Molecular Phylogenetics and Evolution 48 (2008) 563–573

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Exploring tree-building methods and distinct molecular data to recover a known asymmetric phage phylogeny

Ana Sousa [a,*], Líbia Zé-Zé [b], Pedro Silva [c], Rogério Tenreiro [a]

[a] Universidade de Lisboa, Faculdade de Ciências, Centro de Genética e Biologia Molecular and Instituto de Ciência Aplicada e Tecnologia, Edificio C2, Campus da FCUL, Campo Grande, 1749 016 Lisboa, Portugal
[b] Instituto Nacional de Saúde Dr. Ricardo Jorge, Centro de Estudos de Vectores e Doenças Infecciosas, Lisboa, Portugal
[c] Universidade de Lisboa, Faculdade de Ciências, Departamento de Biologia Vegetal, Lisboa, Portugal

ARTICLE INFO

Article history:
Received 3 October 2007
Revised 15 April 2008
Accepted 22 April 2008
Available online 29 April 2008

Keywords:
Experimental phylogeny
Phylogenetic inference methods
Asymmetrical topology
Bacteriophage T7

ABSTRACT

An experimental phylogeny was constructed using bacteriophage T7 and a propagation protocol, in the presence of the mutagen *N*-methyl-*N'*-nitro-*N'*-nitrosoguanidine, based on Hillis et al. [Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., Molineux, I.J., 1992. Experimental phylogenetics, generation of a known phylogeny. Science 255, 589–592]. The topology presented in this study has a considerable variation in branch lengths and is less symmetric than the one presented by Hillis et al. [Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., Molineux, I.J., 1992. Experimental phylogenetics, generation of a known phylogeny. Science 255, 589–592]. These features are known to present additional difficulties to phylogenetic inference methods. The performance of several phylogenetic methods (conventional and less conventional) was tested using restriction site and nucleotide data. Only methods that encompassed a molecular clock or those based on sequence signatures recovered the true phylogeny. Nevertheless a likelihood ratio test rejected the hypothesis of the existence of a molecular clock when the whole sequence data set was considered. This fact or the particular substitution pattern (mainly G → A and C → T) may be related to the unexpected performance of distance methods based on sequence signatures. To test if the results could have been predicted by simulation studies we estimated the evolution parameters from the real phylogeny and used them to simulate evolution along the same tree (parametric bootstrap). We found that simulation could predict most but not all of the problems encountered by phylogenetic inference methods in the real phylogeny. Short interior branches may be more prone to error than predicted by theoretical studies.

1. Introduction

Studies that specifically address the subject of known phylogenies essentially report the cases of known transmission stories for which records have been kept (Leitner et al., 1996), well established phylogenetic relationships (by fossil records and morphological data) (Russo et al., 1996; Steinbachs et al., 2001) and experimental phylogenies (Bull et al., 1993) or pseudo-phylogenies (Sanson et al., 2002) generated for the purpose of testing phylogenetic methods. It can be argued that historical records are severely limited, that such organisms have undergone relatively little genetic divergence (as is the case for experimentally generated phylogenies where mutation rate is usually an issue) or that they cover little diversity of the phylogenies estimated from the real world. Nevertheless they do involve real, evolving biological organisms and situations for which phylogenetic methods are supposed to be applicable. In a simulation a particular mutation model is defined but then there's no way of knowing which substitution model would be tolerated by a real organism or how substitutions in different parts of a gene might interact. This is not a limitation for experimental phylogenies (Hillis et al., 1994).

The experimental model used in this study has been implemented elsewhere (Hillis et al., 1992) and its merits (Bull et al., 1993; Hillis et al., 1993) and demerits (Sober, 1993) extensively argued in the literature. We must add, however, that in the past few years several studies have been conducted that continue to argue on the advantages of this model. For example, the amazing potential of T7 to recover from the most severe conditions was demonstrated by the work of Heineman et al. (2005) that reported the re-evolution of lysis in T7 deleted for its lysin gene or by the experiences of Springman et al. (2005) that showed the regain of wild-type position of the RNApol coding gene ectopically positioned. These results illustrate the extreme plasticity of T7 genome, an essential feature for an experimental phylogeny model, since many of the interesting problems in phylogenetic reconstruction concern organisms that differ by a large percentage of their genome.

* Corresponding author. Fax: +35 121 7500172.
E-mail address: amsousa@fc.ul.pt (A. Sousa).

1055-7903/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2008.04.030

There are a number of papers exploring the applications and pitfalls of Bayesian inference (Holder and Lewis, 2003; Huelsenbeck et al., 2002). Like the maximum-likelihood method, Bayesian estimation of phylogeny is based on the likelihood function which should be an advantage since maximum-likelihood (ML) is known to outperform other methods of phylogenetic estimation under a range of conditions. Bayesian methods for phylogeny inference are now a practical alternative to more traditional methods. The primary analysis of Bayesian inference produces both a tree estimate and measures of uncertainty for the groups of the tree (faster than ML bootstrapping) and in addition allows complex models of sequence evolution to be implemented. Major disadvantages are that the prior distributions for parameters must be specified and that it can be difficult to determine whether the Markov chain Monte Carlo approximation has run long enough. For the above reasons, Bayesian analysis seems unavoidable when the goal is to compare the efficiency of tree-building methods.

Besides traditional and Bayesian methods, the emergence of new approaches to molecular phylogeny that take into account new characteristics of sequences has been rising. Among these approaches is the sequence signature method. A sequence signature is defined as the whole set of frequencies of short oligonucleotides in a sequence (Deschavanne et al., 1999). Species-specificity and conservation of signature in any part of the genome makes sequence signature a promising tool for phylogenetic analysis. It has been hypothesized that a phylogenetic analysis of signatures can reflect genomic changes that shift motif frequencies, yielding higher-order homologies available for phylogenetic analysis (Chapus et al., 2005). In T7, RNApol causes a class of transcription-induced mutations (C → T) that alter the composition of bacteriophage T7 genome and may be a significant force in genome evolution (Beletskii et al., 2000). In addition to this "natural" substitution process, T7 was propagated in the presence of NG, which enhanced several times this kind of mutation rate and also G → A mutations. The previous knowledge of this preferential mutation spectrum led us to the use of the of sequence signature, since we suspected it might be a favourable situation for this method.

Our goal was to test the performance of traditional and a few emergent phylogeny inference methods in the recovery of an experimentally generated phylogeny presenting variation in branch lengths and a less symmetric topology than previously considered (Hillis et al., 1992). The completely symmetric topology had already been tested in the paper of Hillis et al. (1994) and all methods recovered the true tree, so in order to perform a comparison an almost symmetric tree (Fig. 1B) was adapted from the more complex tree (Fig. 1A) and the performance of all the methods with these two phylogenies was compared. Despite having a symmetric branching pattern, tree B (Fig. 1) presents two particularly short interior branches, which allow the discrimination between the effects of symmetry and the effects of short branches.

2. Materials and methods

2.1. Propagation of bacteriophage T7

Escherichia coli strain W3110 was used to propagate bacteriophage T7 strain NCCB 3462 following a protocol similar to the one described by Hillis et al. (1992). The phage was grown in 1 ml cultures of *E. coli* in the presence of 20 μg/ml of the mutagen *N*-methyl-*N*′-nitro-*N*′-nitrosoguanidine (NG). After lysis proceeded to completion a 10 μl aliquot of this lysate was used to infect another culture. This procedure was repeated five times and then the phages were plated on solid medium. Next, a single plaque was randomly chosen, the clonal stock of phages present in this plaque was eluted from the agar and an aliquot was used to infect the first lysate of the next round of five.

