J Mol Evol (2002) 55:65–73
DOI: 10.1007/s00236-001-2304-y

JOURNAL OF MOLECULAR EVOLUTION

rtREV: An Amino Acid Substitution Matrix for Inference of Retrovirus and Reverse Transcriptase Phylogeny

Matthew W. Dimmic,[1] Joshua S. Rest,[2] David P. Mindell,[2] Richard A. Goldstein[1,3,*]

[1] Biophysics Research Division, University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109-1055, USA
[2] Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1055, USA
[3] Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055, USA

Received: 21 September 2001 / Accepted: 11 December 2001

Abstract. Retroviral and other reverse transcriptase (RT)-containing sequences may be subject to unique evolutionary pressures, and models of molecular sequence evolution developed using other kinds of sequences may not be optimal. Here we develop and present a new substitution matrix for maximum likelihood (ML) phylogenetic analysis which has been optimized on a dataset of 33 amino acid sequences from the retroviral Pol proteins. When compared to other matrices, this model (rtREV) yields higher log-likelihood values on a range of datasets including lentiviruses, spumaviruses, betaretroviruses, gammaretroviruses, and other elements containing reverse transcriptase. We provide evidence that rtREV is a more realistic evolutionary model for analyses of the *pol* gene, although it is inapplicable to analyses involving the *gag* gene.

Key words: General reversible transition Markov model — HIV — Lentivirus — Maximum likelihood method — Retroid — *pol* gene

Introduction

Common ancestry for organisms and genes results in the correlation of molecular sequence traits among relatives. Maximum likelihood (ML) models are used to account for these correlations in phylogenetic analyses, and improvements in these models are continually sought. Most such analyses are performed at the DNA level. For protein-coding sequences, however, the DNA-based models are often too simplistic; for example, an A→G substitution in the second position of a Tyr codon can have very different effects on the protein than the same substitution in the third position of a Thr codon. Although some researchers have developed codon-based models (Goldman and Yang 1994; Yang et al. 1998), the most common models for protein evolution are substitution matrices derived from large, general sets of related proteins, such as the JTT model (Jones et al. 1992), and the WAG model (Whelan and Goldman 2001).

It is recognized that different types of proteins are under different selective pressures and might not fit the most general models. While it may be increasingly possible to adjust the parameters of the model to maximize the log-likelihood score for the particular data to be analyzed, a more practical option is to use a substitution matrix that has been developed from a more relevant and specific class of proteins. For example, mitochondrial proteins differ from eukaryotic nuclear-encoded proteins in that most are membrane-bound, their codon table is slightly different, and the recombination rate is lower. To develop their mtREV matrix, Adachi and Hasegawa assumed a set of noncontroversial relationships for a set of mammalian mitochondrial amino acid sequences and adjusted the 189 substitution rate parameters in the model until the like-

*Present address: Siena Biotech SpA, via Florentina 1, 53100 Siena, Italy

Correspondence to: M.W. Dimmic; *email*: mdimmic@umich.edu

Table 1. Taxa used in alignment and phylogeny for optimization of model (see van Regenmortel et al. 2000)

Genus	Species	Abbreviation	Accession	Protein ID
Lentivirus	Feline immunodeficiency virus	FIV		GNLJFP
	Caprine arthritis encephalitis virus	CAEV	NC_001463	NP_04939
	Visna/maedi virus	VISNA	M60610	AAA17529
	Human immunodeficiency virus 1	HIV1u455	M62320	AAA75019
		HIV1eli	K03454	AAA44325
		HIV1mvp5180	L20571	AAA44860
	Human immunodeficiency virus 2	HIV2ben	NC_001722	NP_056838
	Simian immunodeficiency virus	SIVmnd	M27470	AAB49569
		SIVsyk	L06042	AAA74707
		SIVagmtan	U58991	AAC57052
		SIVsun	AF131870	AAD39753
		SIV1hoest	AF075269	AAD12147
		SIVagm155	M29975	AAA91906
		SIVagm3	M30931	AAA91914
		SIVcpzant	U42720	AAB47724
		SIVcpzcam3	AF115393	AAF18575
Alpharetrovirus	Avian leukosis virus	ALV	M37980	AAA91269
	Rous sarcoma virus	RSV	AF052428	AAC08988
Betaretrovirus	Mason-Pfizer monkey virus	MPMV	M12349	AAA47711
	Mouse mammory tumour virus	MMTV	M15122	AAA46542
	Squirrel monkey retrovirus	SMRV	M23385	AAA66453
Deltaretrovirus	Bovine leukemia virus	BLV	K02120	AAA42785
	Primate T-lymphotropic virus 1	HTLV1	D13784	BAA02931
	Primate T-lymphotropic virus 2	HTLV2	M10060	AAB59885
	Primate T-lymphotropic virus 3	STLV3	Y07616	CAA68894
Gammaretrovirus	Feline leukemia virus	FeLVfaids		T10532
	Gibbon ape leukemia virus	GALVsf	NC_001885	NP_056790
	Murine leukemia virus	MLVdg75	NV_002603	NP_068545
		MLVmolon		NNMVIM
Spumavirus	Simian foamy virus type 1	SFV1		P23074
	Simian foamy virus type 3	SFV3	M74895	AAA47796
	Human spumaretrovirus	HSV	U21247	AAB48112
	Feline syncitial virus	FeSFV	U85043	AAC58531

lihood was maximized (Adachi and Hasegawa 1996). This approach has since been repeated for other datasets such as chloroplast-encoded proteins (Adachi et al. 2000).

