Supplementary Information for Hurst et al.: Causes of trends of amino acid gain and loss

Methods

Identification of orthologues, alignment and evolutionary distances

A preliminary set of orthologues was defined by identifying unique pairwise reciprocal best hits, with at least 40% similarity in protein sequence and less than 20% difference in length. This list was then refined by combining the information on the distribution of similarity of these putative orthologs and the data on gene order conservation [1]. The protein sequences were then aligned with ClustalW[2] and then back-translated into DNA. For the multiple alignments of intergenic regions, we selected all large (>50 bp) intergenic regions separating pairs of genes such that the genes and their orthologues are consecutive in all genomes. Only intergenic regions of similar size (<20% difference in length) in all genomes were used for the multiple alignments.

Derivation of nucleotide mutational profiles from intergenic DNA.

The aligned orthologs of each genome were concatenated and a neighbour-joining tree reconstructed for each of the three taxa using MEGA [3] (supplementary Fig. S5 – below)). The trees informed the choice of a single focal lineage in each taxon from which to calculate the nucleotide mutation profile; too little variation results in noisy data and too much variation means that the profile is more likely to have been affected by selection. The mutational profiles of these lineages were then determined by comparing each of these genomes with the other genomes available for the same taxon (total number of genomes used: *Staphylococcus*, n = 6 *Bacillus*, n = 5, *E. coli* / *Shigella*, n= 5). To

maximise confidence in the direction of changes, only unique polymorphisms in the focal lineages were scored in cases where the corresponding site in the other genomes was monomorphic (thus a T->C change was recorded in a given genome when a C was noted at position X in this genome, whilst a T was noted in position X in all other genomes). The number of each substitution type was then divided by the frequency of the ancestral nucleotide (i.e. the total number of C / length of the sequence). These frequencies were normalised to give the mutational profile.

Detection of amino-acid changes

The concatenated orthologues were translated and amino-acid changes in focal lineages were scored in the same manner as the nucleotide changes in intergenic regions (i.e. a change was only accepted if it is monomorphic in all other lineages). In order to examine the effect of divergence time on the amino-acid substitutions we repeated the analysis on the other genomes available for each taxon (including more diverse enterobacterial sequences from *Salmonella*, *Yersinia* and *Erwinia*) (see Fig 2 in manuscript and Fi S2 below). We also increased the timepoints available by the selective exclusion of sequences; given the tree (((A, B), C), D) the exclusion of genome B effectively increases the divergence time in A. The normalised gain/loss ratio, (g-l)/(g+l), was then calculated for each amino acid for each lineage.

Simulation of mutational gain/loss pattern.

From the concatenated file of all orthologues in each focal lineage we derived the total number of each of the 61 codons**.** We randomly picked a codon according to its

frequency in the genome, and randomly picked a base in the codon. We then randomly mutate this base according to the mutational profile frequencies. Stop codons are rejected, but synonymous changes are not. We re-iterated until the total number of non-synonymous changes reached the observed number for the focal lineage. Amino-acids changes were scored as in a real alignment; in the cases of A->B->A or A->B->C the frequency of B is unaffected. This simulation process was repeated 1000 times and the mean (g-l)/(g+l) per simulation for each amino acid determined. The simulations make no account of dinucleotide mutational biases or selection to avoid certain dinucleotides, such as TA.

Calculation of mutational equilibrium amino acid usage.

For each focal lineage we obtained the relative rates of each of the twelve possible point mutations, after allowing for different abundance of the four nucleotides. We could then define the rate of gain and loss of a given nucleotide. For example, for nucleotide A:

Gain of A = freq(T)xmu(T->A) + freq(C)xmu(C->A) +freq(G)xmu(G->A)

Loss of A = freq(A)(mu(A->T) + mu(A->C) + mu(A->G))

Where freq(N) is the frequency of nucleotide N and mu(N->M) is the relative probability of mutation from N to M. We can then solve simultaneously for all nucleotides for the position where gain of a nucleotide is equal to loss. This provides a numerical solution for the equilibrium frequency of all the four bases, for any given set of mutational profiles. Equilibrium codon frequencies were defined as the product of the three relevant mutational equilibrium frequencies. The frequencies were then normalised by dividing by 1- sum of the equilibrium frequencies of the stop codons. For each amino acid we then

considered both the sum of the deviation between observed and expected frequencies and mean per codon bias.

