr/81
L = 456
O1Bfs/68
OMu/82
O1Ca/58
O2Norm/47
OTh/81
OWupp/82
OIsr/81
L = 333
O1Bfs/68
OMu/82
O1Ca/58
O2Norm/47
OWupp/82
OIsr/81
OTh/81
L = 307
O1Bfs/68
OMu/82
O1Ca/58
O2Norm/47
OWupp/82
OIsr/81
OTh/81
L = 297
O1Bfs/68
OMu/82
O1Ca/58
O2Norm/47
OWupp/82
OIsr/81
OTh/81
L = 293
— = 0.01



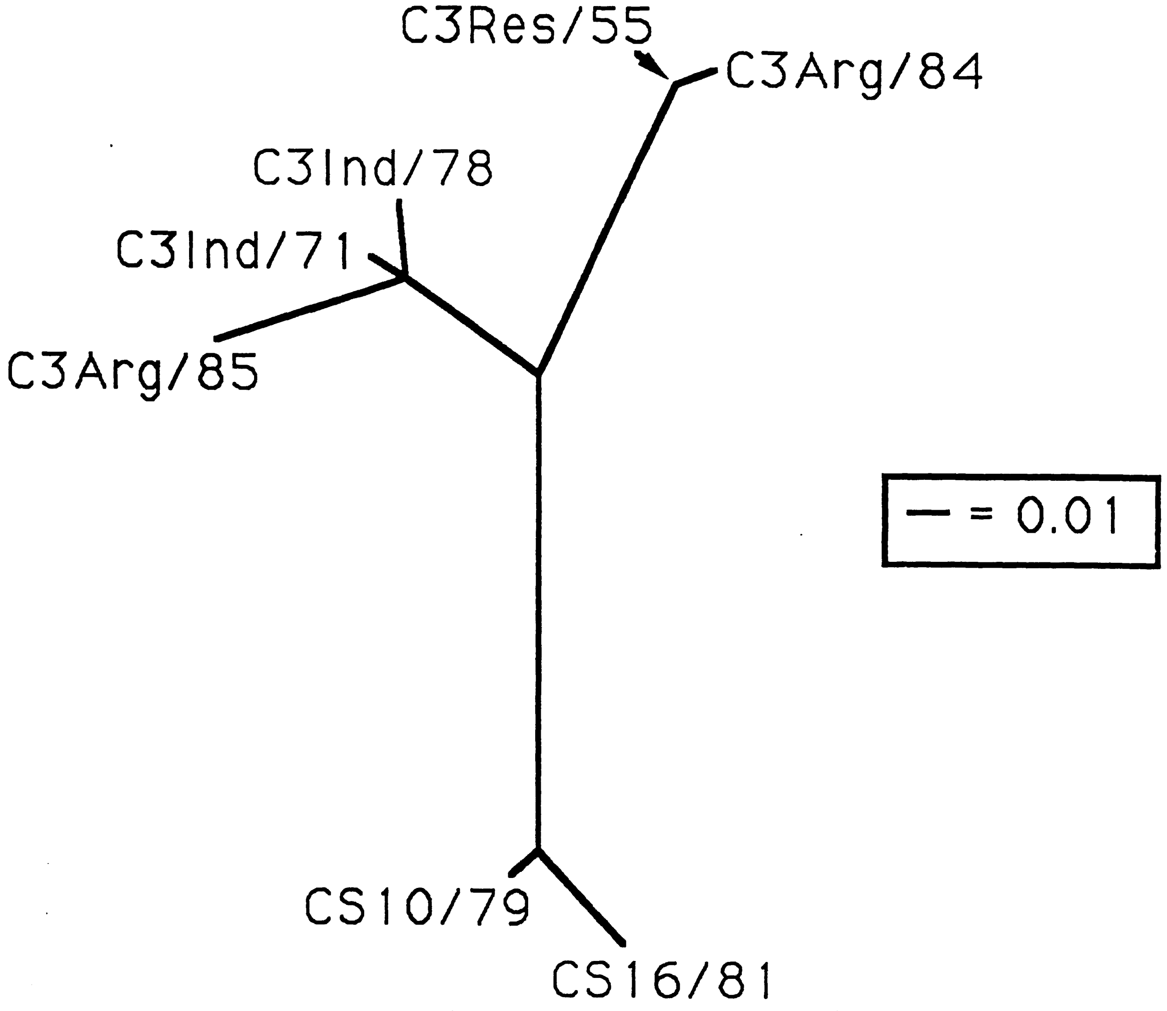
(a)



