Available online at www.sciencedirect.com

SCIENCE DIRECT®

Infection, Genetics and Evolution 6 (2006) 323–330

Infection, Genetics and Evolution

www.elsevier.com/locate/meegid

A computer simulation analysis of the accuracy of partial genome sequencing and restriction fragment analysis in the reconstruction of phylogenetic relationships

Baozhen Qiao [a,b], Tony L. Goldberg [a,c], Gary J. Olsen [d], Ronald M. Weigel [a,e,*]

[a] *Division of Epidemiology, Department of Pathobiology, University of Illinois, 2001 South Lincoln Avenue, Urbana, IL 61802, USA*

[b] *Division of Epidemiologic Studies, Illinois Department of Public Health, 605 West Jefferson Street, Springfield, IL 62761, USA*

[c] *Program in Ecology and Evolutionary Biology, School of Integrative Biology, University of Illinois, 505 South Goodwin, Urbana, IL 61801, USA*

[d] *Department of Microbiology, University of Illinois, 601 South Goodwin, Urbana, IL 61801, USA*

[e] *Department of Community Health, University of Illinois, 1206 South Fourth Street, Champaign, IL 61820, USA*

Received 27 June 2005; received in revised form 23 October 2005; accepted 28 October 2005
Available online 6 January 2006

Abstract

Partial genome sequencing (PGS) and restriction fragment analysis (RFA) are used frequently in molecular epidemiologic investigations. The relative accuracy of PGS and RFA in phylogenetic reconstruction has not been assessed. In this study, 32 model phylogenetic trees with 16 extant lineages were generated, for which DNA sequences were simulated under varying conditions of genome length, nucleotide substitution rate, and between-site substitution rate variation. Genotyping using PGS and RFA was simulated. The effect of tree structure (stemminess, imbalance, lineage variation) on the accuracy of phylogenetic reconstruction (topological and branch length similarity) was evaluated. Overall, PGS was more accurate than RFA. The accuracy of PGS increased with increasing sequence length. The accuracy of RFA increased with the number of restriction enzymes used. In fragment size comparison, the Dice and Nei–Li algorithms differed little, with both more accurate than the Fragment Size Distribution algorithm. For RFA, higher tree stemminess and longer genome length were associated with higher topological accuracy, whereas lower tree stemminess and lower substitution rates were associated with higher branch length accuracy. For PGS, lower tree imbalance was associated with higher topological accuracy, whereas lower tree stemminess, higher substitution rate, and lower between-site substitution rate variation were associated with higher branch length accuracy. RFA had higher topological accuracy than PGS only for the shortest sequence length (200 bps) at a low substitution rate, high tree stemminess, and long genome length. PGS had equal or higher accuracy in branch length reconstruction than RFA under all conditions investigated. Thus, partial genome sequencing is recommended over restriction fragment analysis for conditions within the parameter space examined.

Keywords: Partial genome sequencing; Restriction fragment analysis; Computer simulation; Phylogenetic reconstruction; Disease transmission; Molecular epidemiology

1. Introduction

Accurate reconstruction of phylogenetic relationships between microorganisms is important in determining sources of infection and patterns of transmission in epidemiologic studies (Olive and Bean, 1999). Complete genome sequencing provides the most complete information for phylogenetic reconstruction, yet is unfeasible in most circumstances. Alternative genotyping methods include partial genome sequencing (PGS) and restriction fragment analysis (RFA); both examine only part of the genome.

RFA compares the number and sizes of fragments produced by digestion of DNA with restriction endonuclease enzymes (Dowling et al., 1990). Several quantitative algorithms have been used in RFA. The Dice distance (Dice, 1945; Lynch, 1990) and Nei–Li distance (Nei and Li, 1979) are based on the

* Corresponding author. Tel.: +1 217 244 1365; fax: +1 217 244 7421.
E-mail address: weigel@uiuc.edu (R.M. Weigel).

1567-1348/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2005.10.002

proportion of fragments matching in size. To account for error in estimating fragment sizes, a tolerance range is included in defining a match (Gill et al., 1991; Akbari et al., 2002). In contrast, Weigel and Scherba (1997) developed an algorithm that compares the distribution of fragment sizes in estimating genetic similarity between samples, without relying on determination of fragment matches. Although all three algorithms have been used in research, their relative accuracy in reconstructing phylogenetic relationships is unknown.

Accuracy of phylogenetic reconstruction may be dependent upon characteristics of the genetic material analyzed, e.g., genome or sequence length, nucleotide substitution rate, and between-site variability in substitution rate (Sourdis and Krimbas, 1987; Jin and Nei, 1991; Kuhner and Felsenstein, 1994; Yang, 1996; Buckley et al., 2001), as well as characteristics of the true phylogeny, e.g., variation in nucleotide substitution rates among lineages and associated characteristics of tree structure (Fiala and Sokal, 1985; Rohlf et al., 1990; Hillis, 1995).