Retroviral elements are also suitable candidates for development of a more specific model of amino acid evolution as they are under selective pressures different from those represented in standard models. Distinctive features of RNA virus replication include high mutation rates, short generation times, large numbers of progeny, and frequent bottleneck events influencing local population sizes (Domingo and Holland 1997). Potentially rapid rates of sequence change combined with environmental change due to colonization of new host populations or potential transposition events within a genome provide the opportunity for increased natural selection effects. These and other distinctive aspects of retroviral lifestyles may contribute to difficulties in resolving retroviral phylogenies (e.g. Doolittle et al. 1990; Mindell et al. 1995; Herniou et al. 1998; McClure 1999; Hahn et al. 2000). We believe a model for amino acid change optimized on a retrovirus phylogeny will improve the reconstruction of phylogenetic trees and the estimation of phylogenetic parameters such as the dating of branch points between viral lineages.

Here we present and assess a general model which has been optimized for application to the retroviral *pol* gene, which encodes a polyprotein containing the protease, reverse transcriptase (RT), and integrase proteins. RT has also been identified as a stable genomic or parasitic component of most kinds of eukaryote and eubacteria genomes. It is found in many of the genetic elements that inhabit and readily transport themselves within and among genomes, such as retroposons, retroplasmids, retrons, telomerase, retrointrons, and other retrotransposons (reviewed in McClure 1999). RT sequences tend to be conserved over time despite low fidelity in reverse transcription because of strong functional constraints on essential structural motifs found in all RTs. Thus, this model could help in evolutionary analyses across a diverse range of RTs.

Methods

An alignment and phylogeny of inferred amino acid sequences from retroviral *pol* genes (the 'training set') was used to optimize

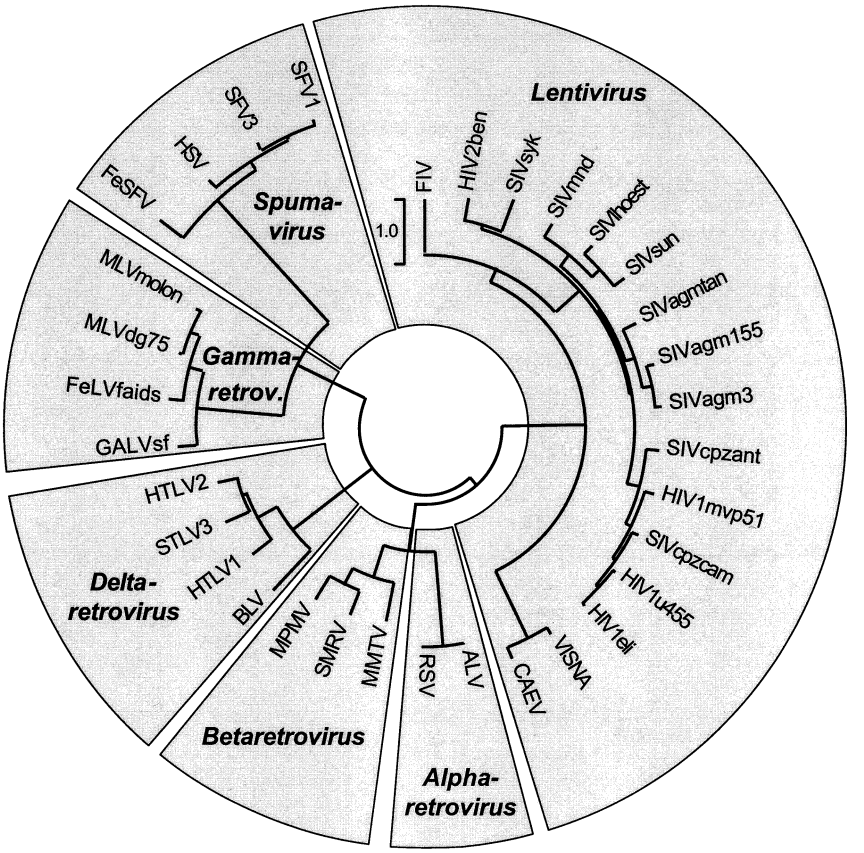

Fig. 1. Phylogeny of retroviruses used for rtREV model optimization (training). Distinctions between retroviral genera are indicated by the shaded regions. See Table 1 for sequence abbreviations.