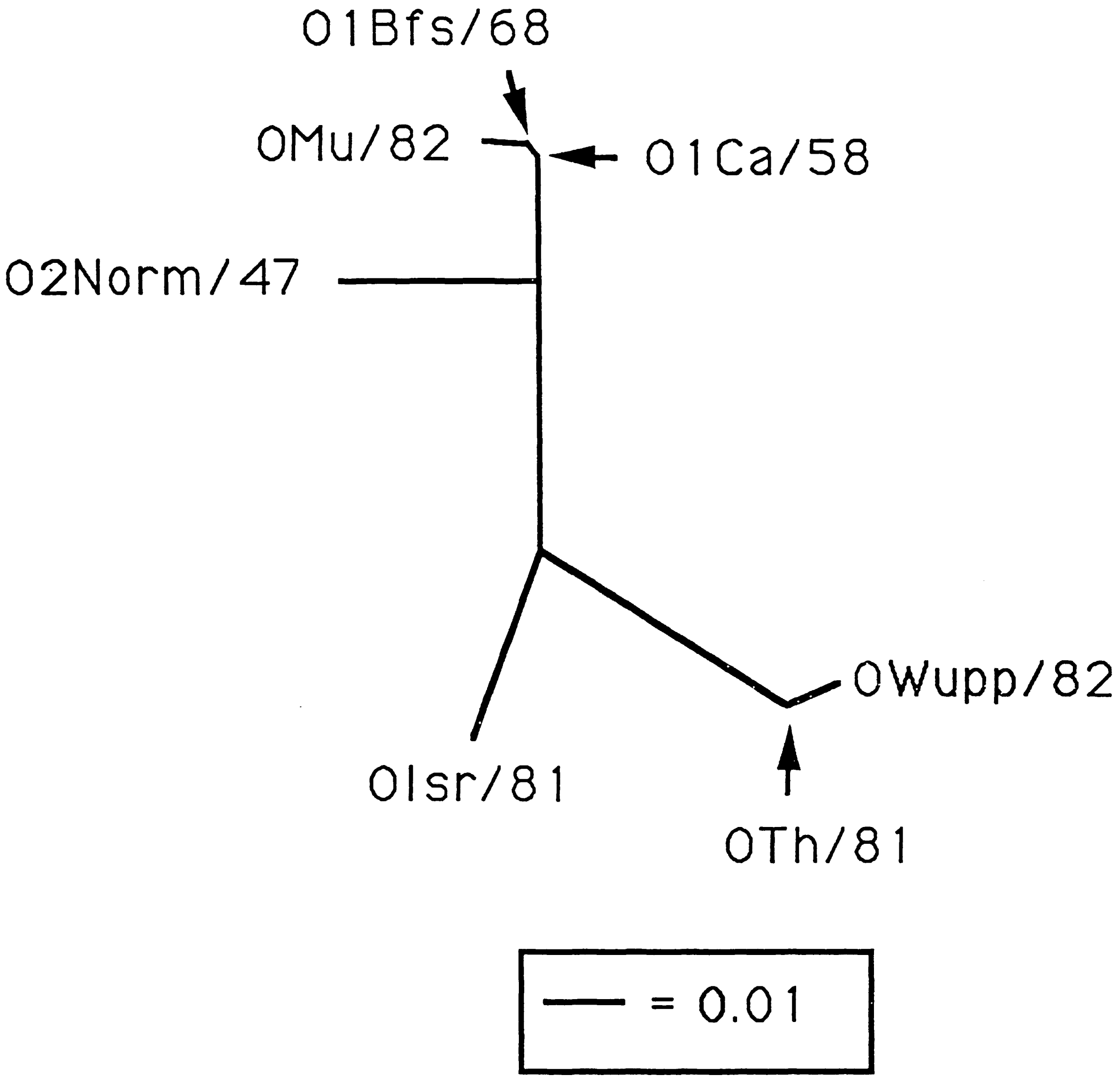
(b)