The relative accuracy of PGS and RFA in estimating genetic relationships has rarely been evaluated. Qiao and Weigel (2004), using 16 completely sequenced papillomavirus isolates from Genbank, found in a computer simulation analysis comparing genetic distance matrices that PGS consistently had a higher correlation with complete genome sequencing than did RFA. The computer simulation study reported below extends the analysis of Qiao and Weigel (2004) by examining the comparison of PGS and RFA within the context of the accuracy of phylogenetic construction (which cannot be evaluated with field samples with an unknown true phylogeny), taking into account the impact of variation in genomic characteristics and phylogenetic tree structure on this accuracy. For RFA, differences among algorithms estimating genetic similarity are also evaluated. The relevance of these results for epidemiologic studies is addressed.

2. Materials and methods

2.1. Definition and selection of model trees

Each model tree, consisting of 16 extant lineages, was randomly generated using a Markovian branching process (Martins, 1996; Housworth and Martins, 2001), modified to achieve a wider range of tree characteristics. This included using different mathematical functions to generate branches between bifurcation events (i.e., using the uniform or exponential functions with the expected time interval between one bifurcation and the next equal to $1/k^2$, $1/k$, $1/\sqrt{k}$, $\sqrt{k}$, and k, respectively, where k was the number of lineages present at that time interval), and modeling branch length as a random variable with different gamma distributions (Takahashi and Nei, 2000). Branch length was measured in units of number of nucleotide substitutions. A total of 12,000 model trees were generated. The structure of each tree was described using three quantitative measures—lineage coefficient of variation, tree imbalance, and tree stemminess.

Lineage coefficient of variation (LCV) was defined as the variation in overall nucleotide substitution rates (from tree root to tip) among different lineages: $\mathrm{LCV} = \left(\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_n})^2/(n-1)}\right)/\overline{x_n}$, where n is the number of lineages in a model tree, x_i is the substitution rate from root to tip for lineage i, and $\overline{x_n}$ is the average substitution rate from root to tip for all lineages. Higher LCV values indicate greater variation in substitution rates among lineages. LCV = 0 indicates a constant substitution rate among lineages, reflected in a tree with equal branch lengths.

Tree imbalance measures the degree to which branch points define subgroups of equal size (Shao and Sokal, 1990). The Colless tree imbalance index (Colless, 1982) ranges from 0 to 1, with 0 indicating a perfectly balanced tree, and 1 indicating a completely imbalanced tree.

The non-cumulative tree stemminess index (Rohlf et al., 1990) is a measure of the average distinctness of all the taxonomic subsets on a tree, which is associated with the relative position of the internal nodes (Rohlf et al., 1990). For trees with varying rates of substitution among lineages ($\mathrm{LCV} \neq 0$), the height of an internal node was calculated as the average length from this node to all its extant descendants. The tree stemminess index ranges from 0 to 1, with higher values indicating more stemminess.

From the 12,000 model trees generated, 32 were selected for further study as follows: the range of values for each tree parameter was divided into three levels with equal intervals. For three tree parameters, this resulted in 27 cells in three dimensions. From each cell, a tree was randomly selected. In addition, five trees with extreme values for tree parameters were also selected.

2.2. Generation of genomic DNA sequences

DNA sequences were generated using the Seq-Gen software (Rambaut and Grassly, 1997). For each model tree, an ancestral sequence with a specified length was randomly generated under conditions of equal probability for each of the four nucleotide bases. The ancestral sequence evolved according to the branching pattern and branch lengths of the model tree. Nucleotide substitutions were assumed to follow the Kimura two-parameter model (1980) with a 2:1 transition/transversion ratio.

For each of the 32 model trees, phylogenies were simulated for varying genetic characteristics. Genome length assumed values of 3000, 9000, and 15,000 bps, respectively, which are within the range of genome lengths of small viruses (Cann, 2001). The average substitution rate from common ancestral to extant DNA samples was alternately set equal to 0.025, 0.05, 0.1, and 0.2 substitutions per nucleotide. Finally, heterogeneity among sites in substitution rates was assumed to follow a gamma distribution (Yang, 1993) with shape parameter values of 0.2, 1, and 2, representing strong, moderate, and weak rate variation among sites, respectively (Gu et al., 1995). Thus, for each model tree, there were 36 simulation conditions. Due to the stochastic nature of the simulations, for each condition three replicates were generated to reflect variation in simulation outcomes. Thus, in total, 3456 sets of genomic DNA sequences were simulated.