the model. Thirty-three sequences from alpharetrovirus, betaretrovirus, deltaretrovirus, gammaretrovirus, spumavirus, and lentivirus genera were downloaded from databases (Table 1). Alignment of the *pol* gene (which encodes protease, reverse transcriptase, and integrase proteins) was performed with the aid of ClustalX (Jeanmougin et al. 1998) and included 1335 aligned amino acid positions. The phylogeny used for optimization (Fig. 1) is based on published analyses (Xiong and Eickbush 1990; Tristem et al. 1996; Broussard et al. 1997; Nakamura et al. 1997; Herniou et al. 1998; Beer et al. 1999; Martin et al. 1999; Yin et al. 1999; Hahn et al. 2000). The alignment and phylogenies are available at http://www.umich.edu/~goldgrp/rtREV.

The procedure for optimizing the parameters follows that of Adachi and Hasegawa (1996). Briefly, the probability of substitution from amino acid i to amino acid j in an instant of evolutionary time is given as

$$M_{ij} = \begin{cases} \delta \pi_j R_{ij}/s & (i \neq j) \\ -\delta \sum_{k=1}^{20} \pi_k R_{ik}/s & (i = j) \end{cases} \quad (1)$$

where the relative substitution rate parameters $R_{ji} = R_{ij}$, the diagonals $R_{ii} = 0, \delta = 0.01, \pi_j$ is the frequency of amino acid j estimated from the data, and

$$s = \sum_{i=1}^{20} \sum_{j=1}^{20} \pi_i \pi_j R_{ij} \quad (2)$$

To find the rate of substitution $P_{ij}(t)$ for any particular amount of evolutionary time t, one can then exponentiate the matrix M as

$$\mathbf{P}(t) = e^{t\mathbf{M}} \quad (3)$$

Eight rate classes were used, and their parameters assigned according to a gamma rate distribution (Yang 1994). The model contains 189 adjustable parameters: the relative substitution rate parameters (R_{ij} above, with one value held constant). In addition, the optimization procedure requires 20 amino acid stationary frequencies, 48 tree branch lengths, and 1 rate parameter. The amino acid stationary frequencies were estimated from the data and then fixed. Using the JTT model for initial parameter estimates, the branch lengths and rate parameter were estimated with PAML (Yang 1994). These were then held fixed during the optimization of the 189 substitution rate parameters, which were constrained at a lower bound of 1. The optimization was performed by adjusting the parameters of the model in order to maximize the likelihood of the training dataset (see Felsenstein 1981). The ML estimate of the gamma shape parameter for the training set was 1.41.

The parameters were adjusted using a combination of programs for likelihood calculation written by the authors as well as by using PAML. The optimization package CFSQP v2.5 (Lawrence et al. 1997) was used for the optimization scheme, with derivatives estimated using finite differences. Gaps were treated as missing data in the sequence in which they occur. Following optimization, the branch lengths and rate parameter were re-estimated, but the values did not change significantly.

Testing the Optimized Matrix, rtREV

To examine performance of the rtREV matrix on various subsets of the training phylogeny, four alignments and phylogenies were constructed. In general, these training subsets are clades from the training dataset augmented by the addition of more taxa. All datasets and phylogenies are available at http://www.umich.edu/~goldgrp/rtREV. The augmented training subsets are:

LENTI—*Lentiviral subset of training set*—We tested the model on the subset of lentiviruses from the training tree (see Table 1, Figs. 1 and 2).

BETA—*Betaretrovirus pol*—A betaretrovirus phylogeny for *pol* was constructed from an alignment of eight sequences from the PIR alignment database, ALN entry M04292 (http://pir.georgetown.edu/pirwww/search/textpiraln.html) (Barker et al. 2000).

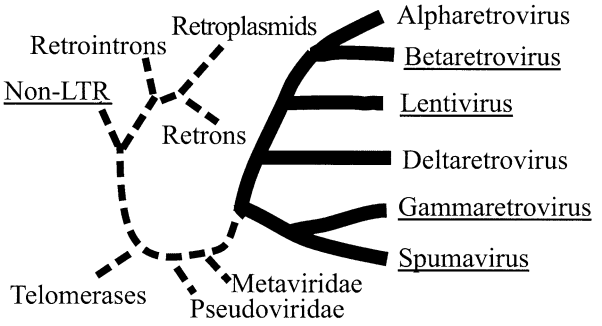

Fig. 2. Diagrammatic phylogenetic hypothesis of the relationship between the analyzed Pol datasets. The rtREV training set is indicated by solid lines. Test sets are underlined: lentivirus (LENTI), betaretrovirus (BETA), spumavirus (SPUMA), gammaretrovirus (GAMMA), and non-LTR retroposon (NONLTR). The endogenous RT (ENDO) test set is indicated with dashed lines.

