Supplementary material

Properties of the artificial data generated with MSAvolve:

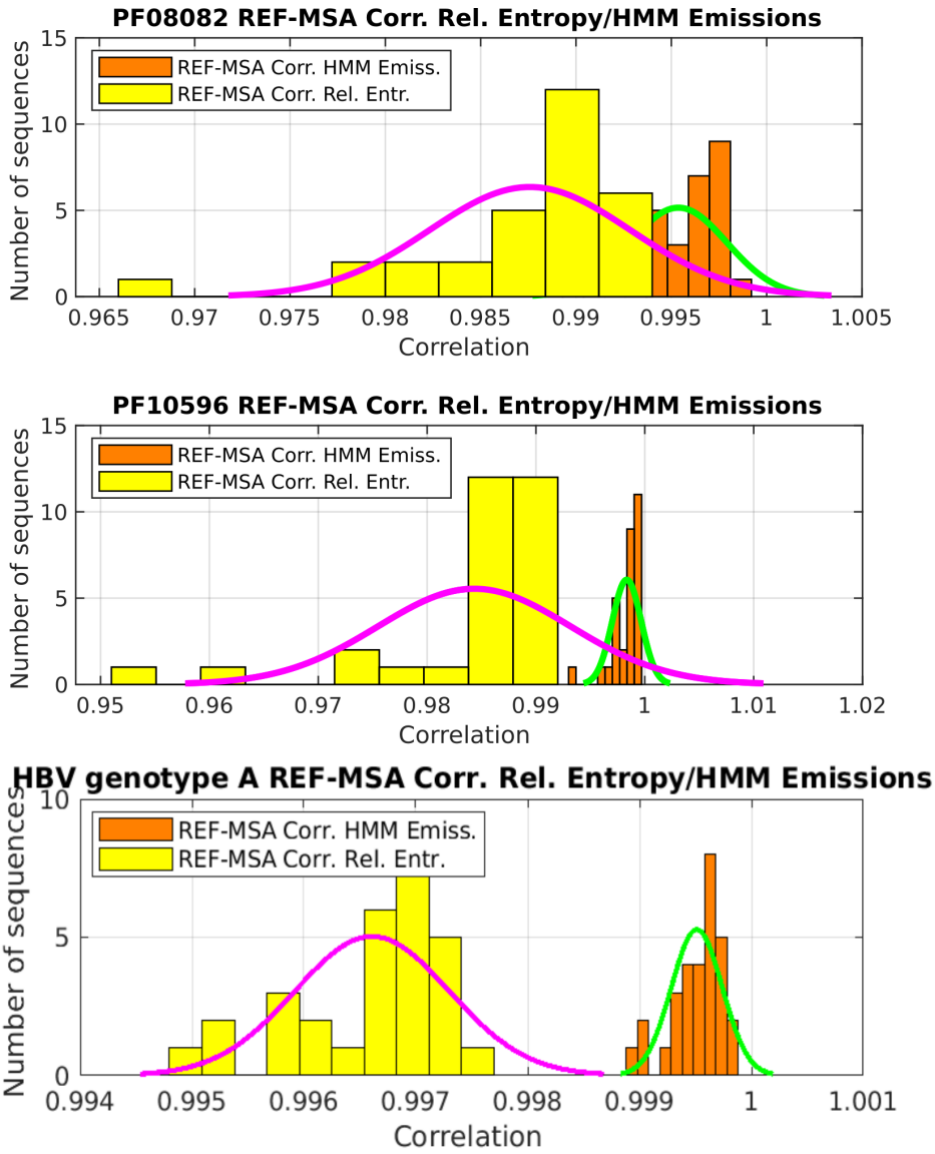

Figure S1: **Correlations between real data and simulated data.**
The distribution of mean correlations computed between the HMM emissions calculated from each MSA of simulated sequences and the HMM emissions calculated from the Reference MSA (orange), and the correlations computed between the relative entropy of each MSA and the relative entropy of the Reference MSA (yellow). Reference MSAs refer to MSAs computed from the sets of sequences named “FULL” downloaded from Pfam for PF08082, PF10596, and from the HBV sequences of genotype A (www.lcqb.upmc.fr/BIS2TreeAnalyzer/).