2.3. Partial genome sequencing analysis

For each phylogeny, three PGS conditions (nucleotide sequence lengths = 200, 600, and 1000 bps) were simulated, with starting positions at bases 100, 500, and 1500 assumed, respectively. For the 16 terminal sequences generated in each lineage, pairwise genetic distances (Kimura, 1980) were calculated using the DNADIST program within the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html).

2.4. Restriction fragment analysis

Using the program DIGEST (http://iubio.bio.indiana.edu/soft/molbio/ibmpc/), digestion of each set of extant genomic DNA sequences was simulated for each of three arbitrarily selected restriction endonuclease enzymes with four-base recognition sites—*Acc*II, *Alu*I, and *Mbo*I. Estimates of genetic distance between extant sequences were obtained by comparing fragment patterns for each enzyme using three different quantitative algorithms. Using the Dice coefficient (Dice, 1945): $S_{xy} = 2N_{xy}/(N_x + N_y)$, where N_{xy} is the number of fragments matched in size between genomic sequences x and y, N_x the number of fragments for sequence x, and N_y is the number of fragments for sequence y; the *D-distance* was calculated as $D_{xy} = 1 - S_{xy}$. The *NL-distance* was derived from the Dice coefficient as described in Nei and Li (1979); it estimates the number of nucleotide substitutions per restriction site separating a pair of organisms. For the D- and NL-distances, the estimated genetic distance over three restriction enzymes was calculated using the same formulas, except that N_x, N_y and N_{xy} were redefined as the sum of the number of fragments and matches over all three enzymes (Nei, 1987). The Fragment Size Distribution (*FSD-distance*) described by Weigel and Scherba (1997), calculates similarity of RFA patterns based on the average differences in fragment sizes. The FSD-distance over three enzymes was calculated as the Euclidean distance in three dimensions.

2.5. Phylogenetic reconstruction

Each genetic distance matrix for each RFA algorithm (with three enzymes combined) and PGS (separately for each of the three sequence lengths) was used to reconstruct a phylogenetic tree, using the neighbor-joining method (Saitou and Nei, 1987). The accuracy of phylogenetic reconstruction was evaluated by comparison with the model trees, using two measures. The partition metric, a measure of the accuracy of the topological structure of a derived tree, counted the number of branches in the reconstructed tree for which there was no equivalent branch on the model tree, and vice versa (Robinson and Foulds, 1981; Penny and Hendy, 1985). The partition metric varies from 0 to 26 for trees with 16 taxa; a lower partition metric value indicates higher accuracy of topological reconstruction. The second measure of accuracy was the branch score (Kuhner and Felsenstein, 1994), which calculated the differences of corresponding branch lengths in the two trees compared. Because branch lengths of the reconstructed and model trees were not always in the same measurement scale, branch lengths were standardized by dividing by the sum of the branch lengths for a tree. A lower branch score indicates higher accuracy in branch length reconstruction.

2.6. Statistical analysis

To evaluate the impact of the genetic factors and model tree parameters on the accuracy of phylogenetic reconstruction for each genotyping condition, multiple linear regression analyses were conducted, with the partition metric or branch score as the dependent variable, and the three genetic factors (genome length, substitution rate, and shape parameter α) and three tree parameters (LCV, tree imbalance, and tree stemminess) as independent variables. The genotyping conditions were the three sequence lengths (200, 600, and 1000 bps) for PGS and the three genetic distance algorithms (D-, NL-, and FSD-distances) for RFA. RFA accuracy was evaluated for the three-enzyme condition.

Subsequently, the relative accuracy of RFA and PGS as a function of the genetic factors and tree parameters was evaluated using a multiple regression model with the arithmetic difference (PGS − RFA) between, alternatively, the partition metric and branch score values as dependent variables. This comparison was made separately for each of the three PGS sequence lengths, using the RFA distance algorithm with the greatest accuracy.

In all regression analyses, model assumptions of normality and homogeneity of residual variance were evaluated, with natural logarithmic transformations of the dependent variable conducted as a corrective measure. Since the sample size for analysis was large and statistical significance could be achieved easily, the importance of each factor in influencing outcome was evaluated based not on its p-value, but on the proportion of variance of outcome uniquely accounted for by each independent variable (i.e., the squared semi-partial correlation coefficient, sr^2) A variable with $\mathrm{sr}^2 \geq 0.05$ was considered an important factor influencing the outcome.

3. Results

Table 1 lists the mean partition metric and branch scores comparing trees generated under the three RFA distance measures to the model trees. Using a single enzyme, the D-distance had the lowest and the FSD-distance the highest values for the partition metric and branch score. There were no differences among restriction enzymes. However, using the three-enzyme distance score resulted in a decreased partition metric for all three distance measures, and in a decreased branch score for the matching algorithms (D- and NL-distances), but a greatly increased branch score for the FSD-distance algorithm.