The alignment includes 928 positions. An exhaustive search using maximum parsimony in PAUP*4.0b4a (Swofford 2000) recovered a single most parsimonious tree 1081 steps in length (Fig. 3a).

SPUMA—*Spumavirus pol*—A spumavirus test phylogeny was constructed by aligning five spumavirus *pol* polyproteins from GenBank with ClustalX. The alignment has 1232 positions, and a single most parsimonious phylogeny 1121 steps in length was recovered from an exhaustive search in PAUP*4.0b4a (Fig. 3b).

GAMMA—*Gammaretrovirus pol*—A gammaretrovirus test phylogeny was constructed by aligning 12 gammaretrovirus *pol* polyproteins from GenBank with the aid of ClustalX. The alignment has 1210 positions, and a single most parsimonious phylogeny 1160 steps in length was recovered from a branch and bound search in PAUP*4.0b4a (Fig. 3c).

In addition, four test alignments and phylogenies were constructed (Fig. 2) to examine performance of the rtREV matrix for sequences and taxa more distantly related to those in the training phylogeny. These "test sets" involve phylogenetic relationships which are non-overlapping with regard to the training set. The performance of rtREV was compared with other models on these four test sets:

ENDO—*Endogenous RT*—Diverse, endogenous retroid elements including telomerase reverse transcriptase, elements from eubacteria, non-LTR elements, and non-retroviral retrotransposons. The alignment and phylogeny is a subset from Xiong and Eickbush (1990) and Nakamura et al. (1997) and includes 47 taxa.

NONLTR—*Non-LTR RT*—An alignment and phylogeny of RT domains from 72 non-LTR retrotransposable elements (Malik et al. 1999). The alignment includes 590 positions.

GAGHIV, GAGGAM—*HIV-gag and Gammaretrovirus-gag*—Phylogenies were also constructed from a *gag* alignment of HIV-1 (49 taxa, 574 characters, PIR ALN entry M04270) and a *gag* alignment of gammaretrovirus from the Megaclass alignment database (12 taxa, 543 characters, http://stateslab.bioinformatics.med.umich.edu/megaclass) using parsimony searches as in datasets BETA, SPUMA, and GAMMA (States et al. 1993).

Log-likelihood scores for each of these topologies under JTT, WAG and rtREV models were calculated in PAML by supplying the topology and alignment. For each test, branch lengths and the gamma shape parameter (eight categories) were optimized, with amino acid frequencies estimated from the data.