2.2. Phylogeny construction

Bacteriophage T7 was serially propagated (as described above) according to the topology and branch lengths of the tree depicted in Fig. 1A. At each internal node, the clonal stock recovered from one plaque was used to infect two independent lineages.

To check possible contamination or swaps between different lineages, the genomes of the phages in every isolated plaque were fully mapped with HpaI and ClaI and partly mapped with Sau3AI. The restriction pattern of these enzymes evolved very quickly in this system (in particular Sau3AI) so the few cases of contamination were immediately detected and lineages were regrown from the last contamination free stock.

2.3. Restriction data (physical mapping)

Both the terminal and internal nodes were mapped for 36 enzymes (the same 34 enzymes used by Hillis et al. (1992): ApaLI, AseI, BamHI, BclI, BglII, BstBI, BstEII, BstNI, ClaI, DraI, EcoNI, EcoRI, EcoRV, HindIII, HpaI, KpnI, MboI, MluI, NcoI, NdeI, NheI, NsiI, PstI, PvuI, PvuII, SacI, SalI, ScaI, SpeI, SspI, StuI, XbaI, XhoI and XmnI, plus Eco72I and SnaBI). Viral DNA from these 26 nodes (internal and terminal) was digested with each of the enzymes, electropho-

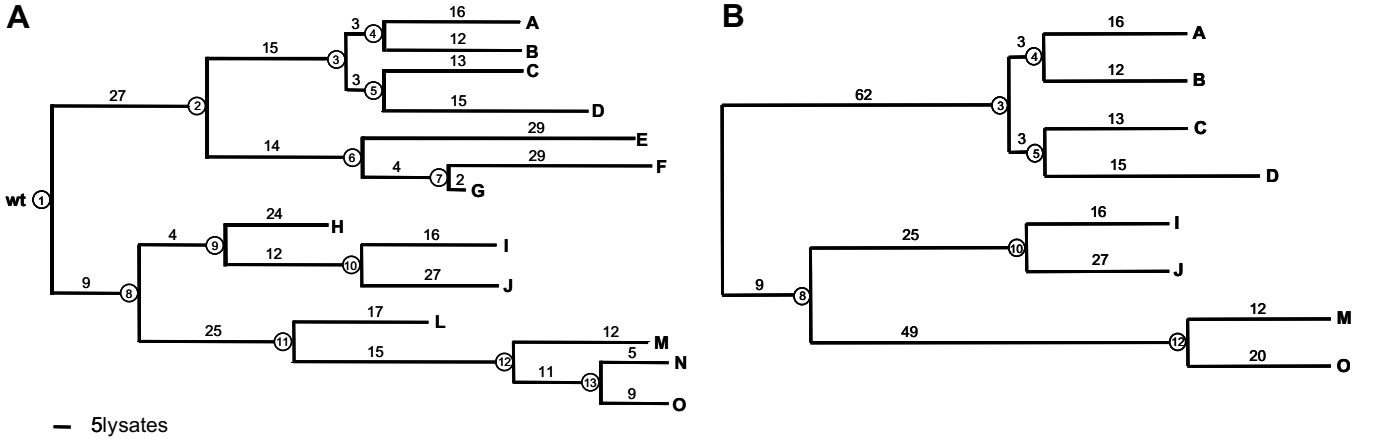

Fig. 1. (A) True asymmetric tree. (B) True symmetric tree. Circled numbers represent interior nodes. Numbers above branches indicate number of differences in restriction sites.

resed on 0.8% agarose gels and then Southern blotted. Fine mapping of the restriction site variation was accomplished by the amplification, for each of the nodes, of 22 partly overlapping fragments ranging from 615 to 4907 bp (Fig. 2) covering the whole genome (except for the first 1363 and the last 63 bp) for each of the nodes (Annex 1). These fragments were either used as probes for Southern blots hybridization or to be digested with restriction enzymes in order to infer the loss or gain of new sites. This methodology allowed a precise location of the majority of the newly created sites. The data produced from the whole set of enzymes, from all the enzymes except Sau3AI and exclusively from Sau3AI were gathered in 3 matrixes (Annex 2) and used for phylogenetic analysis.

2.4. Sequence data

A total of 4824 bp were sequenced for each of the 14 terminal nodes. The sequenced regions (Fig. 2) consisted of 9 fragments scattered through the genome and covering both essential and non essential functions of the phage. This strategy was chosen in light of the finding by Cummings et al. (1995) that "blocks of contigous sites are less likely to lead to the whole-genome tree than samples composed of sites drawn individually from throughout the genome". These fragments were sequenced at least once on both strands using the CEQ Dye Terminator Cycle Sequencing (DTCS) Quick Start Kit (Beckman Coulter). Nucleotide sequences were determined with a CEQ 2000 XL Sequencer (Beckman Coulter) and are available at GenBank under the Accession Nos.: EF516992–EF517117.

All the sequence alignments were performed with CLUSTAL X (version 1.83) (Thompson et al., 1997) with the default options.

2.5. Phylogenetic analysis

2.5.1. Congruence analysis

Measuring and testing the significance of phylogenetic incongruence is necessary when considering genome-scale datasets composed of multiple genes (Planet, 2006). Non significant incongruence can be due to inadequate sample sizes but significant incongruence can arise from different rates of evolution between partitions (Kolaczkowski and Thornton, 2004; Mossel and Vigoda, 2005) (codon position, functional constraints) or from partitions that have had different histories.

Four character incongruence tests were performed to decide whether or not to combine nucleotide and restriction site data. In addition incongruence between each pair of genes and between the fast evolving Sau3AI restriction sites and all the other enzymes was also assessed.

The parsimony-based tests were the incongruence length difference test (ILD) (Farris et al., 1994), Kishino–Hasegawa test (KH) (Kishino and Hasegawa, 1989), winning-sites test (Prager and Wilson, 1988) and Templeton test (Templeton, 1983) as implemented in PAUP* (version 4.0b10) (Swofford, 2003).

2.5.2. Phylogenetic inference

2.5.2.1. Traditional approach. The seven methods of phylogenetic inference evaluated were: unweighted pair-group method of arithmetic averages (UPGMA) (Sokal and Michener, 1958), neighbour joining (NJ) (Saitou and Nei, 1987), minimum-evolution (ME) (Rzhetsky and Nei, 1992), Cavalli-Sforza method (uLS) (Cavalli-Sforza and Edwards, 1967), Fitch–Margoliash method (wLS) (Fitch and Margoliash, 1967), maximum parsimony (MP) (Eck and Dayhoff, 1966) and maximum-likelihood (ML) (Felsenstein, 1981).

For the restriction site data the distance methods UPGMA, NJ, ME, uLS and wLS the mean character difference, total character difference, Upholt distance and Nei–Li distance were used. All the analyses were performed in PAUP* except for the maximum-likelihood analysis that was done with RestML program of the PHYLIP (Felsenstein, 2006) (version 3.66) package.

The same seven phylogenetic inference methods were used for the analysis of sequence data and the distance measures were: *p*, Jukes-Cantor (JC) and Kimura-2-parameter (K2 P) distances. We used these methods as implemented in PAUP*, with heuristic searches for ME, uLS and wLS and the search algorithm branch and bound for MP (weighted with the rescaled consistency index and unweighted). ML analysis was also performed in PAUP* with a full heuristic search with 100 random additions of sequences and the evolutionary model (that best fits the data based on the Akaike information criterion, AIC (Akaike, 1974)) selected by Modeltest (version 3.7) (Posada and Crandall, 1998). For the symmetric tree exhaustive searches were performed for all the analyses.

The robustness of the methods of analysis was evaluated by bootstrapping. The bootstrap procedure was replicated 1000 times for UPGMA, NJ and MP and 100 times for ML.

A combined analysis of restriction site and nucleotide data was also done.