For RFA, higher tree stemminess resulted in a lower partition metric for all three distance measures, and longer genome length resulted in a lower partition metric for the D- and NL-distances (Table 2). For the D- and NL-distances, tree

Table 1
Mean partition metric and branch scores for the different restriction enzymes with the distance measures

Restriction enzyme	Mean partition metric			Mean branch score		
	D-distance	NL-distance	FSD-distance	D-distance	NL-distance	FSD-distance
*Acc*II	12.95 (6.40)	14.16 (6.20)	17.89 (5.04)	19.09 (9.45)	20.00 (13.23)	28.27 (14.80)
*Alu*I	12.84 (6.34)	14.19 (6.15)	17.84 (5.04)	19.11 (9.70)	20.39 (14.09)	27.57 (13.39)
*Mbo*I	12.79 (6.40)	14.09 (6.23)	17.83 (5.12)	19.00 (9.31)	20.06 (13.93)	27.52 (14.52)
Enzymes combined	9.42 (6.23)	10.06 (6.15)	14.43 (5.97)	16.06 (8.82)	12.06 (5.87)	76.23 (24.76)

D-distance, Dice distance; NL-distance, Nei and Li distance; FSD-distance, Fragment Size Distribution distance. Standard deviation in parentheses.

stemminess and substitution rate were positively correlated with branch score, with tree stemminess having the greater effect. Genome length was negatively correlated with branch score for the NL-distance. LCV was positively correlated with branch score for the D-distance.

The mean partition metric and branch score were 11.15 and 12.31 for PGS with 200 bps, 7.20 and 7.77 with 600 bps, and 5.96 and 6.43 with 1000 bps, indicating increased accuracy of phylogenetic reconstruction as sequence length increased. Tree imbalance was positively correlated with the partition metric for all three sequence lengths (Table 3). For 200 bps only, an increased substitution rate resulted in a lower partition metric. Higher tree stemminess resulted in a higher branch score for all three sequence lengths. A higher substitution rate was associated with lower branch scores at 200 and 600 bps only; this effect decreased with longer sequence length. The shape parameter α was negatively correlated with branch score at 600 and 1000 bps. The effect of this parameter increased as sequence length increased.

The regression analyses comparing the relative accuracy of phylogenetic reconstruction for PGS and RFA (using the accurate and most commonly used D-distance) identified the following (Table 4; Fig. 1). The most obvious trend was that for PGS both topological and branch length accuracy increased with longer sequence length, with a greater improvement from 200 to 600 bps than from 600 to 1000 bps. Among the model variables, the most consistent trend was increased accuracy of PGS compared to RFA with a higher substitution rate with these trends stronger for the branch score ($\text{sr}^2 > 0.20$) than for the partition metric ($\text{sr}^2 < 0.15$). For PGS, accuracy of topology and branch lengths both increased with a higher substitution rate. For RFA, topological accuracy was independent of substitution rate (Fig. 1a), whereas accuracy of branch lengths decreased with a higher substitution rate (Fig. 1b). RFA had better topological accuracy than PGS only at the smallest 200 bps sequence length, for all substitution rates except the highest (0.20). RFA had slightly higher branch length accuracy compared to PGS only at the 200 bps sequence length, and only at the lowest substitution rate (0.025).

Tree stemminess had an impact on the relative accuracy of RFA and PGS, but in different directions for topology and branch length. For PGS, there was a slight increase in topological accuracy with higher stemminess for the shortest sequence length (200 bps), but there was a slight decrease with higher stemminess for the longer sequence lengths (600 and 1000 bps) (Fig. 1c). In contrast, branch length accuracy decreased with higher tree stemminess for all sequence lengths (Fig. 1d). For RFA, topological accuracy increased (Fig. 1c) and branch score accuracy decreased (Fig. 1d) with higher tree stemminess. RFA had greater topological accuracy than PGS at the shortest sequence length (200 bps) for tree stemminess in the moderate range (>0.3), but only at the highest tree stemminess (>0.6) for the longer sequence lengths (600 and 100 bps). With respect to branch length accuracy, PGS had

Table 2
Results of multiple linear regression analysis for the association of genetic factors and model tree parameters with the accuracy of phylogenetic reconstruction for the analytic algorithms for restriction fragment data

