Supplementary Material

Consequences of substitution model selection on protein ancestral sequence reconstruction

The supplementary material includes Figures S1-S5 and Tables S1-S2.

Figure S1. Agglomerative clustering of empirical substitution models of protein evolution traditionally used in phylogenetics. The figure shows the clustering of empirical substitution models of protein evolution that are traditionally applied in phylogenetics based on the distance between the normalized amino acid frequencies and exchangeability matrices.

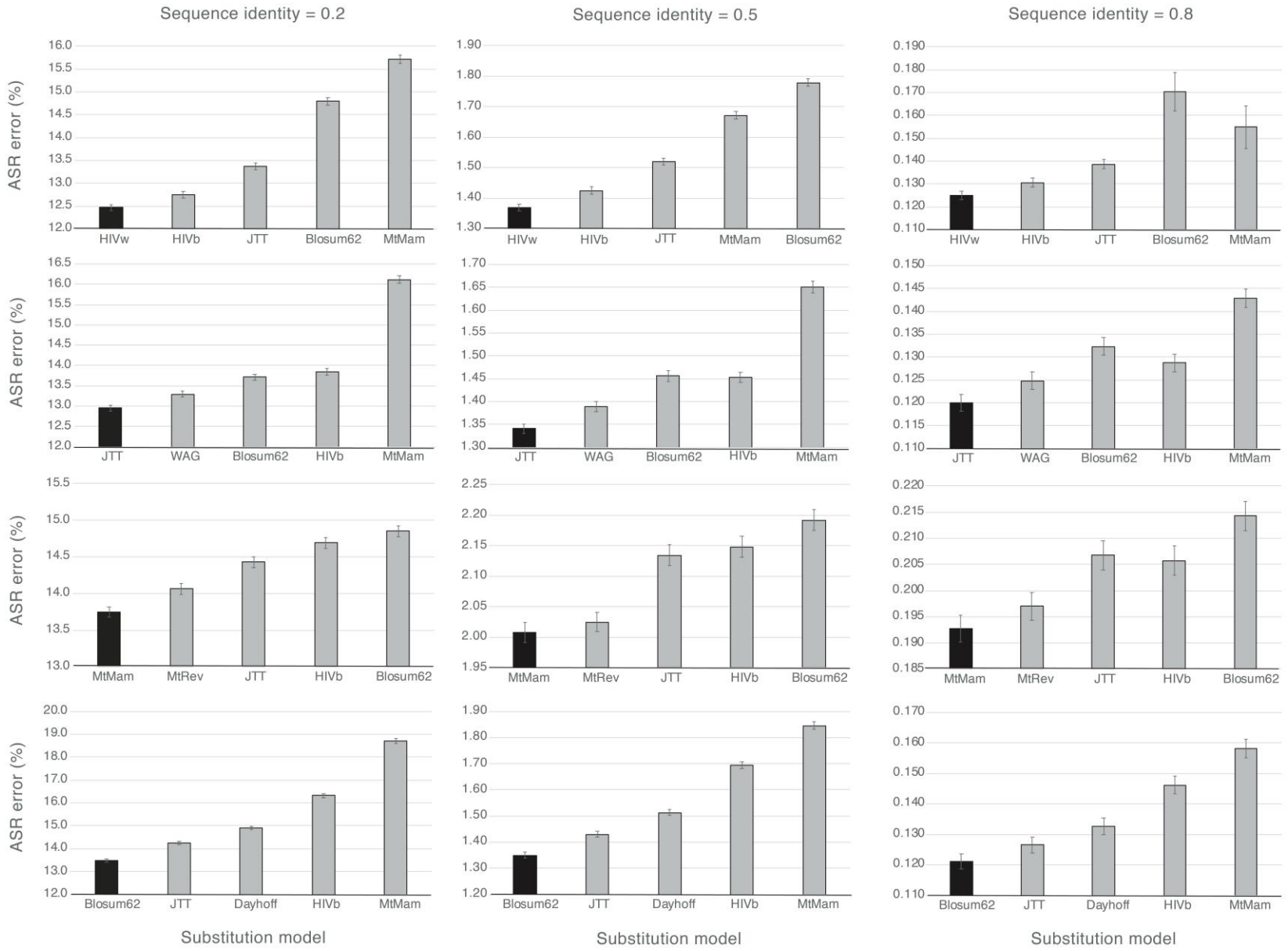
Sequence identity = 0.2
Sequence identity = 0.5
Sequence identity = 0.8
ASR error (%)
HIVw HIVb JTT Blosum62 MtMam
JTT WAG Blosum62 HIVb MtMam
MtMam MtRev JTT HIVb Blosum62
Blosum62 JTT Dayhoff HIVb MtMam
Substitution model