2.5.2.2. Bayesian methods. MrBayes (version 3.1) (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) allows the specification of a partitioned model, making it possible to assign different evolutionary models for each gene partition in a single analysis. It also permits the combined analysis of different data set (e.g. restriction site and nucleotide), so besides the separate analysis of sequence and restriction data we also joined these data in a single analysis. Analysis of individual partitions by MrModeltest (Nylander, 2004) indicated the best fit model for each partition according to the AIC.

For the Bayesian analysis, besides the partitions considered above, we also considered the following: all sequences and a single evolutionary model, all sequences and one evolutionary model per

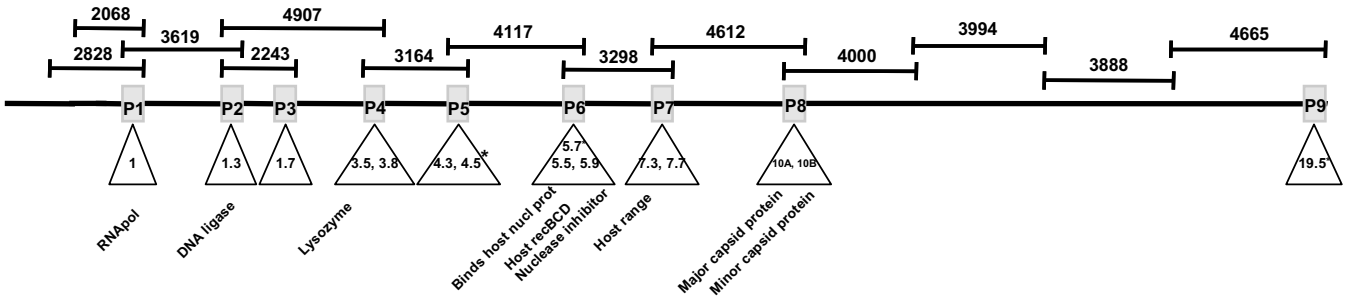

Fig. 2. T7 bacteriophage genome. Gray boxes stand for the PCR amplified regions used for sequencing (P1–P9). Solid lines, above the genome, represent amplified sequences used for fine restriction site mapping (dimension of these sequences, in bp, are showed above the lines). Inside the triangles are the genes present in sequenced fragments. '*' Means fully sequenced gene. The known functions for some of these genes are also shown. All boxes and lines represented are drawn to scale.

sequenced region, all sequences and one evolutionary model per gene, all sequences plus one evolutionary model for first and second codon positions and one evolutionary model for the third position and finally all sequences and one evolutionary model for the first and second codon positions of each gene and one evolutionary model for the third position.

For the combined data set analysis we considered all sequences and a single evolutionary model plus the restriction sites of all enzymes or all enzymes except Sau3AI or only Sau3AI.

For all analyses we treated each partition as "unlinked", so that separate parameter estimates were obtained for all runs. Two independent runs of one million generations were performed, each with 4 chains and trees were sampled every 100 generations. We checked the stationarity of the sampled trees with the Tracer software (version 1.3) (Rambaut and Drummond, 2005) and summarized the posterior distribution of trees by a majority-rule consensus tree.

2.5.2.3. Sequence signatures. Sequence signatures were computed with the algorithm "Chaos game representation" implemented in NASC software (version 4.21) (Vinga and Almeida, 2003). NASC performs different types of sequence comparison based in some distance definitions between frequencies of L-words or L-tuples. For each sequence and each chosen sub-sequence of length L, a vector of counts was extracted. These vectors were used to calculate the Euclidean distance[1] between the sequences and the sample probabilities distributions of the 6-words present. This metric distance and word length have been reported to give the best results for phylogenetic analysis (Chapus et al., 2005). The distance matrices were then used to infer trees with UPGMA, NJ, ME, uLS and wLS algorithms implemented in PAUP*.

For all of the phylogenetic methods whenever a method produced more than one but less than 5 trees the strict consensus tree was obtained; above 5 trees the majority-rule consensus tree was constructed.

2.5.2.4. Accuracy of topology. The accuracy of topology of the trees obtained from each gene was evaluated from the topological distance (d_T) (Penny and Hendy, 1985; Robinson and Foulds, 1981) of each inferred tree to the true tree. This distance reflects the number of internal branches present in one tree but not in the other. In our data set (tree A) d_T ranges from 0 (same topology as the true tree) to 22, as the true tree has 11 internal branches.

3. Results and discussion

As expected, since the true phylogeny is known and all the partitions have had the same history, no major incongruence cases were detected. ILD tests detected no cases of significant ($p < 0.05$) incongruence for all the partitions considered (Annex 3.1 and 3.2) and the KH and Templeton tests considered the nucleotide and restriction site data incongruent with respect to the first partition but not significantly incongruent when the second partition was considered (Annex 3.1). For the non-ILD tests of incongruence (Annex 3.3) most of the 66 pairwise comparisons of genes revealed a non-significant level of incongruence ($p > 0.05$) the only exceptions being the genes *1.3, 1.7*, *7.7*, and *10B* (without a Bonferroni correction[2]). For genes *1.3*, *1.7* and *10B* this might be the result of these genes having twice the length of the remaining but this was not the case for gene *7.7*. In fact, although the probability of rejecting the null hypothesis of congruence for all the pairwise gene comparisons involving this gene was never significant with the ILD test (Annex 3.2), these had the lowest p values. These results compelled us to test the accuracy of the phylogeny inferred when this gene was omitted. Five equally parsimonious trees were produced (in opposition to the single MP tree inferred by the whole data set) when gene *7.7* was discarded from the global analysis, but the strict consensus tree had half the d_T to the true tree. The topological accuracy did not improve for the other inference methods when gene *7.7* was deleted.

The two partitions in the restriction site data were considered significantly incongruent ($p < 0.05$) by the KH and Templeton test with respect to the first partition (Annex 3.1). By contrast to the gene data, the accuracy of tree reconstruction suffered from combining both partitions since when the restriction site data (omitting Sau3AI sites) were considered alone the true tree was recovered (UPGMA, NJ, ME and wLS) or at least the d_T was smaller than for the combined analysis. In comparison, the Sau3AI partition performed much worse when considered isolated (Table 1).

3.1. Inferred and actual phylogeny comparison

3.1.1. Accuracy of topology

As expected from the incongruence analysis results, combining all the restriction site data did not improve the accuracy of topology. In fact the d_T values were between 2 and 6 for the complete data set but dropped to a minimum of 0 and a maximum of 3 when the Sau3AI sites were omitted from the analysis (Table 1). When these sites were considered alone the trees inferred departed from the true tree by a d_T of 4 to 10. As Bull et al. (1993) pointed out in a similar study, Sau3AI recognition sites have unique features in T7 genome that distinguishes them from the other enzymes. Based on the wild-type genome and on the knowledge of the mutation spectrum they identified the sites that are a single mutation G → A or C → T from becoming a recognition sequence (1-off sites) for all the enzymes. These sites are much more abundant for Sau3AI than they are for the other enzymes, since the recognition sequence of this enzyme is counter selected in the wild-type (Bull et al., 1993). The recognition sequence of Sau3AI, GATC, is statistically expected 156 times in a 40 kb genome but occurs only 6 times in T7 (Granoff and Webster, 1999). As in the work by Hillis et al. (1992), the gains of new restriction sites for this enzyme were almost half of the total gains (131 new sites only for Sau3AI against 144 for all the other enzymes). In our work, the mapping strategy of Sau3AI was improved but a high level of convergent gains is still expected (that is two different gains being scored as only one due to their proximity or to the same mutation arising in two individuals by chance alone). This may disturb phylogenetic inference, particularly for methods based on the Nei–Li distance (Nei and Li, 1979). This model makes the assumption that all sites that are shared between two species were present at a common ancestor halfway between them.