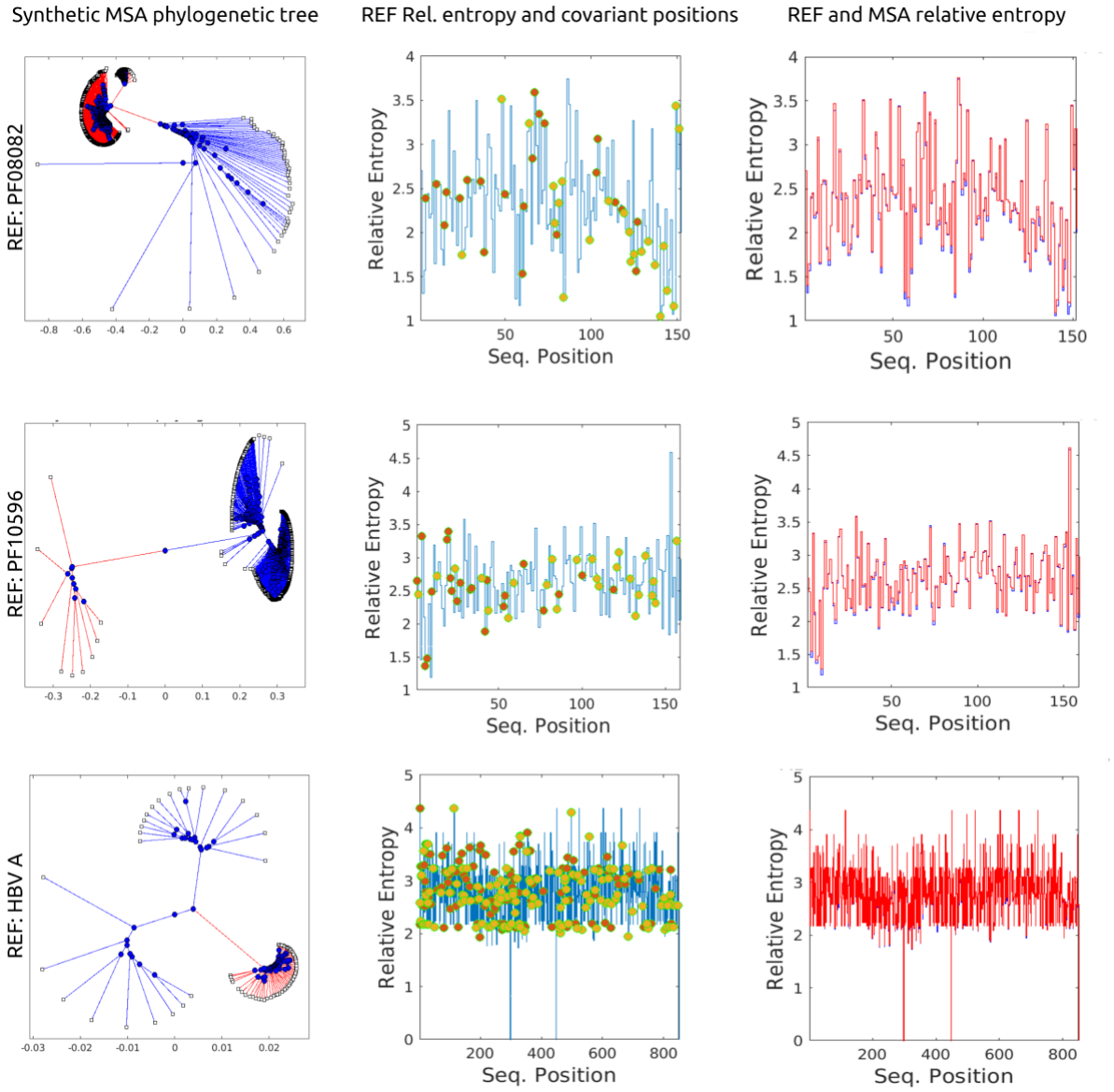

Figure S2: **Statistical features of the three groups of simulated MSAs.**
Each row describes features for one (randomly selected) of the synthetic MSAs among the 30 alignments of 500 sequences generated for the PF08082 family (top), the PF10596 family (middle), and the HBV Polymerase of genotype A (bottom). The **first** panel shows the phylogenetic tree of sequences for the randomly selected synthetic MSA. The **second** panel shows the relative entropy at each position of the real MSA (in blue). Given the group of pairs of positions set to coevolve in the simulated MSA, we represented the positions in the pairs by circles where one of the positions is colored orange and the other yellow; the plot shows that pairs of coevolving positions are distributed in the MSA with different entropy values. The group of pairs of coevolving positions, randomly