Figure S2. Influence of substitution model selection on ancestral sequence reconstruction using data simulated with 100 protein sequences. Distances between true ancestral sequences and ancestral sequences reconstructed under true (black bars) and other substitution models (grey bars; including from the left to the right a model that is similar, intermediate and far from the true model). The distances are shown in percentage. The study is based on 1000 simulated datasets of 100 protein sequences with sequence identity of 0.2 (large genetic diversity; plots on the left), 0.5 (intermediate genetic diversity; middle plots) and 0.8 (low genetic diversity; plots on the right). Error bars indicate 95% confidence intervals. The same results showing ASR error (Y-axis) from zero is presented in the Figure S4 (Supplementary Material).

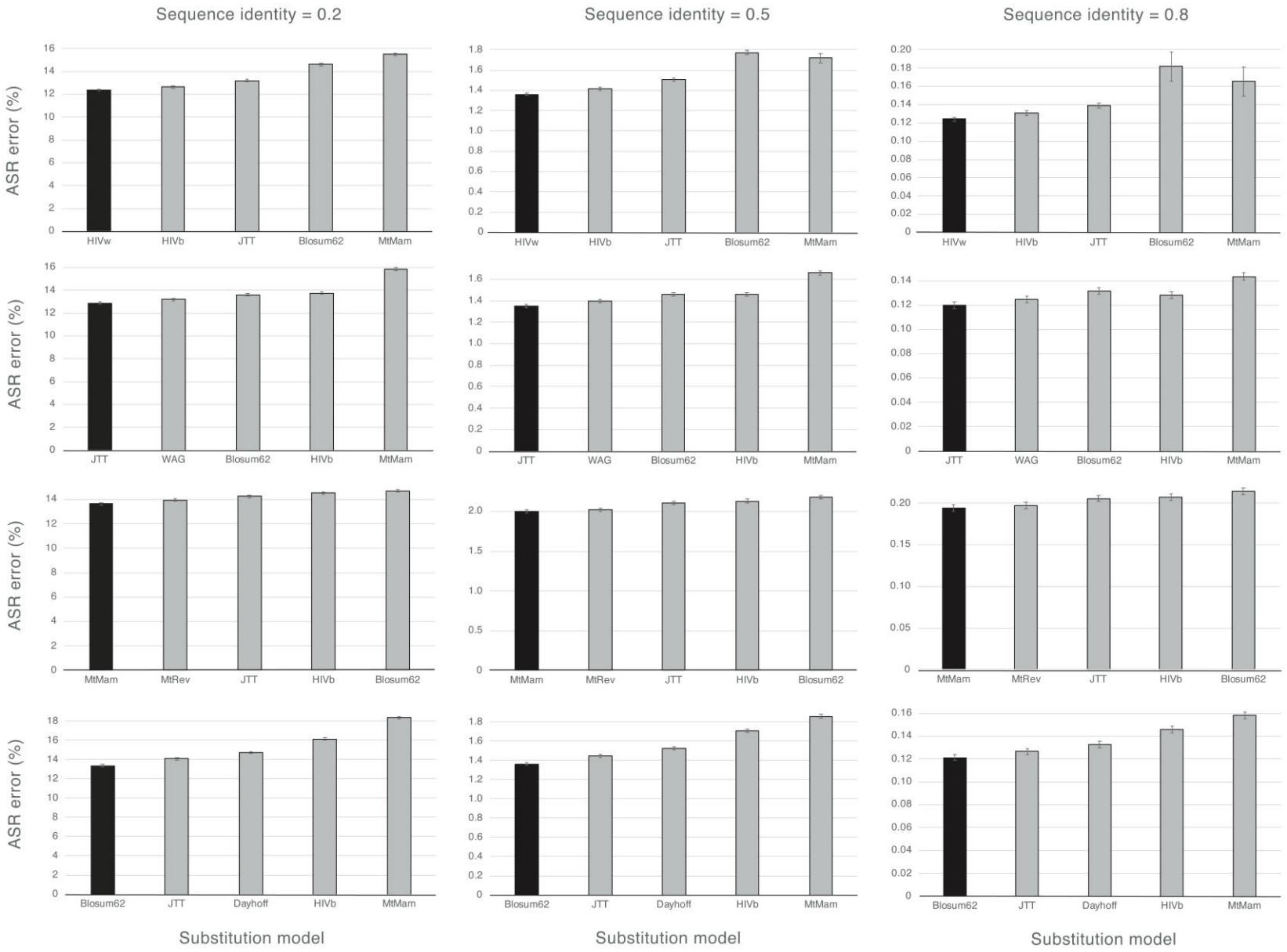
Sequence identity = 0.2
Sequence identity = 0.5
Sequence identity = 0.8
ASR error (%)
HIVw
HIVb
JTT
Blosum62
MtMam
WAG
MtRev
Dayhoff
Substitution model