Sau3AI sites did not affect phylogeny inference in Hillis et al. (1992) probably due to the simple chosen topology. Bull et al. (1993), in their study of the molecular evolution of T7, estimated probabilities of gains and losses for Sau3AI sites and found that these were significantly different from the rest of the enzymes. The use of these parameters in complex phylogenies, such as the one presented in this paper, might be a real necessity.

Ranking the methods by their overall performance with restriction site data we get almost the opposite order than by doing the same procedure with nucleotide data. With restriction site data, UPGMA and ME had the best performance, Bayesian and likelihood criteria produced the worst topologies.

If instead we consider the data set composed of all the restriction sites except those of Sau3AI than all methods performed almost equally well but only the distance methods retrieved the exact true topology ($d_T = 0$). MP, ML, and Bayesian methods consis-

[1] The square root of the sum of the square of the differences in frequency of strings between species.

[2] If this correction was applied only genes *1.3*, *1.7* and *10B* showed a few cases of significant incongruence.

Table 1
Topological distances (d_T) of inferred trees from the true tree for restriction site data and both restriction site data and sequence data combined in a single data set

Data sets	All enz[a]	All enz b S[b]	Sau[c]	Sum	All enz + all seqs	All enz b S + all seqs	Sau + all seqs	Sum
UPGMA								
Upholt[d]	2	0	6	8				
Nei–Li[e]	2	0	6	8				
Tot. diff.[f]	2	0	4	6	0	0	0	0
Mean diff.[g]	2	0	4	6	0	0	0	0
NJ								
Upholt	4	0	10	14				
Nei–Li	4	0	10	14				
Tot. diff.	4	0	6	10	2	2	2	6
Mean diff.	4	0	6	10	2	2	2	6
ME								
Upholt	4	0	6	10				
Nei–Li	4	0	6	10				
Tot. diff.	2	0	5	7	2	2	2	6
Mean diff.	2	0	6	8	2	2	2	6
uLS								
Upholt	4	2	8	14				
Nei–Li	4	2	8	14				
Tot. diff.	4	2	5	11	2	2	4	8
Mean diff.	4	2	6	12	3	3	5	8
wLS								
Upholt	2	0	10	12				
Nei–Li	2	0	10	12				
Tot. diff.	2	2	4	8	2	2	2	6
Mean diff.	2	2	4	8	2	2	2	6
MP								
Unweighted	4	3	6	13	2	2	2	6
Weighted	4	2	5	11	2	2	2	6
ML								
	6	2	10	18				
Bayesian								
	6	2	6	14	2	3	4	9
Sum	80	21	157		23	21	29	

[a] Restriction sites from all the enzymes.
[b] Restriction sites from all the enzymes except Sau3AI.
[c] Restriction sites from Sau3AI.
[d] Upholt distance.
[e] Nei–Li distance.
[f] Total difference distance.
[g] Mean difference distance.

tently failed to correctly place lineage H and MP (unweighted), uLS and wLS (with total and mean differences distance) could not resolve or correctly infer branch 3–5.

The distance matrices produced by the sequence signature method led to the inference of trees that differed from the true tree by very high d_{Ts} (data not shown).

Gene length is usually considered an important factor in recovering the true topology. In our study, omitting gene *1* from the regression analysis, 80% of the variance of the performance of the genes could be explained by gene length (results based on regression analysis of the sum of d_{Ts} of the trees produced by each gene versus its length). Therefore a dependence of the accuracy of the inferred topology on sequence length is apparent. Gene *1* will be discussed later.

Table 2 summarises some statistical properties of the genes included in this analysis. The g_1 statistics (Huelsenbeck, 1991) is related to phylogenetic signal. A more negative g_1 value (left-skewed distribution) indicates a stronger phylogenetic signal. Consistency of information among individual parsimony informative sites in the true tree is apparent from average consistency indices (CI), average retention indices (RI) and rescaled consistency indices (RC). The range of these indices is 0–1, a higher value indicates a higher agreement between the characters in the data set.

The values of mean pairwise proportion of differences in percent (Table 2) are relatively small so that synonymous sites are most certainly not saturated. Despite this, we examined the efficiency in obtaining the correct tree by using all three codon

Table 2
Statistical properties of genes included in this study

Genes	n_{tot}[a]	n_{print}[b]	n_{var}[c]	p (%)[d]	g1[e]	CI[f]	RI[g]	RC[h]
1	532	16	33	1.77 (0–3)	−0.58	0.94	0.97	0.91
1.3	553	27	49	2.63 (0–4.3)	−0.65	0.85	0.90	0.76
1.7	564	27	54	2.5 (0.5–3.9)	−0.47	0.81	0.83	0.67
3.5	292	9	28	2.2 (0–4.5)	−0.60	0.93	0.93	0.87
3.8	186	8	21	2.68 (0–6.5)	−1.15	0.88	0.85	0.74
4.5	270	6	18	1.66 (0–3.7)	−0.29	0.90	0.92	0.83
5.5	239	12	25	2.6 (0–5.5)	−0.90	0.86	0.88	0.76
5.7	210	9	19	2.5 (0–3.79)	−0.68	0.86	0.91	0.79
7.3	238	13	22	2.99 (0–6.8)	−0.45	0.85	0.92	0.78
7.7	275	15	33	3.12 (0–5.81)	−0.63	0.83	0.84	0.69
10B	565	22	48	2.3 (0.2–3.9)	−0.57	0.87	0.90	0.79
19.5	150	5	15	2.2 (0–4.8)	−0.55	0.76	0.70	0.54
All	4824	200	434	2.4 (0.5–3.7)	−0.59	0.85	0.88	0.75

[a] Total number of nucleotides of each gene.
[b] Number of parsimony informative sites.
[c] Number of variable positions.
[d] Mean pair wise p distance in percent and range.
[e] g1 statistic.
[f] Average consistency index.
[g] Average retention index.
[h] Rescaled consistency index.

positions versus only first and second codon positions. As predicted, trees inferred with the first approach showed generally lower d_T values (data not shown). Interestingly the use of all codon positions but with one evolutionary model for the first and second positions and a different model for the third positions (Bayesian inference with all sequence data) reduced d_T from 4 to 2. This may indicate that although there's no saturation in synonymous positions their evolution under different constraints is worth taking into account.

In Table 3, n_c stands for the number of genes that produced a tree with $d_T \leqslant 4$. For the distance methods this was accomplished 38 times, 12 of them when the Euclidean distance was used (between six base sequence signatures).

If we now consider the global analysis, all methods recovered the true topology, at least once, except for uLS, MP and Bayesian criterion. UPGMA is the only method that assumes a molecular clock, and was also the distance method that recovered more often the true tree. The best-fit model (TIM + G) selected by AIC in Modeltest accounts for base frequency differences, substitution rate variation among sites and bias in substitution types. These features are neglected by simpler models such as JC or K2P, yet the accuracy of the topologies inferred did not benefit from this more sophisticated model ($d_T = 4$, for both models of nucleotide substitution). Excluding the distance methods, the true topology was only recovered when a molecular clock was enforced. This seems in agreement with UPGMA results. Nevertheless a likelihood ratio test (LRT) rejected the null hypotheses of the existence of a molecular clock ($p < 0.01$). This test was also made for each gene separately and none rejected the hypothesis of a molecular clock, probably because the trees inferred from each of the genes alone were based on few informative positions which led to frequent politomies. So there is a greater probability that there are several trees that explain equally well the data, among those being the tree that assumes a molecular clock. The same argument justifies the fact that the LRT, based on all sequences concatenated, rejected the existence of a molecular clock, since when enough informative positions are considered the inferred tree is better supported making the difference between this and the tree assuming a molecular clock statistically significant.