Model variables	Partition metric						Branch score					
	D-distance ($R^2 = 0.265$)		NL-distance ($R^2 = 0.286$)		FSD-distance ($R^2 = 0.428$)		D-distance ($R^2 = 0.716$)		NL-distance ($R^2 = 0.523$)		FSD-distance ($R^2 = 0.018$)	
	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2
Substitution rate	0.66	0.00005	1.37	0.00022	9.35	0.011	3.18	0.172^*	2.10	0.099^*	0.48	0.007
Shape parameter (α)	−0.32	0.0015	−0.18	0.00045	−0.11	0.00018	−0.04	0.003	−0.07	0.014	−0.02	0.0009
Genome length	−0.0003	0.072^*	−0.0005	0.126^*	−0.0002	0.019	−0.00002	0.022	−0.00004	0.173^*	0.0000001	0.000001
Lineage coefficient of variation	2.65	0.009	2.47	0.008	0.07	0.000007	0.80	0.119^*	0.45	0.049	0.04	0.0005
Tree imbalance	6.64	0.032	6.56	0.032	4.91	0.019	0.43	0.019	0.12	0.002	−0.004	0.000003
Tree stemminess	−13.05	0.093^*	−11.21	0.071^*	−21.95	0.287^*	2.23	0.402^*	1.30	0.179^*	−0.26	0.009

A natural logarithmic transformation of branch score was performed to correct for violations of model assumptions of normality and homogeneity of variance of residuals. After transformation, model violations were corrected and model fit was improved. D-distance, Dice distance; NL-distance, Nei and Li distance; FSD-distance, fragment size distribution distance; R^2, multiple R-square; β, regression coefficient; sr^2, squared semi-partial correlation coefficient.

[*] Indicates importance of variable based on criterion of $\text{sr}^2 \geq 0.05$ (all with $p < 0.001$).

Table 3

Results of multiple linear regression analysis of the association of genetic factors and model tree parameters with the accuracy of phylogenetic reconstruction for the partial genome sequencing conditions

Model variables	Partition metric						Branch score					
	Sequence length: 200 bps ($R^2 = 0.223$)		Sequence length: 600 bps ($R^2 = 0.159$)		Sequence length: 1000 bps ($R^2 = 0.146$)		Sequence length: 200 bps ($R^2 = 0.468$)		Sequence length: 600 bps ($R^2 = 0.430$)		Sequence length: 1000 bps ($R^2 = 0.437$)	
	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2
Substitution rate	−29.17	0.110*	−15.52	0.035	−10.05	0.015	−3.08	0.258*	−2.00	0.10*	−1.28	0.037
Shape parameter (α)	−0.77	0.009	−0.60	0.006	−0.66	0.008	−0.12	0.045	−0.22	0.15*	−0.27	0.21*
Genome length	−0.00001	0.00005	0.000007	0.00003	−0.000003	0.000008	0.0000002	0.000004	0.0000006	0.00004	0.00000001	0.00000003
Lineage coefficient of variation	1.87	0.005	3.38	0.018	3.67	0.022	0.02	0.00017	0.12	0.004	0.19	0.009
Tree imbalance	9.58	0.074*	10.18	0.093*	10.11	0.097*	0.09	0.0014	0.10	0.002	0.14	0.003
Tree stemminess	−2.25	0.003	1.15	0.0009	2.23	0.004	1.10	0.155*	1.20	0.171*	1.29	0.179*

A natural logarithmic transformation of branch score was performed to correct for violations of model assumptions of normality and homogeneity of variance of residuals. After transformation, model violations were corrected and model fit was improved. R^2, multiple R-square; β, regression coefficient; sr^2, squared semi-partial correlation coefficient.

* Indicates importance of variable based on criterion of sr$^2 \geq 0.05$ (all with $p < 0.001$).

Table 4

Results of multiple linear regression analysis evaluating the association of genetic factors and model tree parameters with the difference between partial genome sequencing and restriction fragment analysis in the accuracy of phylogenetic reconstruction

Model variables	Partition metric						Branch score					
	Sequence length of 200 bps vs. RFA ($R^2 = 0.309$)		Sequence length of 600 bps vs. RFA ($R^2 = 0.355$)		Sequence length of 1000 bps vs. RFA ($R^2 = 0.367$)		Sequence length of 200 bps vs. RFA ($R^2 = 0.578$)		Sequence length of 600 bps vs. RFA ($R^2 = 0.579$)		Sequence length of 1000 bps vs.RFA ($R^2 = 0.577$)	
	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2
Substitution rate	−29.83	0.133*	−16.18	0.05*	−10.70	0.023	−91.06	0.386*	−67.83	0.284*	−59.87	0.242*
Shape parameter (α)	−0.45	0.004	−0.28	0.002	−0.34	0.003	−0.91	0.0046	−1.28	0.0123	−1.39	0.0156
Genome length	0.0003	0.088*	0.0003	0.122*	0.0003	0.121*	0.00018	0.0079	0.00018	0.0111	0.00018	0.0111
Lineage coefficient of variation	−0.79	0.001	0.72	0.001	1.02	0.002	−13.04	0.0872*	−11.79	0.0947*	−11.39	0.0966*
Tree imbalance	2.94	0.008	3.54	0.015	3.47	0.015	−6.37	0.0118	−6.74	0.0176	−6.53	0.018
Tree stemminess	10.80	0.082*	14.20	0.18*	15.28	0.218*	−20.76	0.0948*	−24.80	0.18*	−25.90	0.214*

The D-distance provided the most accurate estimates for the topological reconstruction in restriction fragment analysis, and thus, its results were compared with the three partial genome sequencing conditions. The dependent variable in the regression analysis is the difference of the measured accuracy (partition metric or branch score) between each of partial genome sequencing conditions and restriction fragment analysis. R^2, multiple R-square; β, regression coefficient; sr^2, squared semi-partial correlation coefficient.