Further evidence

For the relationship between dS/dN alluded to in the text, see Fig S1. For evidence that parsimony and maximum likelihood recontruction provide qualitatively similar results see Fig S3. For SNP based evidence that even in humans rarer amino acids are being purged by purifying selection.

References

1. Rocha, E. P. C. et al. Comparisons of dN/dS are time-dependent for closely related bacterial genomes. *J Theor Biol* **238**, doi:10.1016/j.jtbi.2005.08.037 (2005).
2. Thomson, J. D., Higgins, D. G. & Gibson, T. J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. **22**, 4673-4680 (1994).
3. Kumar, S., Tamura, K. & Nei, M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150-63 (2004).

Table S1 - Genomes used in this analysis. Optimal estimates of the mutation profile correspond to lower levels of divergence than the profile of amino-acid gain and loss, and hence slightly different genomes were used for each analysis in the Staphylococcal and enterobacterial comparisons. * Used only as comparitors to increase confidence in the directionality of the mutational profile (of STAUA and ESCOD). ** Used to calculate amino-acid loss and gain in ESCOD, and for examination of time dependence in the Enterobacteria, but not used to calculate intergenic mutational profile in ESCOD. The genomes in bold are the three focal lineages (STAUA, ESCOD and BATUA).

Species	Strain	id	Reference or URL	Accession
Staphylococcus aureus	MSSA-476	STAUE*	PNAS 101:9786-9791	BX571857
	MRSA-252	STAUD	PNAS 101:9786-9791	BX571856
	MW2	STAUC*	Lancet 359: 1819-1827	NC_003923
	Mu50	STAUB*	Lancet 357: 1225-1240	NC_002758
	N315	**STAUA**	Lancet 357: 1225-1240	BA000018
	COL	STAUF	J. Bact. 187(7):2426-38	CP000046
Escherichia coli	EAEC-42	ESCOE	http://www.sanger.ac.uk	
	CFT073	**ESCOD**	PNAS 99:17020-4.	CFT073v17o
	K-12 MG1655	ESCOA*	Science 277:1453-74	U00096
	O157:H7 sakai	ESCOB*	DNA Res. 8:11-22	BA000007
Shigella flexneri	2a301	SHFLA	NAR 30:4432-41	NC_004337
Yersinia pestis	CO92	YEPEA**	Nature 413:523-7	AL590842
Yersinia pseudotuberculosa	IP32953	YEPSA**	PNAS 101: 13826–31	NC_006155
Erwinia carotovora	SCRI1043	ERCAA**	PNAS 101:11105–10	BX950851
Salmonella enterica typhi	CT18	SATYB**	Nature 413:848-852	AL513382
Salmonella enterica typhimurium	LT2	SATYA**	Nature 413:852-6	AE006468
Bacillus anthracis	Sterne	BAAND	Science 296:2028-33	AE017225
Bacillus thuringiensis	9727	**BATUA**	Unpublished	AE017355
Bacillus cereus	ATCC10987	BACEB	NAR 32:977-88.	NC_003909
	ATCC14579	BACEA	Nature 423:87-91	NC_004722
	ZK	BACEC	Unpublished	CP000001