Even when we consider the best estimates of the known phylogeny, every method, except those assuming a molecular clock (UPGMA, ME and ML with a molecular clock) seem to have difficulty inferring the small internal branch (slow evolving) that leads to H, I and J cluster (branch 8–9) and branch 9–10.

It has been verified, by simulation studies (Nei et al., 1998), that these small interior branches tend to be frequently misinferred. On top of that, if we look at the number of differences in restriction sites in branch 8–9, we have reasons to suspect that, in reality, this branch is even smaller than planned, i.e. fewer differences than expected arise. To test if this fact was related to the consistently failure of most of the methods we conducted a simulation study (parametric bootstrapping) generating 100 data sets using the true tree plus the evolution parameters estimated from the real data (model, general time reversible; frequency of bases, A = 0.2931, C = 0.2162, G = 0.2339, T = 0.2568; rate matrix, 1.0000, 233.4292, 4.2426, 4.2426, 303.5897, 1.0000; gamma rate heterogeneity with shape = 0.2172). All the branch lengths were the same as in the real phylogeny, except for the branch 8–9, which was shrunken to fit

Table 3
Topological distances (d_T) of inferred trees from the true tree for nucleotide sequence data

Genes	*1*	*1.3*	*1.7*	3.5	*3.8*	*4.5*	5.5	5.7	*7.3*	7.7	*10B*	*19.5*	All	Sum	n_c[a]
UPGMA															
P	12	4	10	12	16	16	14	14	16	12	*0*	18	0	130	2
JC	12	4	10	12	16	14	14	14	16	8	*0*	18	0	124	2
K2P	12	4	10	12	16	14	14	14	16	8	*0*	18	0	124	2
L6	12	4	10	12	16	16	14	14	16	10	*0*	20	4	148	2
NJ															
P	12	8	4	8	14	14	8	12	12	8	4	18	4	106	2
JC	14	8	4	6	14	12	8	12	14	8	4	18	4	106	2
K2P	14	8	4	6	14	12	8	12	14	8	4	18	4	106	2
L6	10	4	4	12	14	14	10	12	10	8	0	18	0	116	3
ME															
P	8	5	4	9	10	11	7	9	11	9	4	16	4	80	2
JC	10	5	4	6	11	13	7	11	12	8	4	16	4	88	2
K2P	10	7	5	8	11	13	7	11	12	8	4	16	4	92	1
L6	9	2	4	8	13	14	10	12	8	8	0	14	0	102	3
uLS															
P	11	6	4	8	10	10	8	9	11	9	4	13	3	89	2
JC	11	6	4	8	10	11	9	9	11	9	3	13	3	94	2
K2P	11	6	4	8	10	10	8	9	11	8	3	13	3	91	2
L6	10	4	6	7	11	12	12	12	10	10	1	14	4	113	2
wLS															
P	11	7	6	7	10	13	7	11	11	9	4	15	4	97	1
JC	11	7	6	6	12	13	7	10	11	8	4	15	4	98	1
K2P	11	7	6	6	12	13	7	10	11	8	4	15	4	98	1
L6	9	4	8	10	13	14	10	14	8	10	0	14	0	114	2
MP															
Unweighted	7	5	3	5	8	8	7	7	9	8	4	12	4	66	2
Weighted	7	5	3	5	8	8	7	7	9	8	4	12	4	79	2
ML															
	8	7	5	5	8	8	7	7	9	8	4	13	4	59	1
Clock	nd[b]	1	5	6	10	9	7	9	12	9	4	13	0	85	2
Bayesian	9	7	4	5	7	9	7	7	7	8	4	12	4	90	2

[a] Number of genes that produced a tree with a $d_T < 4$ from the true tree.
[b] Non determined because >1000 optimal trees were found.

the relative size observed from the number of differences of restriction sites. One hundred data sets were produced and used to infer the UPGMA (JC distance), NJ (JC distance), MP and ML (JC distance) trees. UPGMA produced the true tree 4% of the time, NJ 43%, MP 30% and ML 54%.

Fig. 3 shows a comparison between the results of the simulation with those from real data. UPGMA results were reasonably in agreement with those of the real data. As stated before, UPGMA produced the true topology for all the distances except for sequence signatures. In the later case one or both of the two branches misinferred were also wrong in 65% ($d_T = 2$ and $d_T = 4$) of the trees estimated from simulated data (including Bayesian analysis that misinferred these two branches with a posterior probability of 0.5 and 0.64, respectively). The rest of the methods were all wrong about branches 9–10 and 8–9 in the real phylogeny. Branch 8–9 was also misinferred in a high percentage of trees by all the methods in the simulation study, but this was not the case for branch 9–10, which was never incorrectly predicted. Branches 3–5 and 3–4 were also a problem for the simulation study probably because these branches were larger in the true phylogeny (relying on restriction site data). Branch 6–7 was also a small problem for MP and ML, probably because of a phenomenon known as "long branch attraction". Parametric bootstrapping failed to infer the misplacing of branch 9–10, yet in the real phylogeny this branch was consistently misinferred by all methods but UPGMA, with moderate to high support (bootstrap indices from 52% to 86%). This fact may indicate that theoretical studies (even with empirical parameters) cannot predict all the details of "real life" and some other explanation must be found to account for this phenomenon.

Branches 3–4 and 3–5 were also predicted to be misinferred a significant number of times, yet they were both correctly inferred