Figure S3. Influence of substitution model selection on ancestral sequence reconstruction using simulated data showing ASR error (Y-axis) from zero. Distances between true ancestral sequences and ancestral sequences reconstructed under true (black bars) and other substitution models (grey bars; including from the left to the right a model that is similar, intermediate and far from the true model). The distances are shown in percentage. The study is based on 1000 simulated datasets of 50 protein sequences with sequence identity 0.2 (large genetic diversity; plots on the left), 0.5 (intermediate genetic diversity; middle plots) and 0.8 (low genetic diversity; plots on the right). Error bars indicate 95% confidence intervals. The statistical differences among the ASR error caused by the studied substitution models are better visualized in Figure 1.

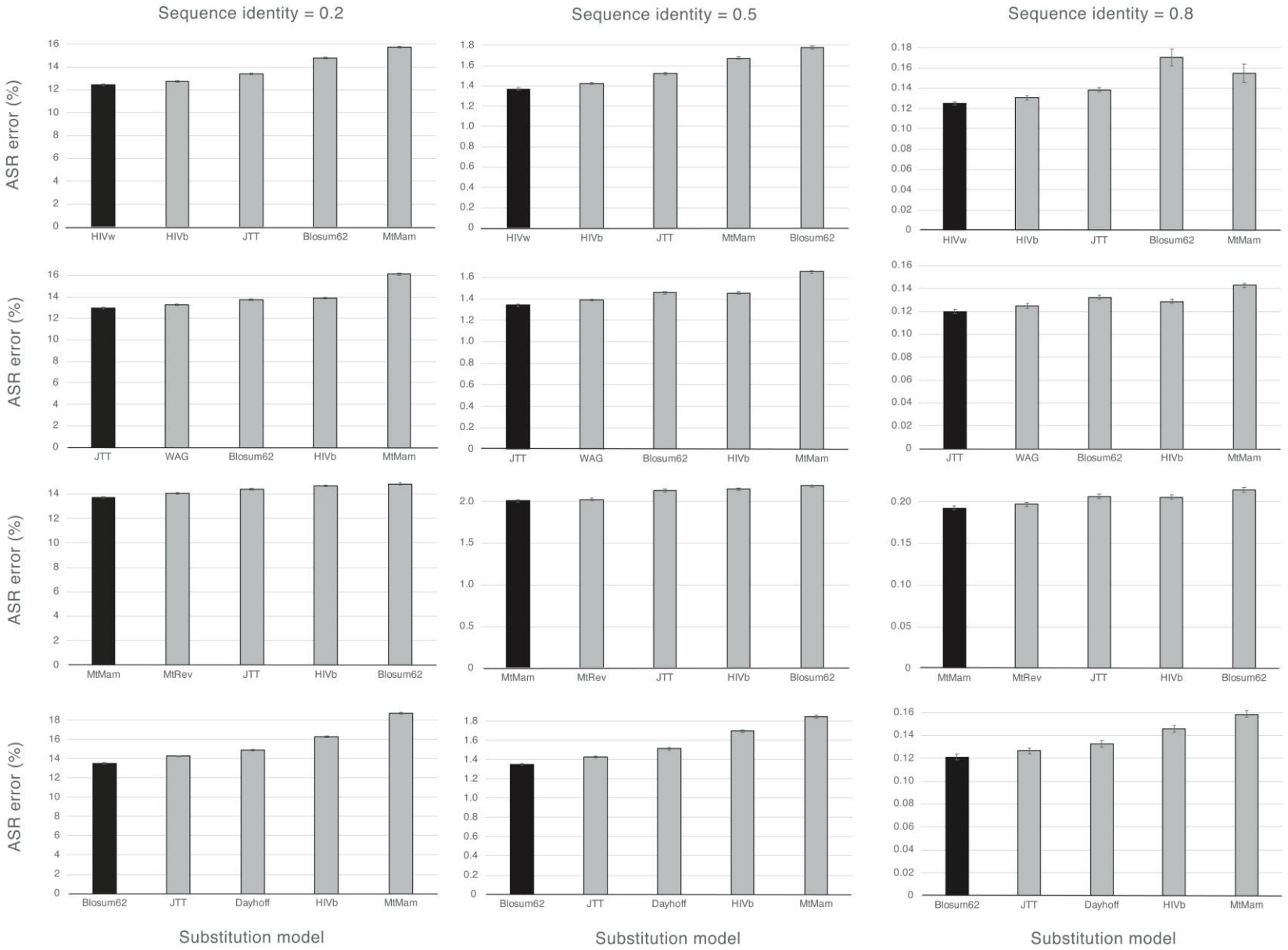
Sequence identity = 0.2
Sequence identity = 0.5
Sequence identity = 0.8
ASR error (%)
HIVw
HIVb
JTT
Blosum62
MtMam
WAG
MtRev
Dayhoff
Substitution model