* Indicates importance of variable based on criterion of sr$^2 \geq 0.05$ (all with $p < 0.001$).

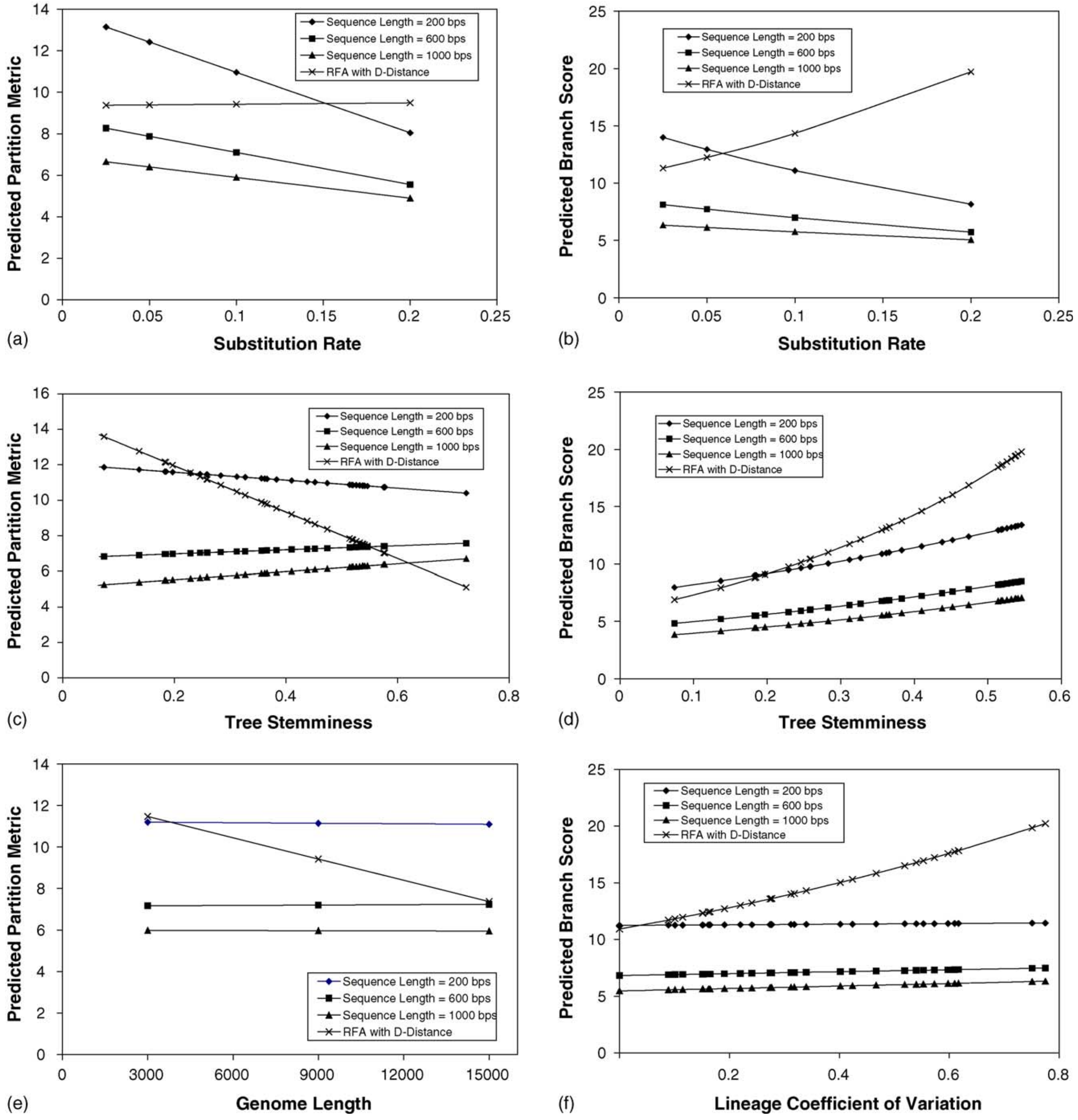

Fig. 1. (a–f) Predicted effect of model variables on the accuracy of phylogenetic reconstruction for restriction fragment analysis using the D-distance and for the partial genome sequencing conditions, as determined from multiple regression analyses.

nearly equal or greater accuracy for all sequence lengths at all values of tree stemminess.