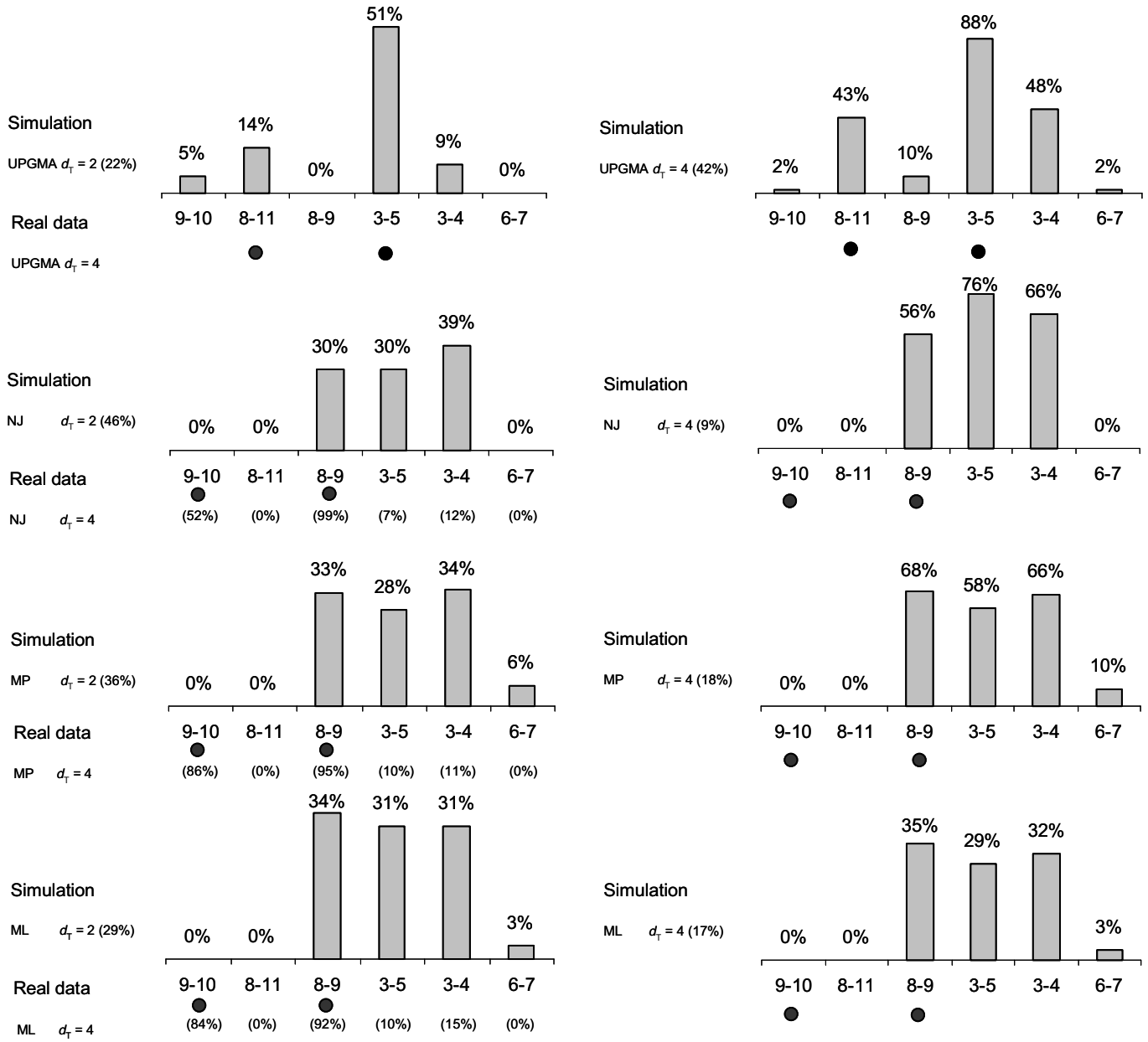

Fig. 3. Number of times (in percent) on top of columns, that the branches indicated below the line were misinferred by the simulation study. '●' Indicate branches that were also misinferred from real data. Percentages below the line are bootstrap values.

by NJ, MP and ML in the real phylogeny. This was probably due to a phenomenon inverse to what led to the misinference of 8–9 in the real phylogeny, i.e. a higher number of substitutions accumulate in these branches than were expected by the known length (number of lysates) of these branches (number of differences in branches 3–4 and 3–5 in Fig. 1). It is known that variable rates among lineages are also a source of error (Lyons-Weiler and Takahashi, 1999).

Taken together the overall performance of the methods, ML and MP produced the smaller sum of d_{Ts}. By this criterion NJ and UPGMA performed the worst and the other methods were fairly similar. The use of different evolutionary models for each gene or each sequenced region, in the global analysis, by the Bayesian criterion did not improve the accuracy of the inferred phylogeny (data not shown).

Since heuristic searches were done for all the methods except parsimony, the ME and ML scores were calculated for both the true tree and the tree inferred for each method. The true tree had always a higher ME score and a lower likelihood (so even if an exhaustive search was conducted it would never converge to the true topology).

Ranking the genes by their performance in recovering the most accurate trees and relating it to their known features, such

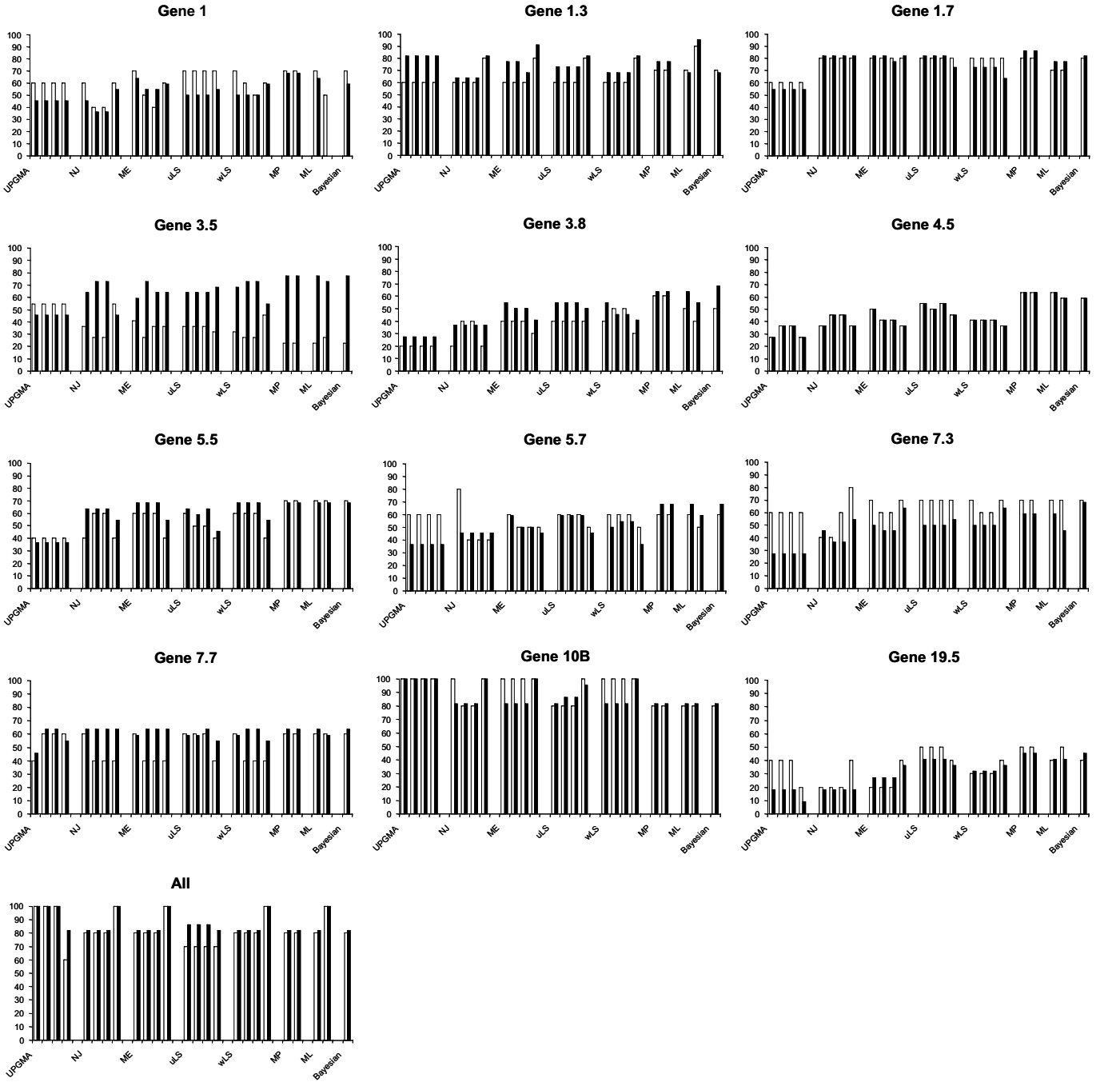

Fig. 4. Percentage of branches correctly inferred for the symmetric tree (white columns) and asymmetric tree (black columns). Each column set of distance methods represent a different distance measure (from left to right: *p*, JC, K2P and sequence signature based distance). MP is depicted as unweighted and weighted and ML is considered without and with a molecular clock enforced.

as length, proportion of differences (*p* distance), phylogenetic signal or biological function is not straightforward matter. Lengthier genes usually produced more accurate trees, but this was not true for gene *1*, which performed worst than expected, or for gene *3.8* which was the second smaller but the fifth in performance.

Approximately half of the 59 genes in T7 are considered nonessential, or at least conditionally essential. These genes are designated with fractional numbers (Dunn and Studier, 1983). Not surprisingly gene *7.7*, *5.5*, and *1.3* presented 3, 1 and 1 nonsense mutations, respectively, and gene *1.7* presented one dinucleotide (AT) insertion.