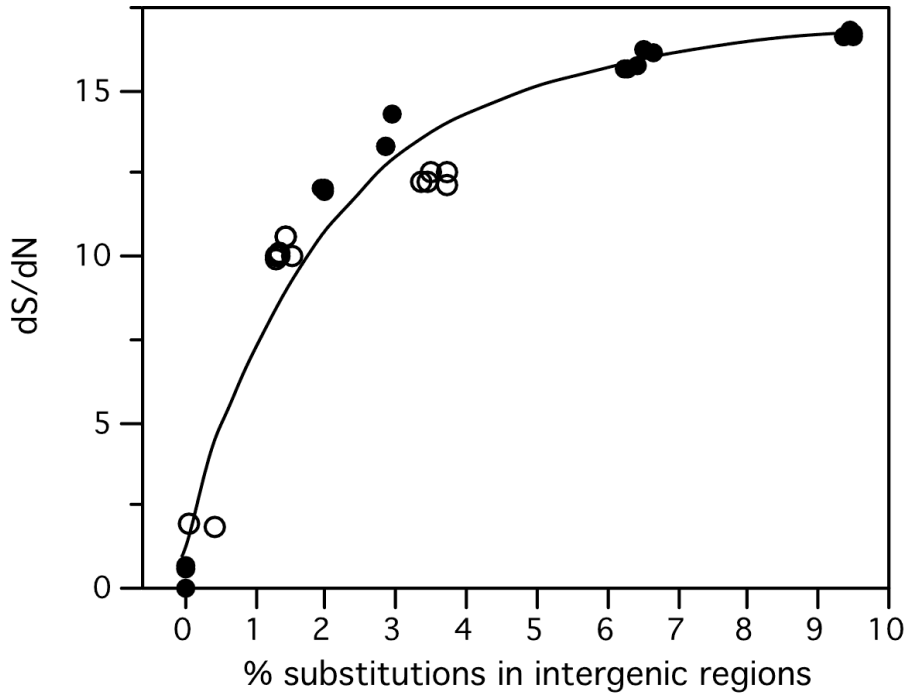


Figure S1. The dS/dN ratio between two taxa as a function of the divergence in their intergenic DNA (approximately a measure of evolutionary distance). Full circles: Bacillus data, open circles: Staphylococcus data. Distance between intergenic sequences was computed as the percentage of sequence identity. Adapted from [1].

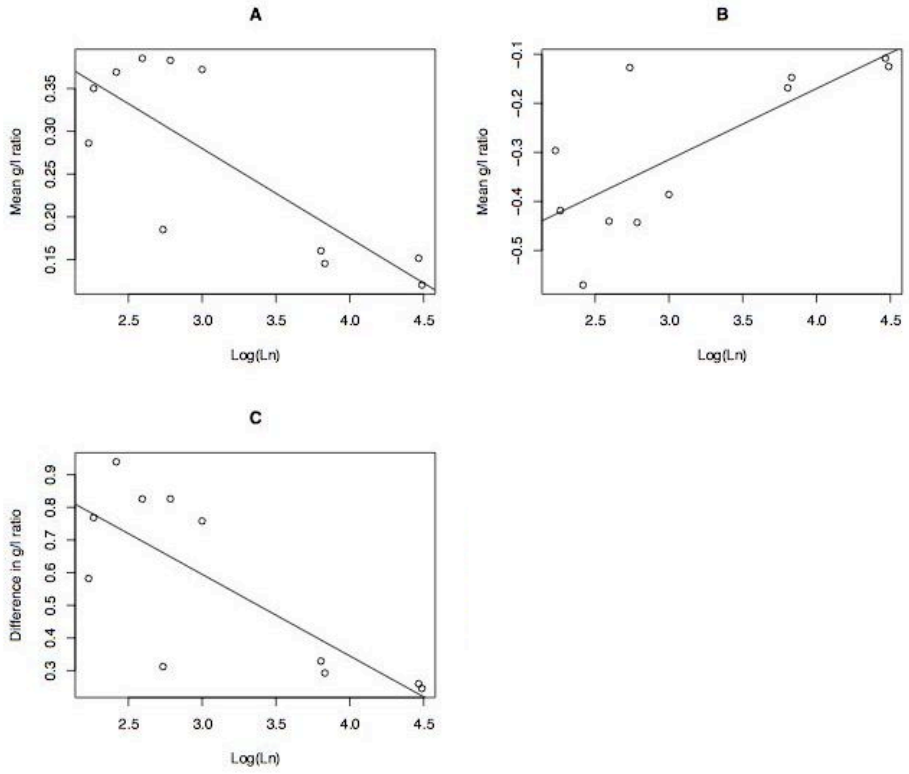


Figure S2 The observed mean normalised gain loss difference in the enterobacterial comparisons for the amino acids that are short term gainers (a) and short term losers (b) and the difference between mean gain and mean loss (c), as a function of the number of non-synonymous changes (Ln) in the relevant comparisons. Note that when the comparator species are more divergent there is a significant decrease in the net gain or loss. The relevant statistics are: a) Average gainers: $R^2 = 0.641$, $P = 0.003$; b) Average losers: $R^2 = 0.559$, $P = 0.008$; c) Difference $R^2 = 0.606$, $P = 0.005$. The combined data from other taxa are shown in Figure 2A.