There were differences between RFA and PGS in the effect of genome length on topological accuracy (Table 4, Fig. 1e). For RFA, accuracy increased with longer genome length; for PGS, accuracy was independent of genome length. RFA had higher accuracy than PGS for the shortest sequence length (200 bps) at the higher genome lengths (9000 and 15,000 bps). For all other conditions investigated, PGS had the same or higher accuracy than RFA. As indicated by the sr^2 values, the impact of genome length on the relative topological accuracy of RFA and PGS was greater than the impact of substitution rate, but less than the impact of tree stemminess.

There were differences in the relative branch length accuracy of PGS and RFA as a function of LCV (Table 4, Fig. 1f). RFA accuracy decreased with increasing LCV; PGS accuracy was independent of LCV. Only at low LCV for the shortest 200 bps sequence length did RFA have the same accuracy as PGS; otherwise, PGS was more accurate. However, the sr^2 scores indicate that LCV was less influential than tree stemminess in affecting the relative branch length accuracy of RFA and PGS, and substitution rate was the most influential factor.

4. Discussion

Qiao and Weigel (2004), in using computer simulations of RFA and PGS for 16 completely sequenced papillomavirus isolates to determine the similarity of genetic distance matrices and phylogenetic tree structure, found that PGS agreed more with complete genome sequencing than did RFA. The simulations conducted here investigated these issues further by simulating evolution within known phylogenies, and specifically examining the relationship of variation in genetic and phylogenetic tree parameters to the accuracy of phylogenetic reconstruction.

The simulations showed that, for partial genome sequencing, longer sequence length was associated with greater accuracy in phylogenetic reconstruction, as measured in both topology and branch lengths. Longer sequences contain more genetic information and less nucleotide sampling error (Kumar and Gadagkar, 2000). Increased accuracy was more apparent as sequence length increased from 200 to 600 bps than from 600 to 1000 bps. This implies that sequencing 200 nucleotide bases may be inadequate for accurate phylogenetic reconstruction, whereas 600 bps may be sufficient. Based on simulations, Kumar and Gadagkar (2000) also found higher accuracy of reconstructed phylogenetic relationships for sequences greater than 500 bps and lower accuracy for shorter sequences.

Accuracy of phylogenetic reconstruction in restriction fragment analysis was dependent upon the distance measure employed. The D- and NL-distance algorithms, based on fragment matching, consistently performed better than the FSD-distance algorithm based on comparison of the distribution of fragment sizes, for both topology and branch length reconstructions. Differences in accuracy between the D- and NL-distances were minor. The FSD-distance measure may be associated with lower accuracy because it uses superfluous information (i.e., comparison of all differences in fragment sizes) in determining genetic similarity. For the D- and NL-distance measures, combining fragment results from three enzymes increased the accuracy of both topology and branch length estimates. Each enzyme detected different genetic variation and thus contributed more to the differentiation of genomes. For the FSD-distance, combining information from three enzymes increased the accuracy of topological reconstruction but, inexplicably, substantially decreased the accuracy of branch length reconstruction.

For restriction fragment analysis, lack of measurement error in estimating fragment sizes was assumed. The FSD-distance measure was derived to take measurement error into account (Weigel and Scherba, 1997). Thus, the degree to which these results generalize to the more realistic condition of laboratory measurement error is unknown.

The primary purpose of this investigation was to determine whether restriction fragment analysis or partial genone sequencing has greater accuracy in phylogenetic reconstruction. With respect to branch length reconstruction, PGS achieved equal or greater accuracy than RFA under essentially all conditions examined. Thus, PGS is recommended over RFA whenever estimating the degree of genetic change since divergence from a common ancestor is important.

Even considering only topological reconstruction, RFA was more accurate than PGS only under genomic conditions of low nucleotide substitution rate and large genome size, and the phylogenetic condition of high tree stemminess, with this last factor being the most influential. Conditions under which this could be achieved in nature are rare. A low nucleotide substitution rate in a lineage would be associated with low directional selection (i.e., a constant environment). High tree stemminess represents separate recent radiations from distantly related ancestors. These criteria might be met under certain conditions, e.g., sampling from separate populations where within population divergence was recent but the common ancestor for all populations existed in the distant past. The long-term separation of lineages may have resulted from geographic separation after individual migration events, although incomplete sampling could also be responsible.