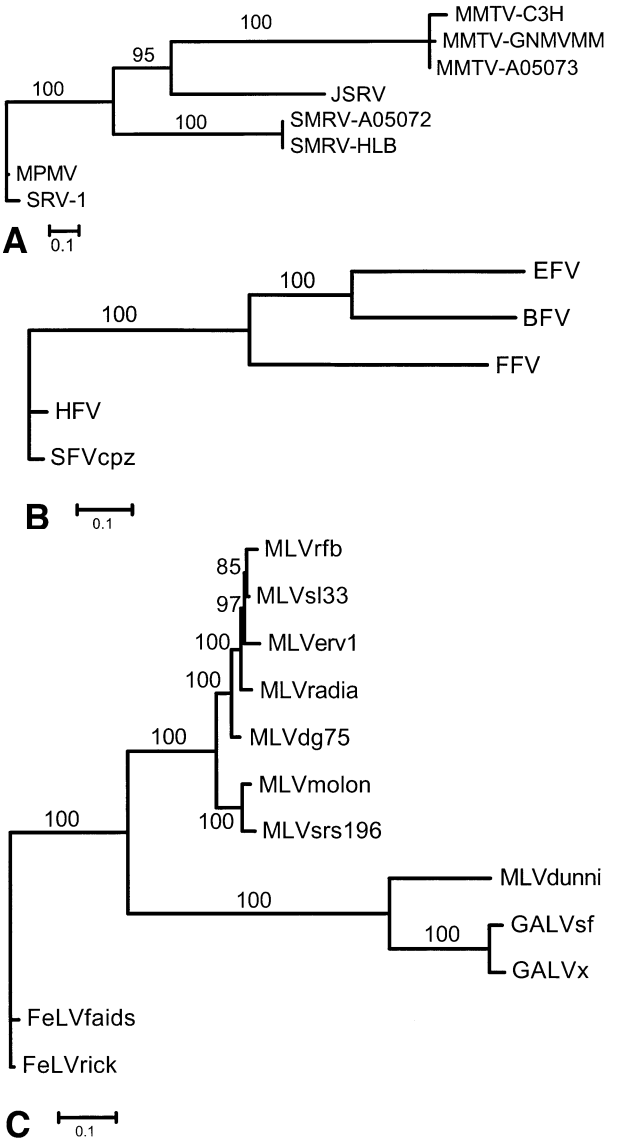

Fig. 3. Maximum parsimony phylogeny of (**A**) betaretrovirus, (**B**) spumavirus, and (**C**) gammaretrovirus Pol proteins. Relative support for each topology is indicated at the nodes by bootstrap values based on 1000 replicates. Branch lengths shown are the ML branch lengths using the rtREV model. Trees are shown here with an arbitrary root, but no rooting was assumed in the analyses. Abbreviations are (**A**) mouse mammory tumor virus (MMTV), Mason-Pfizer monkey virus (MPMV), simian AIDS retrovirus (SRV-1), ovine pulmonary adenocarcinoma virus (JSRV), and squirrel monkey retrovirus (SMRV). Additional details can be found in the text. Abbreviations and accession numbers for (**B**) bovine syncitial virus (BFV, GenBank accession AAB68770), simian foamy virus (SFVcpz, AAA19978), human foamy virus (HFV, CAA69003), feline foamy virus (FFV, CAA70075), equine foamy virus (EFV, NP_054716). (**C**) Murine leukemia virus (MLVrfb, AAK50381; MLVsl33, AAD55051; MLVervl, AAD34536; MLVradia, GNMVRV; MLVdg75, NP_068545; MLVsrsl96, AAC98548; MLVmolon, GNMV1M; MLVdunni, AAC31805), Feline leukemia virus (FeLVrick, NP_047255; FeLVfaids, T10532), Gibbon ape leukemia virus (GALVsf, NP_056790; GALVx, AAC80264).

Comparing Models

On the training dataset, the mathematical form of the rtREV model is equivalent to the form of JTT and WAG. The only difference is that the 189 parameters of rtREV's relative substitution rate matrix (the R-matrix) are adjusted to maximize the likelihood. Therefore, on the training set rtREV is nested with these models with 189 degrees of freedom, and the likelihood ratio test (LRT)

can be used to compare them (e.g., Huelsenbeck and Crandall 1997).

On all other datasets the 189 parameters of rtREV's R-matrix are not adjusted; instead, the values obtained from model optimization on the training set are used. Because there are 0 degrees of freedom, the LRT is not useful when comparing the models on these datasets. In the case of the independent test sets (ENDO, NONLTR, GAGHIV, and GAGGAM), the consistency of the results are compared using a paired *t*-test for means. In this test, the likelihood score (L) is calculated for each column in the alignment for the two models indicated, and the difference between the sitewise L values is calculated. The null hypothesis in this test is that the mean difference is zero, and the *p*-value indicates the significance with which we can reject this null hypothesis. It is similar to a tree-testing method used in the PAUP* software package (Swofford 2000) and described as the PriNPncs test by later authors (Goldman et al. 2000), although here we apply it to model-testing rather than tree-testing.

On the augmented training subsets (LENTI, BETA, SPUMA, and GAMMA), statistical significance is more difficult to assess. These datasets share some sequences with the training set, and therefore are not truly independent, with "true" degrees of freedom somewhere between 0 and 189. Most widely-used methods of model testing in phylogenetics require either specification of degrees of freedom (e.g., the LRT) or the assumption of a null generating model (e.g., parametric bootstrapping), and therefore have questionable applicability on these subsets. We know of no statistical test in phylogenetics which describes how to scale the degrees of freedom appropriately, and the field would certainly benefit from systematic examination of this issue. For this reason, no presumption of significance is made in these cases, and no *P*-value is shown. These augmented training subsets are simply included to examine whether the results on the retroviral training set are consistent across the various genera and with the addition of new sequences.

Results

The Training Set

The final optimized model (rtREV) is shown in Table 2. On the training set, the final log-likelihood (L) score was −46066 for rtREV and −46418 for WAG, an improvement of 352. (The L scores for other models were significantly lower.) To examine whether this improvement is significant, the likelihood ratio test is used. Since in this case rtREV has 189 more parameters than the WAG model, using the LRT WAG can be rejected in favor of rtREV with $>99.99\%$ significance. Although some of the assumptions of the LRT have been questioned in phylogenetic model comparison (Yang et al. 1995; Zhang 1999), other tests lead to the same conclusion. The Akaike Information Criterion (AIC) (Akaike 1974; see also Hasegawa et al. 1990), for example, also suggests that rtREV should be chosen.

The Training Subsets

It is clear that rtREV is a significant improvement over WAG for the training tree. To examine whether rtREV's improvement arises from just one part of the training phylogeny or if the improvement is uniform over the whole phylogeny, four datasets were assembled as described above and tested with several different models. These datasets represent subclades of the training phylogeny which have been augmented with the addition of further taxa. All were assembled using the so-called "model-free" maximum parsimony approach, to avoid biasing the results in favor of one model or another. Tests were also performed on suboptimal trees for some datasets (not shown), with qualitatively similar results.