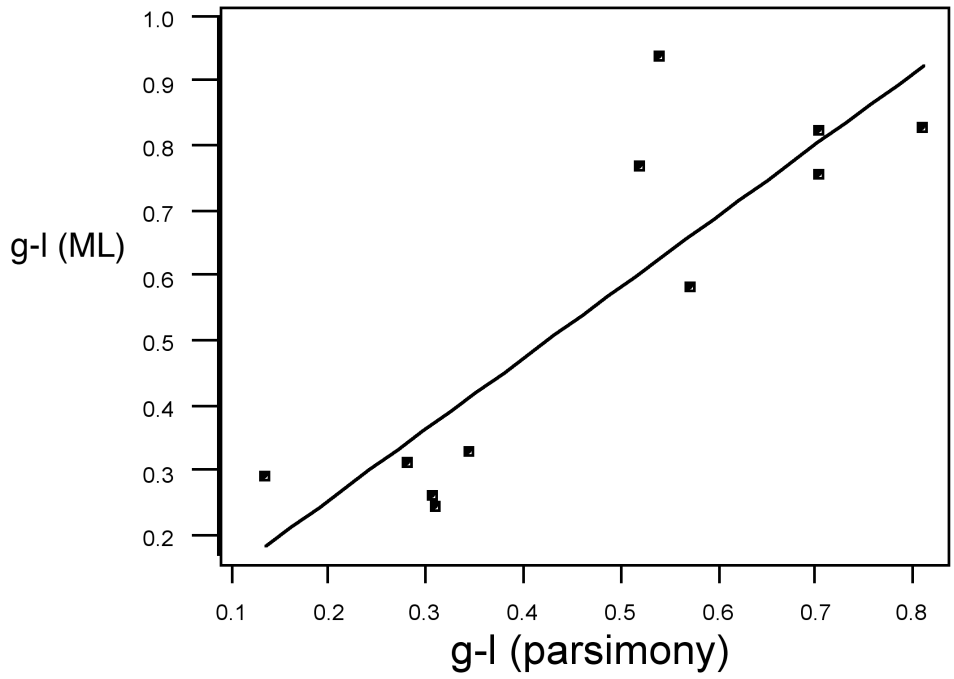


Figure S3. A comparison of the differences in bias between the main gainers and the main losers (g-l) as calculated using the parsimony method described in *methods* and a maximum-likelihood approach as implemented in PAML ($R^2 = 0.74$, P=0.001).