In epidemiologic investigations, genotyping of pathogenic microorganisms is used frequently to make inferences about disease transmission. Thus, the relative accuracy of genotyping methods has important implications for disease control programs. In the simulations conducted, RFA had accuracy equal to or greater than PGS only for topological reconstruction under conditions of high tree stemminess, low nucleotide substitution rates, and large genome size. High tree stemminess might occur when separate epidemics are compared (e.g., human influenza outbreaks from separate years). However, the sudden appearance of disease implies exposure of a susceptible population to a pathogen, which is usually due to rapid genetic change in the pathogen, and thus, is outside the advantageous conditions for RFA established in the simulations. Thus, it is difficult to identify conditions under which RFA would be advantageous over PGS for epidemiologic investigations. Also, increasing the sequence length for PGS would eliminate any advantage for RFA.

If there are any conditions where RFA would gain an advantage over PGS, these might be under genome lengths larger than those examined here. The genome sizes examined were relatively small (≤15,000 bp), in part because of the limitation of computer software in handling longer nucleotide sequences. Bacteria have larger genomes, although many viruses fall within the range of genome sizes investigated.

Laboratory costs may be a limiting factor affecting selection of a genotyping method. The costs of preparation of samples and gel electrophoresis are likely to be similar for RFA and PGS. Increased material costs are incurred in RFA with the use of multiple enzymes. Increased costs for PGS are incurred in sequencing the targeted gene, with these costs varying from institution to institution. Thus, widespread use of the more accurate PGS is dependent on decreased sequencing costs.

Nevertheless, this study, in investigating relationships among factors affecting the accuracy of genotyping methods in phylogenetic reconstruction, has identified within a multi-dimensional parameter space reflecting a subset of natural variation, that partial genome sequencing will be more accurate than restriction fragment analysis in phylogenetic reconstruction almost everywhere in that space.

References

Akbari, A., Albregtsen, F., Lingjærde, O.C., 2002. Adaptive weighted least squares method for the estimation of DNA fragment lengths from agarose gels. Electrophoresis 23, 176–181.

Buckley, T.R., Simon, C., Chambers, G.K., 2001. Exploring among-site rate variation models in maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50, 67–86.

Cann, A.J., 2001. Principles of Molecular Virology, third ed. Academic Press, San Diego.

Colless, D.H., 1982. Phylogenetics: the theory and practice of phylogenetic systematics. Part II [book review] Syst. Zool. 31, 100–104.

Dice, L.R., 1945. Measures of the amount of ecological association between species. J. Ecol. 26, 297–302.

Dowling, T.E., Moritz, C., Palmer, J.D., 1990. Nucleic acids. Part II: restriction site analysis. In: Hillis, D.M., Moritz, C. (Eds.), Molecular Systematics. Sinauer, Sunderland, Massachusetts, pp. 250–317.

Fiala, K.L., Sokal, R.R., 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution 39, 609–622.

Gill, P., Evett, I.W., Woodroffe, S., Lygo, J.E., Millican, E., Webster, M., 1991. Databases, quality control and interpretation of DNA profiling in the Home Office Forensic Science Service. Electrophoresis 12, 204–209.

Gu, X., Fu, Y., Li, W.H., 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12, 546–557.

Hillis, D.M., 1995. Approaches for assessing phylogenetic accuracy. Syst. Biol. 44, 3–16.

Housworth, E.A., Martins, E.P., 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. Syst. Biol. 50, 628–639.

Jin, L., Nei, M., 1991. Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. Mol. Biol. Evol. 8, 356–365.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kuhner, M., Felsenstein, K.J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.

Kumar, S., Gadagkar, S.R., 2000. Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies. J. Mol. Evol. 51, 544–553.

Lynch, M., 1990. The similarity index and DNA fingerprinting. Mol. Biol. Evol. 7, 478–484.

Martins, E.P., 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. Evolution 50, 12–22.

Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. U.S.A. 76, 5269–5273.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Olive, D.M., Bean, P., 1999. Principles and applications of methods for DNA-based typing of microbial organisms. J. Clin. Microbiol. 37, 1661–1669.

Penny, D., Hendy, M.D., 1985. The use of tree comparison metrics. Syst. Zool. 34, 75–82.

Qiao, B., Weigel, R.M., 2004. A computer simulation of the accuracy of partial genome sequencing and restriction fragment analysis in estimating genetic relationships: an application to papillomavirus DNA sequences. BMC Bioinf. 5, 102 (http://www.biomedcentral/1471-2105/5/102).

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS 13, 235–238.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

Rohlf, F.J., Chang, W.S., Sokal, R.R., Kim, J., 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. Evolution 44, 671–684.

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

Shao, K., Sokal, R.R., 1990. Tree balance. Syst. Zool. 39, 266–276.

Sourdis, J., Krimbas, C., 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. 4, 159–166.

Takahashi, K., Nei, M., 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. 17, 1251–1258.

Weigel, R.M., Scherba, G., 1997. Quantitative assessment of genomic similarity from restriction fragment patterns. Prev. Vet. Med. 32, 95–110.

Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analyses. Tree 11, 367–372.