The results are shown in Table 3a. On each of these subsets, rtREV has a higher L value than either the JTT matrix or the WAG matrix. Note that in each case, the values of the rtREV matrix are frozen and not allowed to adjust. While technically this means that all the models have the same number of adjustable parameters on these subsets, because of their relatedness to the training set it is difficult to estimate the significance of the increase in L in these cases (see Methods). Despite this difficulty, these results do have predictive value; they indicate that the increased likelihood of rtREV is not confined to any particular family of retroviruses in the training set, and that the increase is seen with the addition of new sequences.

The Test Sets

To examine whether rtREV can be applied to retroviruses or other retroid elements outside the training phylogeny, four datasets were assembled as described above and tested with several different models. The taxa used in these sets represent phylogenetic relationships which do not overlap with the training set. For this reason, these datasets arc called "test sets" to differentiate them from the 33-taxa training set where there are 189 degrees of freedom. The results are given in Table 3b, with the phylogenetic relationship between the sets shown in Fig. 2.

Test sets ENDO and NONLTR involve non-retroviral sequences which are relatively distantly related to the training set, and again the likelihood using rtREV is greater than the likelihood using other models. This is evidence that the likelihood increase is broadly distributed across the RT topology (Fig. 2), and that rtREV may be applicable to testing phylogenetic hypotheses across a diverse range of retroids. As with the augmented training subsets, when examining each of these datasets the values of the rtREV matrix are frozen and not allowed to adjust. Therefore JTT, WAG, and rtREV all have the same number of adjustable parameters, and the *p*-values shown indicate whether the sitewise mean increase in L is significant. The mean increase on ENDO appears to be significant, while the significance of the increase on NONLTR is more marginal.

Table 2. The optimized rtREV model

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	—	15	19	11	88	16	14	17	24	—	10	11	62	—	14	47	30	10	20	23
Arg	34	—	25	19	45	33	13	12	34	12	8	47	45	11	10	31	21	11	19	13
Asn	51	35	—	52	62	38	23	20	64	14	9	23	—	22	14	50	31	14	22	14
Asp	10	30	384	—	—	23	48	14	33	0	5	13	—	—	14	32	18	—	—	7
Cys	439	92	128	1	—	38	—	27	83	48	35	22	147	67	—	58	53	45	68	76
Gln	32	221	236	78	70	—	33	9	53	12	10	29	52	—	13	35	25	10	18	14
Glu	81	10	79	542	1	372	—	12	19	—	6	17	—	—	11	23	17	10	0	10
Gly	135	41	94	61	48	18	70	—	21	—	3	8	23	8	6	23	8	10	10	4
His	30	90	320	91	124	387	34	68	—	17	16	24	71	45	15	48	26	28	59	—
He	1	24	35	1	104	33	1	1	34	—	27	9	74	24	6	8	16	18	15	69
Leu	45	18	15	5	110	54	21	3	51	385	—	6	73	25	5	9	9	13	16	18
Lys	38	593	123	20	16	309	141	30	76	34	23	—	38	9	9	20	18	7	—	9
Met	235	57	1	1	156	158	1	37	116	375	581	134	—	94	—	—	52	60	56	68
Phe	1	7	49	1	70	1	1	7	141	64	179	14	247	—	8	21	21	47	114	30
Pro	97	24	33	55	1	68	52	17	44	10	22	43	1	11	—	21	13	7	9	8
Ser	460	102	294	163	75	225	95	152	183	4	24	77	1	20	134	—	61	13	22	15
Thr	258	64	148	55	117	146	82	7	49	72	25	110	131	69	62	671	—	—	18	25
Trp	5	13	16	1	55	10	17	23	48	39	47	6	111	182	9	14	1	—	43	19
Tyr	55	47	28	1	131	45	1	21	307	26	64	1	74	1017	14	31	34	176	—	23
Val	197	29	21	6	295	36	35	3	1	1048	112	19	236	92	25	39	196	26	59	—
Freq	0.0646	0.0453	0.0376	0.0422	0.0114	0.0606	0.0607	0.0639	0.0273	0.0679	0.1018	0.0751	0.0150	0.0287	0.0681	0.0488	0.0622	0.0251	0.0318	0.0619

Parameters are shown in the lower triangle as relative substitution rates, and can be converted to a substitution matrix using the method given in the text. The upper triangle of the matrix shows standard errors estimated by second derivative inversion. Standard errors of parameters at the lower bound were not estimated.