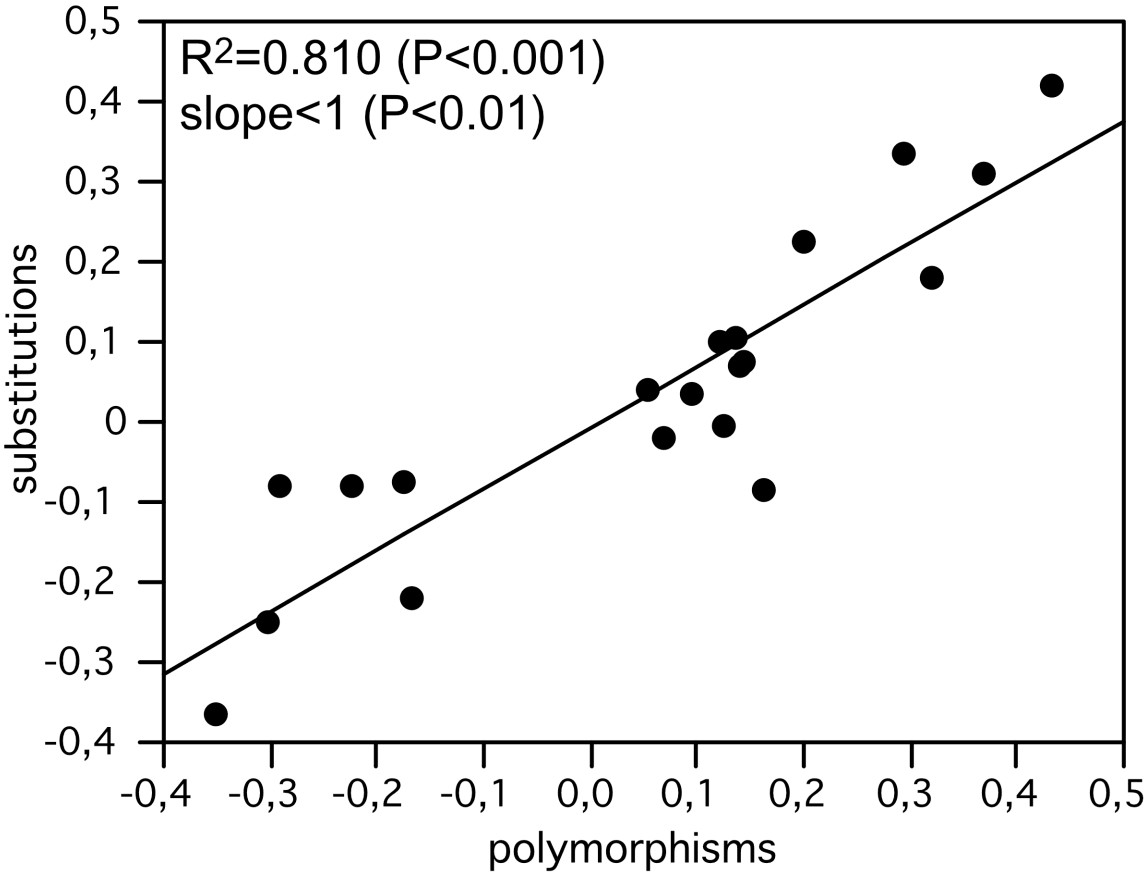
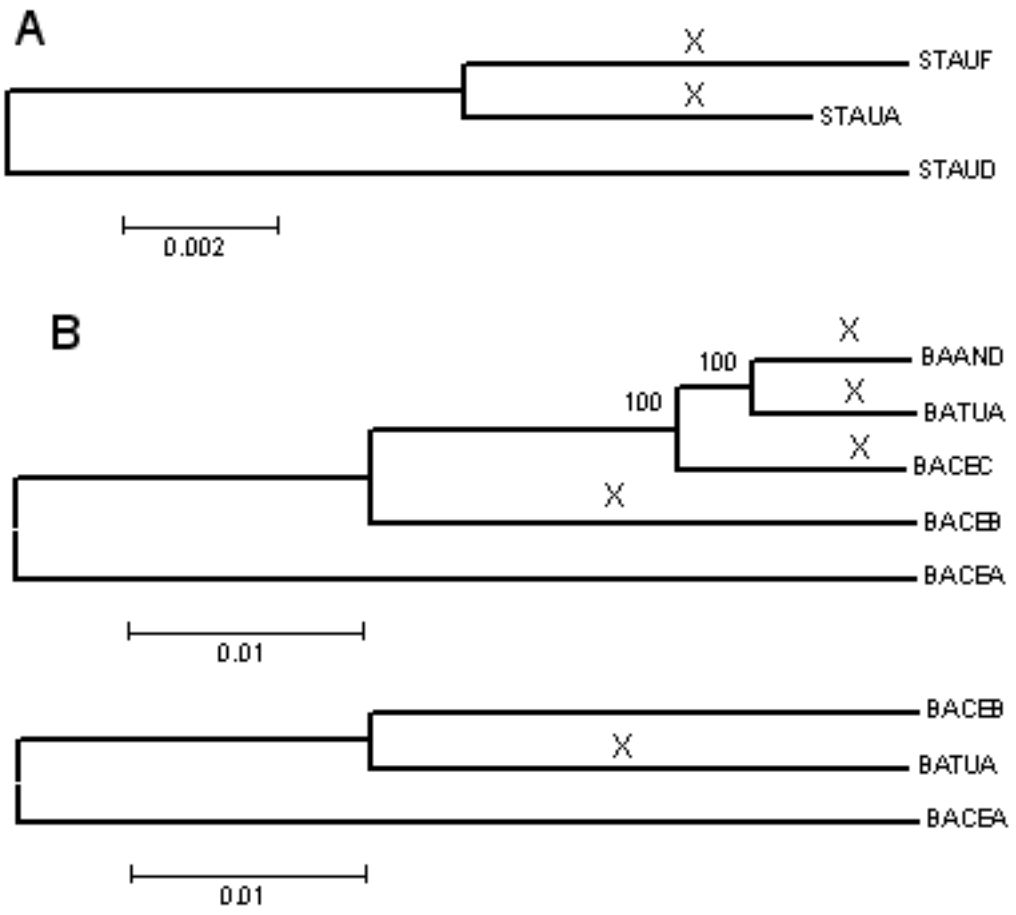