selected in each simulation, covers the 15% of the full set of positions. The **third** panel shows that the relative entropy of the simulated MSA (red line) superimposes almost perfectly to that of the real MSA (blue line).

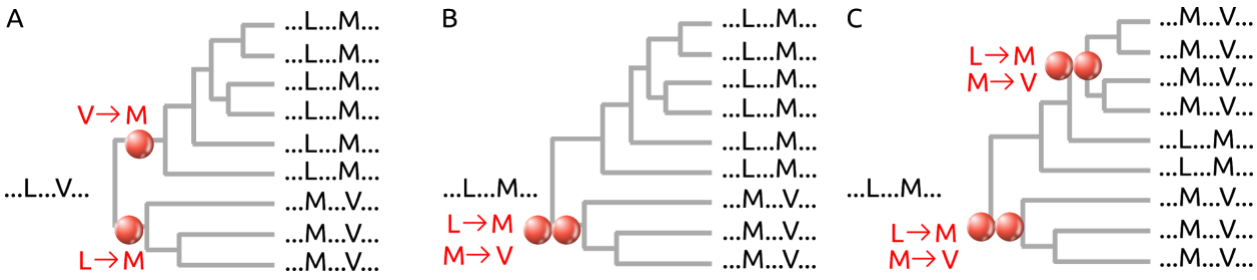

Figure S3: **Different evolutionary scenarios for a covariation signal between pairs of positions.**

A: The phylogenetic tree induces the identification of covariation between positions in an alignment when two independent mutational events (red circles) take place in different branches of the tree. We consider this case as “phylogenetic signal”. B: The covariation between positions results from a double substitution in a branch of the tree. C: The covariation between positions results from two (or more) double substitutions in different branches of the tree. This provides an increasing evidence of coevolution.