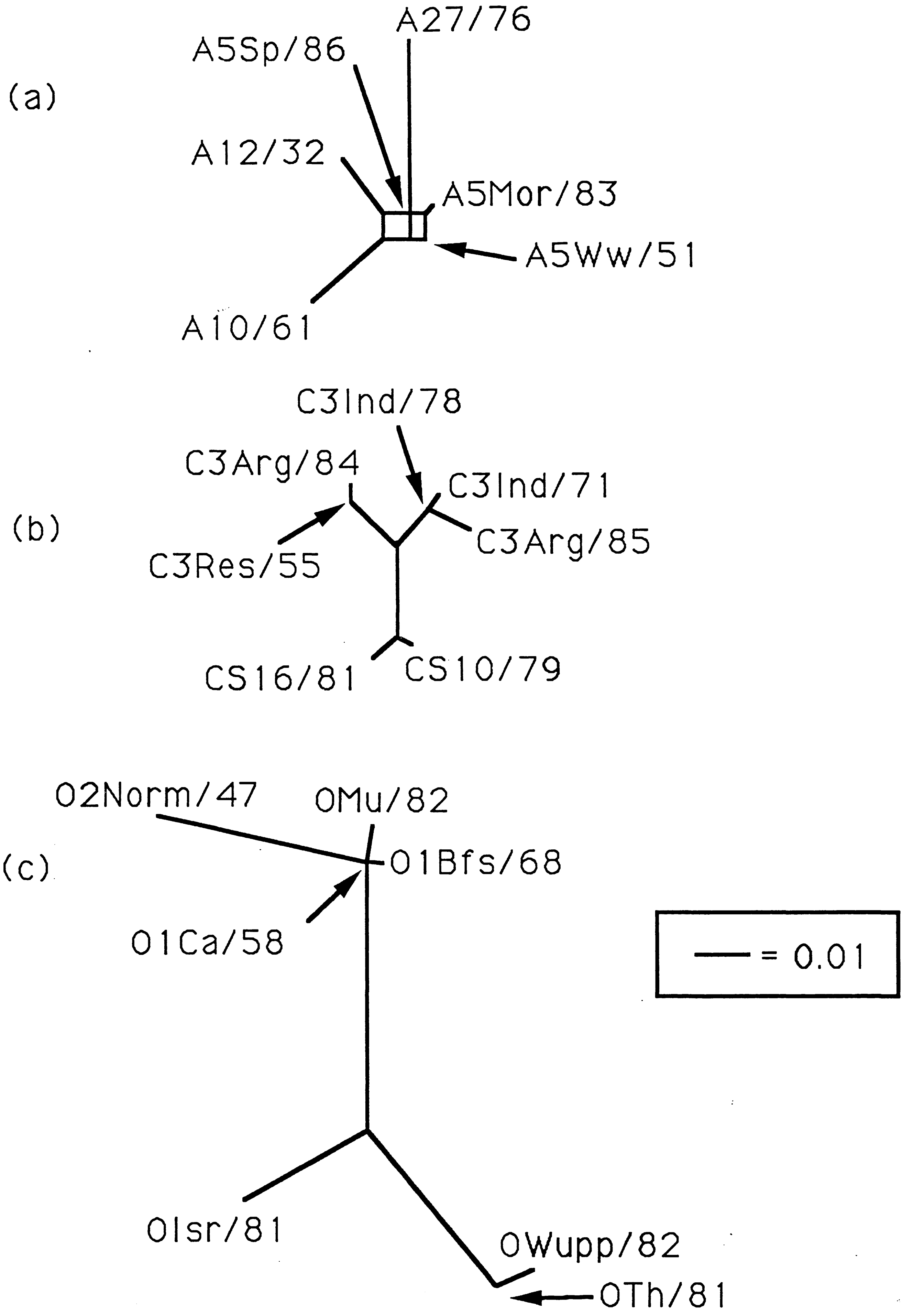

C3Res/55
C3Arg/84
C3Ind/78
C3Arg/85
C3Ind/71
— = 0.001
CS10/79
CS16/81

(a)


C3Res/55
C3Arg/84
C3Ind/78
C3Ind/71
C3Arg/85
— = 0.01
CS10/79
CS16/81

(b)


O1Bfs/68
OMu/82
O1Ca/58
O2Norm/47
OIsr/81
OWupp/82
OTh/81
= 0.01


(a)
A27/76
A5Sp/86
A12/32
A5Mor/83
A5Ww/51
A10/61
(b)
C3Ind/78
C3Arg/84
C3Ind/71
C3Arg/85
C3Res/55
CS16/81
CS10/79
(c)
O2Norm/47
OMu/82
O1Bfs/68
O1Ca/58
OIsr/81
OWupp/82
OTh/81
= 0.01

ZOOLOGISCHES INSTITUT
DER
UNIVERSITÄT MÜNCHEN

Luisenstraße 14
8000 MÜNCHEN 2, den
Telefon (089) 5 90 20 Durchwahl 5 902____
Telefax (089) 590 24 50

Zoologisches Institut Luisenstraße 14 8 München 2

Gary!

Here it is! the final version of the famous nw-paper.

I have submitted it to JME today.

As potential reviewers I suggested: Penny, Hasegawa & Felsenstein.

Now, we can lean back and wait for $ comments on the ms.

By the way thanks for the reprints

Say hello to your family

Yours

Arndt.