Figure S4. Influence of substitution model selection on ancestral sequence reconstruction using data simulated with 100 protein sequences including ASR error (Y-axis) from zero. Distances between true ancestral sequences and ancestral sequences reconstructed under true (black bars) and other substitution models (grey bars; including from the left to the right a model that is similar, intermediate and far from the true model). The distances are shown in percentage. The study is based on 1000 simulated datasets of 100 protein sequences with sequence identity of 0.2 (large genetic diversity; plots on the left), 0.5 (intermediate genetic diversity; middle plots) and 0.8 (low genetic diversity; plots on the right). Error bars indicate 95% confidence intervals. The statistical differences among the ASR error caused by the studied substitution models are better visualized in Figure S2.

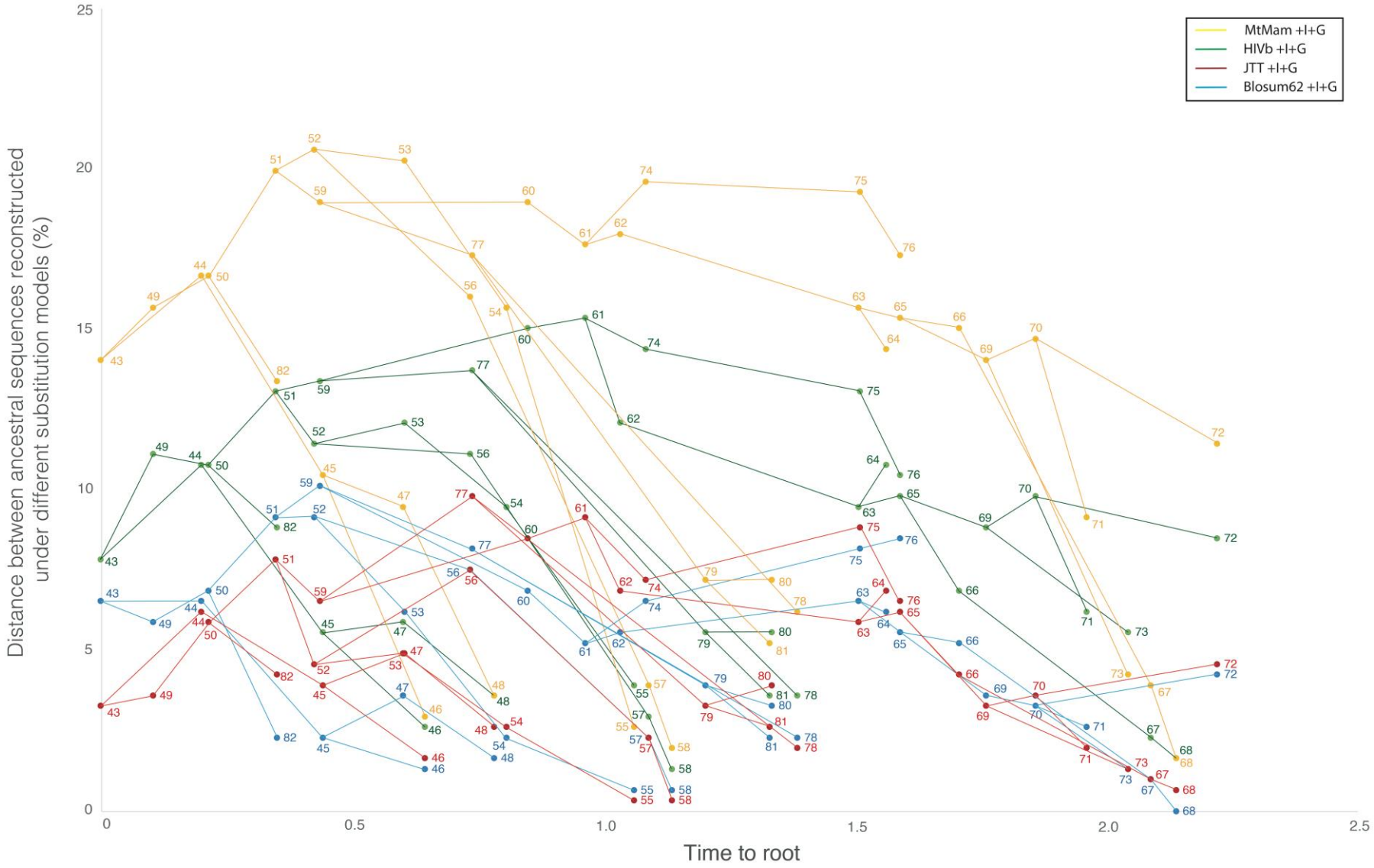
MtMam +I+G
HIVb +I+G
JTT +I+G
Blosum62 +I+G
Distance between ancestral sequences reconstructed under different substitution models (%)
Time to root
0
5
10
15
20
25
0
0.5
1.0
1.5
2.0
2.5

Figure S5. Influence of substitution model selection on ancestral sequence reconstruction of the DDL protein family. The figure shows the distance between ancestral sequences reconstructed under the best-fitting substitution model (LG +I+G) and other substitution models (MtMam +I+G, HIVb +I+G, JTT +I+G and Blosum62 +I+G; shown with different colors) at every internal node and as a function of the time to root. Note that all the nodes shown in the figure are internal nodes, the tip nodes are excluded because their sequences are given (thus, they are not reconstructed).

Table S1. Inferred CTL epitopes from ancestral sequences reconstructed under different substitution models for the HIV-1 group M reference alignment of the HIV-1 *env* region. Numbers are the CTL epitopes estimated by *MHCPred* for all available alleles in March of 2022. The cut-off value for the IC50 was 50, which only returns CTLs with high affinity. The best-fitting substitution model for the dataset was HIVb +I+G (numbers of epitopes in bold) and results from ancestral sequences reconstructed under other substitution models are included. Below, sum of absolute differences (per HLA allele) of number of epitopes in ancestral sequences reconstructed under the best-fitting substitution model and under other studied substitution models. Note that the substitution model with the exchangeability matrix more similar (HIVw +I+G) to the exchangeability matrix of the best-fitting substitution model (HIVb +I+G) also shows the number of epitopes more similar to the number of epitopes of the best-fitting substitution model.