Table 3. Summary of model performance on various datasets: **a:** Model performance on the training tree and on augmented subsets of the training tree; **b:** Model performance on test sets which do not phylogenetically overlap with the training tree

	JTT (L)	WAG (L)	rtREV (L)	ΔL	$<\Delta L/\text{site}>$	p-value
a						
Training tree (*pol*)	−46763	*−46418*	**−46066**	+352	+0.263	<0.001
LENTI (*pol*)	−19759	*−19608*	**−19439**	+169	+0.155	—
BETA (*pol*)	−7497	*−7466*	**−7440**	+26	+0.028	—
SPUMA (*pol*)	−8838	*−8794*	**−8755**	+40	+0.032	—
GAMMA (*pol*)	−9475	*−9472*	**−9467**	+5	+0.004	—
b						
ENDO (*pol*)	−17251	*−17121*	**−17088**	+33	+0.179	0.031
NONLTR (*pol*)	−55865	*−55506*	**−55364**	+142	+0.246	0.083
GAGHIV (*gag*)	**−10904**	−10956	*−10949*	−45	−0.078	<0.001
GAGGAM (*gag*)	**−3393**	−3433	*−3456*	−63	−0.116	0.070

Highest-L model for each set is **bold**, and this is compared with the model in *italics* to obtain the ΔL and associated values for each dataset. p-values are obtained using either the LRT or the paired t-test for means (see Methods).

The *pol* gene is not the only gene used in retroviral phylogenetics; the *gag* (group specific antigen) gene is also commonly used. As a rough examination of whether the evolutionary pressures on *pol* are similar to those on *gag*, test sets GAGHIV and GAGGAM represent topologies identical to those in other test sets, but here sequences from the *gag* gene are used for testing. On these test sets JTT performs better than rtREV, which indicates that rtREV contains *pol* and RT-specific information, and care should be taken not to use rtREV in analysis of *gag* sequences.

Discussion

In examining the substitution parameters of the rtREV model more closely (Table 2), no clear patterns emerge to explain the reasons for its improved ML performance on such a wide variety of retroid datasets. Ideally, an evolutionary model should provide parameters which are biologically meaningful and can lead to testable hypotheses. Such "mechanistic" models are currently being developed on both the codon level and the amino acid level (Yang et al. 1998; Dimmic et al. 2000; Soyer et al. 2002; see also Thorne 2000) and rtREV may serve as a helpful null hypothesis (or "baseline") model in future comparisons.

In what phylogenetic analyses is rtREV most useful? rtREV appears to be the preferred model on every *pol* and retroid-element dataset examined, its performance rivaled by other models only on gammaretroviruses. Perhaps gammaretroviruses have evolved under unique selective pressures relative to the rest of the retroviruses. Another interesting result which bears further scrutiny is that WAG outperforms JTT on every dataset except two, both of which are *gag* datasets. Again, the use of models with more readily-interpretable parameters would be helpful here.

Two assumptions made in the optimization of rtREV are that the sequence alignment is appropriate, and that the original training tree represents an accurate depiction of the evolution of retroviruses. How sensitive is the estimation of model parameters to these assumptions? Small errors in the alignment can be considered analogously to errors in measurement; they should be swamped by the large regions of the sequences which are more easily aligned. Furthermore, the sections of sequence which are the most difficult to align are often the sections undergoing the fastest sequence changes. Since the use of a rate distribution parameter accounts for the higher probability of multiple substitutions in these regions, this should also serve to mitigate the effect of alignment errors on parameter estimates. Regarding possible errors in the training tree topology, while Morrison and Ellis (1997) found that different alignment methods can cause variation in the reconstructed tree, the basic tree structure is still often maintained regardless of the exact alignment (Goldman 1998). Because the likelihood calculation represents a continuous (i.e., non-discrete) process, any small changes in branching order of the tree should not greatly affect the final model. However, the assumptions mentioned should be kept in mind, and sensitivity of results to their violation should be further examined.

The improved performance of rtREV relative to JTT and WAG on the wide variety of retroid test trees derived by different methods and researchers also indicates that any errors in tree topology or alignment did not compromise the final optimized model. This supposition is supported by our observation that several different optimized models for rtREV using slightly different tree structures and alignments (unpublished data) yielded qualitatively similar results.

Conclusions

When performing ML analysis on phylogenetic trees, it is almost always desirable to adjust the substitution probabilities in an effort to maximize the likelihood. When comparing many taxa, however, the computational cost of such adjustment can be prohibitive. The next-best choice is then to use a fixed model with parameters determined on an evolutionarily similar dataset. For researchers interested in phylogenetic questions on retroids including retroviruses, there is compelling evidence that rtREV will provide improvement over models which have been optimized on more general datasets, such as the JTT and WAG models. rtREV appears useful in improving ML analyses of the *pol* gene in retroviruses and RT proteins in other retroid elements, especially in cases where it is not feasible to adjust individual amino acid substitution probabilities.