Figure S4 - Reanalysis of Table 2 in Jordan et al. We plot the normalized gain-loss of each type of amino acid in human polymorphisms (SNP) (X) and substitutions since human-chimpanzee divergence (Y). If normalized gain-loss was similar among both sets, then the slope should be 1, if amino acids were yet to achieve a compositional equilibrium the slope should be higher than 1, if rarer amino acids are being purged by purifying selection then the slope should be smaller than 1. The statistical tests reject all but the latter hypothesis (P<0.01, one tailed t-test).


A
STAUF
STAUA
STAUD
0.002
B
100
100
BAAND
BATUA
BACEC
BACEB
BACEA
0.01
BACEB
BATUA
BACEA
0.01

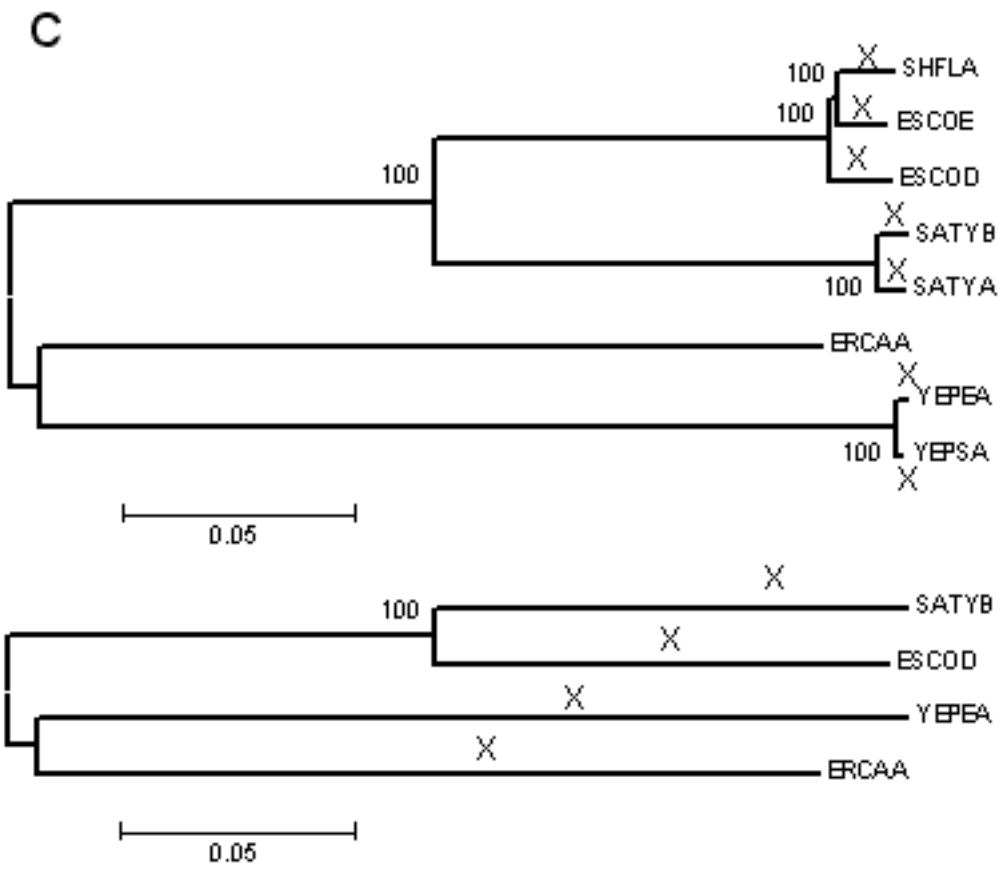


Figure S5 – Neighbour-Joining trees reconstructed from the genomes used to measure time-dependency of the amino-acid biases (A = *S. aureus* sequences, B = *Bacillus* sequences and C = Enterobacterial sequences). The unique amino-acid changes along each branch marked with an "X" were scored if the corresponding residue was monomorphic in all other lineages. The direction of each change was assigned on the principal of parsimony.