Gene *1* and *10* are the only essential genes for viability in this study. Gene *1* codes for T7 RNA polymerase, which is highly specific for phage promoters, being responsible for the expression of class II and III genes and for the translocation of 81% of the phage genome into the bacterium. Gene *10* codes for the major capsid protein gp10A and, by programmed ribosomal frameshift originates the minor capsid protein gp10B. Both are assembled into wild-type particles but either alone suffices for viability. Being essential, gene *1* has a lower proportion of differences which makes this a poor gene for phylogenetic inference. As expected gene *10A* has a lower proportion of differences too, but after the frameshift, that permits the transcription of gene *10B,* we see a greater accumulation of mutations than expected (27 different mutations were seen in the first 441 bp of gene *10* and 19 mutations in the 124 bp sequenced after the frameshift). To ensure that at least the smaller capsid protein maintained its structure with relatively few errors was probably enough for phage viability.

Gene *10B* produced by far the most accurate trees, almost as well as the global analysis. Probably this result arises from stochastic effects, since neither of the above reasons seems to justify it. Despite its exceptional performance, gene *10B* encountered the same problems of misplacing lineages H, I and J.

A great part of T7's genome was well represented by restriction site data, since a large set of restriction sites was determined for each individual. This kind of information (except for Sau3AI recognition sites) is less prone to sampling errors that arise from bad choices of genes. In fact, at least in this case, joining the two data sets (restriction and nucleotide) improved the accuracy of the analysis except for the data set composed of all the enzymes but Sau3AI (Tables 1 and 3).

The overall results obtained with the symmetric topology are illustrated in Fig. 4. The percentage of branches correctly inferred was usually very similar for both topologies (except for gene *3.5*, for which the symmetric topology performed much worse). Branches 3–4 and 3–5 were poorly resolved by both topologies as well as branch 8–10 (9–10 in the asymmetric tree) which was still consistently misinferred by many of the algorithms. Lineages I and J differ from the wild-type (wt, common ancestor of all lineages) in 38 and 47 positions, respectively (70 mutations were expected given the distance between these nodes and the wt). The observed and expected number of differences between all the terminal nodes and the wt were considered significantly different by a χ^2 test ($p < 0.005$), but this difference became not significant if lineages I and J were removed from the analysis. Taken together these results may indicate that the sampled sequences of lineages I and J evolved at a sufficiently lower rate as to confound the majority of the inference methods even when a simple topology was considered.

It seems plausible to conclude that in this case, short branches were more difficult to infer than a less symmetric branching pattern. That is, even a symmetric branching pattern could not prevent short branches from being consistently misinferred by the generality of the methods.

3.1.2. Branch lengths and ancestral states

In order to assess the accuracy of branch length estimates a correlation analysis between estimated and known/observed branch lengths was done (Annex 4-A). The branch score (*Bs*; Kunher and Felsenstein, 1994) between the inferred trees and the true tree (with known/observed branch lengths) was also calculated (Annex 4-B). This distance reflects simultaneously topological and branch length differences between trees (the branch score value increases with the distance between trees).

The correlation (*r*), for all the data partitions, was always greater (and conversely the branch score was always smaller) when the branch lengths were measured in restriction site differences (observed lengths) than in number of lysates (real lengths). Estimates of branch lengths are known to be particularly sensitive to the choice of model so the evaluation of the reliability of an estimated tree may be misleading if oversimplified models are used (Leitner et al., 1997). This was particularly true for UPGMA, which always produced the worst estimates of branch lengths regardless of the distance measure, data partition or branch length units (correlation ranging from 0.4 to 0.7). Total and mean differences distance usually gave better estimates of branch lengths (and smaller *Bs*) than Upholt and Nei–Li distances, especially when the branches were measured in number of lysates. Euclidean distance was the worst distance measure. Distance methods were all fairly equivalent but slightly worse than MP and ML. Overall restriction site ($r = 0.53$–0.75, number of lysates; $r = 0.8$–0.96, number of differences) and combined data sets produced better estimates of branch lengths than sequence data ($r = 0.46$–0.69, number of lysates; $r = 0.63$–0.83, number of differences).

Concerning the branch score results, UPGMA was fairly equivalent to the other methods since this measure reflects also the d_T and this was the algorithm that more often inferred the true tree. Nevertheless having a similar *Bs* but a higher d_T means that branch lengths were more accurately estimated. This explains the results for some distance methods and MP. Like in the correlation analysis, Euclidean distance was the worst distance measure, leading to greater *Bs* values even when the correct topology was inferred.

Parsimony correctly inferred 97.4% of ancestral states (12 internal nodes) from restriction site data.

4. Conclusions

The purpose of this study was to compare different data and different tree-building methods with a well studied experimental model in the recovery of a known phylogeny. The novelty of this study consists in the choice of a tree with unequal evolutionary rates and the use of an alignment free method such as the sequence signatures that still is in a primordial phase in the phylogenetic context.

Parametric bootstrapping offers a method of producing independent replicates of observed data sets, which can be used to test the performance of competing methods or to extend the conclusions of experimental studies (Hillis, 1995). Though this might be a valuable tool in predicting in what aspects most methods will fail, our results show that there will always be some details of "natural history" difficult to incorporate in simulation.

Overall restriction site produced more accurate trees in respect to topology and branch length estimates than nucleotide data. This is not an unexpected result since the former data represents the genome more broadly than sequence data. Only 12% of the genome was sequenced, so the bias coming from sampling errors is probably larger than from restriction sites.

For nucleotide data (allseqs) only methods that assume a molecular clock (UPGMA), had a molecular clock enforced (ML and ME—using FITCH from PHYLIP, data not shown) or used sequence signature based distance recovered the true tree. These

results might seem in disagreement with the work published by Cunningham et al. (1997, 1998) where they reported a superior performance of ML in relation to ME and a very important role for the correction for among-site rate variation in overcoming long-branch attraction. However, our work and theirs differ in the experimental protocol. Despite using the same phage they bottlenecked it every 50 lysates while we did it every 5 lysates. Botlenecking so often diminishes the action of natural selection and increases the effect of genetic drift which allows the system to evolve more like a molecular clock and explains the overall success of the algorithms that assume this kind of evolution.

For restriction site (all enzymes but Sau3AI) all the distance methods infer the true tree except uLS. So we can conclude that when all sequences are considered, distance methods (UPGMA being the best) performed better for both data sets.

If we consider the individual performance of each gene, then the order is almost reversed with UPGMA and NJ being the worst, ML and MP the best and the others intermediate. It is important to note that many of the high values of d_T produced for most of the genes are due to polytomies (lack of resolution) rather than to errors in the branching pattern.

A previous study (Hillis et al., 1994) compared the ability of methods to infer the correct phylogeny from restriction sites versus nucleotide sequences. Restriction sites proved to be somehow superior, yet it was partly attributed to the fact that the number of variable sites almost tripled the number of variable positions in the nucleotide sequence. In the present work, the number of variable restriction sites (304) for the symmetric tree was equivalent to variable nucleotide positions (312), nevertheless the performance of restriction sites was undoubtedly superior since all the methods produced the correct tree. As stated before (Hillis et al., 1994) this might be explained by the independence of evolution (an assumption of most of the methods) being less affected in restriction sites than in nucleotide sequences. Although restriction mapping implies a much bigger effort than sequencing it may be rewarding.