Acknowledgments. We thank Sarah Ingalls for programming advice and help with algorithms. We would also like to gratefully acknowledge Shirley Lee for her assistance on the training set sequence alignments, Toru Nakamura for the RT alignment, Ziheng Yang, Jianzhi G. Zhang, and Simon Whelan for helpful discussions, and Todd Raeker and Mike Kitson for computer support. This research was supported with funding from the NIH (grants GM08270 and T32-HG00040 to MWD and JSR), the NSF (grants DEB-9726427 and DBI-9974525 to DPM and RAG), and the Horace H. Rackham School of Graduate Studies.

References

Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42:459–468

Adachi J, Waddell PJ, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50:348–358

Akaike H (1974) New look at statistical-model identification. IEEE Trans Automat Control AC 19:716–723

Barker WC, Garavelli J, Huang H, et al. (2000) The protein information resource (PIR). Nucleic Acids Res 28:41–44

Beer B, Bailes E, Sharp P, Hirsch V (1999) Diversity and evolution of primate lentiviruses. Hum Retrovir AIDS 1999:460–474

Broussard SR, Comuzzie AG, Leighton KL, Leland MM, Whitehead EM, Allan JS (1997) Characterization of new simian foamy viruses from African nonhuman primates. Virology 237:349–359

Dimmic MW, Mindell DP, Goldstein RA (2000) Modeling evolution at the protein level using an adjustable amino acid fitness model. Pac Symp Biocomp 2000:18–29

Domingo E, Holland JJ (1997) RNA virus mutations and fitness for survival. Annu Rev Microbiol 51:151–178

Doolittle RF, Feng DF, McClure MA, Johnson MS (1990) Retrovirus phylogeny and evolution. Curr Top Microbiol Immunol 157:1–18

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

Goldman N (1998) Effects of sequence alignment procedures on estimates of phytogeny. BioEssays 20:287–290

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. Syst Biol 49:652–670

Hahn B, Shaw G, DeCock K, Sharp P (2000) AIDS as a zoonosis: scientific and public health implications. Science 287:607–614

Hasegawa M, Kishino H, Hayasaka K, Horai S (1990) Mitochondrial-DNA evolution in primates-transition rate has been extremely low in the lemur. J Mol Evol 31:113–121

Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M (1998) Retroviral diversity and distribution in vertebrates. J Virol 72:5955–5966

Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Syst 28:437–466

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trend Biochem Sci 23:403–405

Jones D, Taylor W, Thornton J (1992) The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282

Lawrence CT, Zhou JL, Tits AL (1997) User's guide for CFSQP version 2.5: a C code for solving (large scale) constrained nonlinear (minimax) optimization problems, generating iterates satisfying all inequality constraints. Institute for Systems Research, Technical Report TR-94-16rl, College Park, MD

Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16:793–805

Martin J, Herniou E, Cook J, O'Neill RW, Tristem M (1999) Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. J Virol 73:2442–2449

McClure MA (1999) The retroid agents: disease, function, and evolution. In: Domingo E, Webster R, Holland J (eds) Origin and evolution of viruses. Academic Press, London, pp 163–195

Mindell DP, Shultz JW, Ewald PW (1995) The AIDS pandemic is new, but is HIV new? Syst Biol 44:77–92

Morrison DA, Ellis JT (1997) Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol Biol Evol 14:428–441

Nakamura T, Morin G, Chapman K, Weinrich S, Andrews W, Lingner J, Harley C, Cech T (1997) Telomerase catalytic subunit homologs from fission yeast and human. Science 277:955–959

van Regenmortel MHV, Fauquet CM, Bishop DHL, et al. (2000) Virus taxonomy, VIIth report of the ICTV. Academic Press, San Diego, pp 369–387

States DJ, Harris NL, Hunter L (1993) Computationally efficient cluster representation in molecular sequence megaclassification. ISMB 1:387–394

Soyer O, Dimmic MW, Neubig RR, Goldstein RA (2002) Using evolutionary methods to study G-protein coupled receptors. Pac Symp Biocomp, pp 625–636

Swofford D (2000) PAUP* phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA

Thorne JL (2000) Models of protein sequence evolution and their applications. Curr Opin Genet Devel 10:602–605

Tristem M, Herniou E, Summers K, Cook J (1996) Three retroviral sequences in amphibians are distinct from those in mammals and birds. J Virol 70:4864–4870

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families

using a maximum-likelihood approach. Mol Biol Evol 18: 691–699

Xiong Y, Eickbush T (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z, Goldman N, Friday A (1995) Maximum likelihood trees from DNA-sequences: a peculiar statistical estimation problem. Syst Biol 44:384–399

Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600–l611

Yin H, Medstrand P, Krisofferson A, Dietrich U, Aman P, Blomberg J (1999) Characterization of human MMTV-like (HML) elements to a sequence that was highly expressed in a human breast cancer: Further definition of the HML-6 group. Virology 256:22–35

Zhang J (1999) Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol Biol Evol 16:868–875