HLA Allele	HIVb +I+G	HIVw +I+G	JTT +I+G	WAG +I+G	Blosum62 +I+G	MtMam +I+G
A0201	**13**	13	7	7	7	7
H2Db	**5**	5	6	6	7	5
H2Kb	**97**	95	107	108	104	106
H2Kk	**47**	47	50	51	52	45
A0101	**3**	4	5	8	7	4
A0202	**15**	15	26	25	23	27
A0203	**222**	224	225	228	223	227
A0206	**45**	43	44	48	46	47
A0301	**34**	43	32	30	27	31
A1101	**348**	351	374	359	359	369
A3101	**1**	1	2	2	2	2
A6801	**26**	26	25	24	25	25
A6802	**18**	18	13	15	12	14
B3501	**0**	0	0	0	0	0
DRB00101	**342**	339	326	329	327	324

DRB0401	**4**	4	5	9	7	6
DRB0701	**11**	11	16	13	13	18
IAb	**0**	0	0	0	0	0
IAk	**61**	58	49	50	49	49
IEg	**0**	0	0	0	0	0
IEk	**0**	0	0	0	0	0
IAd	**61**	59	56	55	58	58
IAs	**174**	177	189	189	183	183
IEd	**0**	0	0	1	1	0
TAP	**174**	175	171	180	168	167
All	***1701***	*1708*	*1728*	*1737*	*1700*	*1714*

	HIVb +I+G - HIVw +I+G	HIVb +I+G - JTT +I+G	HIVb +I+G - WAG +I+G	HIVb +I+G - Blosum62 +I+G	HIVb +I+G - MtMam +I+G
Differences	***31***	*129*	*126*	*111*	*125*

Table S2. Inferred CTL epitopes from ancestral sequences reconstructed under different substitution models for the HIV-1 subtype B reference alignment of the HIV-1 *env* region. Numbers are the CTL epitopes estimated by *MHCPred* for all available alleles in March of 2022. The cut-off value for the IC50 was 50, which only returns CTLs with high affinity. The best-fitting substitution model for the dataset was HIVw +I+G (numbers of epitopes in bold) and results from ancestral sequences reconstructed under other substitution models are included. Below, sum of absolute differences (per HLA allele) of number of epitopes in ancestral sequences reconstructed under the best-fitting substitution model and under other studied substitution models. Note that the substitution model with the exchangeability matrix more similar (HIVb +I+G) to the exchangeability matrix of the best-fitting substitution model (HIVw +I+G) also shows the number of epitopes more similar to the number of epitopes of the best-fitting substitution model.

HLA Allele	HIVw +I+G	HIVb +I+G	JTT +I+G	WAG +I+G	Blosum62 +I+G	MtMam +I+G
A0201	**7**	7	13	9	9	13
H2Db	**4**	3	3	4	4	4
H2Kb	**113**	114	104	108	107	105
H2Kk	**59**	60	63	63	63	62
A0101	**5**	5	4	4	3	4
A0202	**21**	21	21	22	20	22
A0203	**239**	240	234	229	230	235
A0206	**39**	39	32	34	33	33
A0301	**28**	28	26	26	24	27
A1101	**383**	377	391	394	392	397
A3101	**3**	3	1	2	3	1
A6801	**25**	24	24	25	26	26
A6802	**19**	19	26	26	25	26
B3501	**0**	0	0	0	0	0
DRB00101	**326**	321	314	304	304	311

DRB0401	**3**	3	2	2	2	2
DRB0701	**16**	16	14	13	14	13
IAb	**0**	0	0	0	0	0
IAk	**54**	54	64	66	67	63
IEg	**0**	0	0	0	0	0
IEk	**0**	0	0	0	0	0
IAd	**59**	57	53	55	55	53
IAs	**202**	201	193	188	189	192
IEd	**0**	0	0	0	0	0
TAP	**134**	131	136	134	137	136
All	***1739***	*1723*	*1718*	*1708*	*1707*	*1725*

	HIVw +I+G - HIVb +I+G	HIVw +I+G - JTT +I+G	HIVw +I+G - WAG +I+G	HIVw +I+G - Blosum62 +I+G	HIVw +I+G - MtMam +I+G
Differences	***22***	*95*	*105*	*108*	*100*