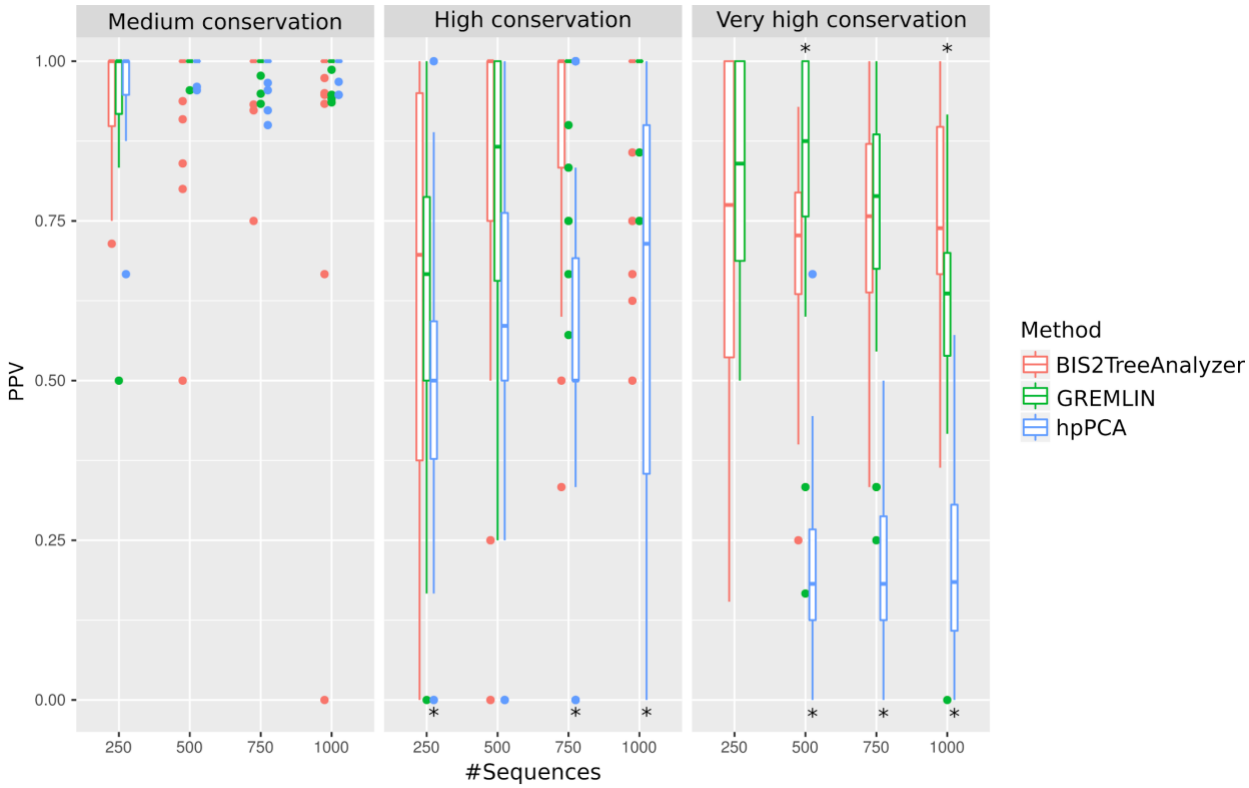

Figure S4: PPV distribution for BIS2TreeAnalyzer, GREMLIN and hpPCA.
PPV were computed on 360 MSAs for the 12 conditions reported in the columns of the plot (30 MSAs are considered for each condition). Distributions were compared using Mann–Whitney U test. Statistically significant differences (P-value < 0.05) between BIS2TreeAnalyzer and GREMLIN are labelled "*" at the top of the boxplot (for very high conservation, on 500 sequences, in favor of GREMLIN and on 1000 sequences in favor of BIS2TreeAnalyzer), and differences between BIS2TreeAnalyzer and hpPCA are labelled "*" at the bottom (all in favor of BIS2TreeAnalyzer). The median and median absolute deviation values are reported in Table S1.

	Medium conservation (PF08082)				High conservation (PF10596)				Very high conservation (HBV Polymerase of genotype A)			
Alignment size	250	500	750	1000	250	500	750	1000	250	500	750	1000
BIS2TreeAnalyzer	1±0	1±0	1±0	1±0	0.697 ± 0.449	1 ± 0	1 ± 0	1 ± 0	0.775 ± 0.334	0.727 ± 0.130	0.757 ± 0.177	0.739 ± 0.189
GREMLIN	1±0	1±0	1±0	1±0	0.667 ± 0.222	0.866 ± 0.199	1 ± 0	1 ± 0	0.840 ± 0.238	0.875 ± 0.185	0.789 ± 0.165	0.636 ± 0.120
hpPCA	1±0	1±0	1±0	1±0	0.500 ± 0.178	0.586 ± 0.244	0.500 ± 0.247	0.714 ± 0.424	NA	0.182 ± 0.103	0.182 ± 0.111	0.185 ± 0.135

Table S1: **PPV comparison of BIS2TreeAnalyzer, GREMLIN and hpPCA.**
We report the median and the median absolute deviation of PPV obtained by BIS2TreeAnalyzer, GREMLIN and hpPCA in simulated alignments. Rows represent the coevolution method, and columns are organized with respect to the number of sequences in the simulated MSAs generated for three distinguished seed sequences presenting different levels of conservation. Values were computed over 30 MSAs, whose properties are described in Figures S1 and S2. A method failing in the calculation is indicated as NA (not available). Compare with Fig. S4.