We must also emphasize the performance of sequence signature based distance in the global analysis. As stated before we are convinced that the mutation bias of this system (page T7 propagated in the presence of NG) can shift motif frequencies and this may be reflected in the Euclidean distance matrices. Sequence signature has been described as a fast tool for exploring phylogenetic data since it avoids the alignment step and allows the use of numerous sequences of varying size that need not be homologous. It has also been demonstrated that long word frequencies describes DNA sequence information more accurately, but long words are difficult to apply to short sequences because word frequencies are poorly estimated (Chapus et al., 2005). In the system described in this paper, all the sequences are homologous, their mean pairwise proportion of differences in percent is small (about 2.4%) and the genes size is also small between 150 bp and 565 bp. In spite of these limiting features, sequence signatures based methods were the only methods (apart from those assuming a molecular clock) capable of inferring the true phylogeny, even for a small gene like *10B*. Another interesting result that supports our findings was the fact that sequence signatures based methods (NJ, ME, uLS) (data not shown) were able to infer the true tree with 1091 bp (Hillis et al., 1994) of Hillis et al. (1992) phylogeny. In the study of Hillis et al. (1994) all the tested methods (UPGMA, NJ and ML) except MP (that produced two equally parsimonious trees, one of which was the true tree) inferred an incorrect tree.

There are several other systems that have similar mutation spectrums, such as eukaryotic pseudogenes (Gojobori et al., 1982; Li et al., 1984) and HIV virus (Leitner et al., 1997; Moriyama et al., 1991; Vartanian et al., 1991), and preliminary studies (data not shown) have suggested that this might also be a good approach for these cases.

Nevertheless some other method like Bayesian or maximum-likelihood methods must be used to infer branch lengths since the sequence signature performed very poorly in this matter.

Acknowledgment

A. Sousa thankfully acknowledges the financial support by Grant PRAXIS XXI/BD/19793/99 of Fundação para a Ciência e Tecnologia (FCT), Portugal.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2008.04.030.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. AC-19, 716–723.
Granoff, A., Webster, R.G., 1999. Encyclopedia of Virology. Three volume Set, 1–3, second ed., Elsevier.
Beletskii, A., Grigoriev, A., Joyce, S., Bhagwat, A.S., 2000. Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. J. Mol. Biol. 300, 1057–1065.
Bull, J.J., Cunningham, C.W., Molineux, I.J., Badgett, M.R., Hillis, D.M., 1993. Experimental molecular evolution of bacteriophage T7. Evolution 47, 993–1007.
Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation procedures. Am. J. Hum. Genet. 19, 233–257.
Chapus, C., Dufraigne, C., Edwards, S., Giron, A., Fertil, B., Deschavanne, P., 2005. Exploration of phylogenetic data using a global sequence analysis method. BMC Biol. Evol. 5, 63.
Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12, 814–822.
Cunningham, C.W., Jeng, K., Husti, J., Badgett, M., Molineux, I., Hillis, D.M., Bull, J.J., 1997. Parallel molecular evolution of deletions and nonsense mutations. Mol. Biol. Evol. 14, 113–116.
Cunningham, C.W., Zhu, H., Hillis, D.M., 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52, 978–987.
Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature, characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.
Dunn, J.J., Studier, F.W., 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J. Mol. Biol. 166, 477–535.
Eck, R.V., Dayhoff, M.O., 1966. Atlas of protein sequence and stucture. Silverspring, Maryland. Natl. Biomed. Res. Found.
Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1994. Testing significance of incongruence. Cladistics 10, 315–319.
Felsenstein, J., 1981. Evolutionary trees from DNA sequences, a maximum likelihood approach. J. Mol. Evol. 17, 368–376.
Felsenstein, J., 2006. PHYLIP (Phylogeny Inference Package). Department of Genetics, University of Washington, Seattle. (distributed by the author).
Fitch, W.M., Margoliash, E., 1967. Construction of phylogenetic trees. Science 155, 279–284.
Gojobori, T., Li, W.H., Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18, 360–369.
Heineman, R.H., Molineux, I.J., Bull, J.J., 2005. Evolutionary robustness of an optimal phenotype, re-evolution of lysis in a bacteriophage deleted for its lysin gene. J. Mol. Evol. 61, 181–191.
Hillis, D.M., 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44, 3–16.
Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., Molineux, I.J., 1992. Experimental phylogenetics, generation of a known phylogeny. Science 255, 589–592.
Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R., Molineux, I.J., 1993. Experimental approaches to phylogenetic analysis. Syst. Biol. 42, 90–92.
Hillis, D.M., Huelsenbeck, J.P., Cunningham, C.W., 1994. Application and accuracy of molecular phylogenies. Science 264, 671–677.
Holder, M., Lewis, P.O., 2003. Phylogeny estimation, traditional and Bayesian approaches. Nat. Rev. Genet. 4, 275–284.
Huelsenbeck, J.P., 1991. Tree-length distribution skewness, an indicator of phylogenetic information. Syst. Zool. 40, 257–270.
Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673–688.
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES, Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.
Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29, 170–179.
Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980–984.

Kunher, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.
Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., Albert, J., 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA 93, 10864–10869.
Leitner, T., Kumar, S., Albert, J., 1997. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J. Virol. 71, 4761–4770.
Li, W.H., Wu, C.I., Luo, C.C., 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. 21, 58–71.
Lyons-Weiler, J., Takahashi, K., 1999. Branch length heterogeneity leads to nonindependent branch length estimates and can decrease the efficiency of methods of phylogenetic inference. J. Mol. Evol. 49, 392–405.
Moriyama, E.N., Ina, Y., Ikeo, K., Shimizu, N., Gojobori, T., 1991. Mutation pattern of human immunodeficiency virus gene. J. Mol. Evol. 32, 360–363.
Mossel, E., Vigoda, E., 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309, 2207–2209.
Nei, M., Kumar, S., Takahashi, K., 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc. Natl. Acad. Sci. USA 95, 12390–12397.
Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76, 5269–5273.
Nylander, J.A.A., 2004. MrModeltest. Evolutionary Biology Centre. Uppsala University.
Penny, D., Hendy, M.D., 1985. The use of tree comparison metrics. Syst. Zool. 34, 75–82.
Planet, P.J., 2006. Tree disagreement, measuring and testing incongruence in phylogenies. J. Biomed. Inform. 39, 86–102.
Posada, D., Crandall, K.A., 1998. MODELTEST, testing the model of DNA substitution. Bioinformatics 14, 817–818.
Prager, E.M., Wilson, A.C., 1988. Ancient origin of lactalbumin from lysozyme, analysis of DNA and amino acid sequences. J. Mol. Evol. 27, 326–335.
Rambaut, A., Drummond, A., 2005. Tracer, MCMC Trace File Analyser. University of Oxford.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3, Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Russo, C.A., Takezaki, N., Nei, M., 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13, 525–536.
Rzhetsky, A., Nei, M., 1992. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. J. Mol. Evol. 35, 367–375.
Saitou, N., Nei, M., 1987. The neighbor-joining method, a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Sanson, G.F., Kawashita, S.Y., Brunstein, A., Briones, M.R., 2002. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. Mol. Biol. Evol 19, 170–178.
Sober, E., 1993. Experimental tests of phylogenetic inference methods. Syst. Biol. 42, 85–89.
Sokal, R.R., Michener, C.D., 1958. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 38, 1409–1438.
Springman, R., Badgett, M.R., Molineux, I.J., Bull, J.J., 2005. Gene order constrains adaptation in bacteriophage T7. Virology 341, 141–152.
Steinbachs, J.E., Schizas, N.V., Ballard, J.W., 2001. Efficiencies of genes and accuracy of tree-building methods in recovering a known Drosophila genealogy. Pac. Symp. Biocomput., 606–617.
Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods). Sunderland, Massachusetts, Sinauer Associates.
Templeton, A.R., 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution 37, 221–244.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface, flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.
Vartanian, J.P., Meyerhans, A., Asjo, B., Wain-Hobson, S., 1991. Selection, recombination, and G(A hypermutation of human immunodeficiency virus type 1 genomes. J. Virol. 65, 1779–1788.
Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19, 513–523.