Supplementary material for:

Towards more accurate ancestral protein genotype–phenotype reconstructions with the use of species tree-aware gene trees.

Mathieu Groussin[1,a,‡,*], Joanne K Hobbs[2,b,*], Gergely J Szöllősi[1,3],
Simonetta Gribaldo[4], Vickery L. Arcus[2], and Manolo Gouy[1]

[1] – *Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France*

[2] – *Department of Biological Sciences, University of Waikato, Hamilton, New Zealand*

[3] – *ELTE-MTA "Lendület" Biophysics Research Group, Pázmány P. stny. 1A., H-1117 Budapest, Hungary*

[4] – *Institut Pasteur, Département de Microbiologie, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, 25-28 rue du Dr Roux, 75724 Paris cedex 15, France*

[a] – Present address: Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[b] – Present address: Department of Biochemistry and Microbiology, University of Victoria, British Columbia, Canada

[*] – These authors equally contributed to the work

[‡] – Corresponding author

Email: mgroussi@mit.edu

Supplementary Material

Homogeneous and site-heterogeneous models employed in this study

Site-homogeneous substitution models assume that the evolutionary process is constant among sites. When the Markovian process is time-reversible, the transition probability matrix is computed by multiplying the matrix of exchangeabilities with the diagonal matrix of equilibrium frequencies (Whelan and Goldman, 2001). Hereafter, the latter matrix is referred to as a profile of equilibrium frequencies. In the case of site-homogeneous models, both the exchangeabilities and the profile are constant among sites. The LG site-homogeneous model (Le and Gascuel, 2008) is one of the models employed in our study.

To model the heterogeneity of the process among sites, mixture models were considered. These approaches use sets of different models in which each model is assigned a particular weight. The likelihood of a given site is then the sum of all weighted likelihoods computed with each model of the mixture (Le et al., 2008b). The models of the mixture may have been learned to take into account protein properties that are heterogeneous along the sequence and that influence the substitution process, such as solvent exposure or secondary structure. In line with this, Le et al. (2008b) and Le and Gascuel (2010) proposed a series of empirical mixture models that outperform any site-homogeneous models. They learned their models on the HSSP database (Schneider et al., 1997) of aligned protein sequences in a supervised or unsupervised way. In the supervised way, sites were *a priori* assigned to a component of the mixture given knowledge about their localization in the protein, and the exhangeabilities and equilibrium frequencies of each model were subsequently learned from these sites. Le et al. (2008b) and Le and Gascuel (2010) inferred four models in this way:

- EX2, which is composed of two matrices corresponding to exposed/buried sites
- EX3, which is composed of three matrices corresponding to highly exposed/intermediate/buried sites
- EHO, which is composed of three matrices corresponding to extended/helix/other sites
- EX_EHO, which is composed of six matrices corresponding to the combination of EX2 and EHO.

In the unsupervised way, both site partitions and their corresponding matrices were directly learned from the data. Two models were proposed by Le et al. (2008b):

- UL2, which is composed of two matrices
- UL3, which is composed of three matrices

Note that all these models are mixtures of matrices with both exchangeabilities and equilibrium frequencies varying among components.

Mixture of profiles were also previously proposed (Le et al., 2008a). In these site-heterogeneous models, only the profiles vary among the components of the mixture, which share the same exchangeabilities. The components of these mixtures were learned in an unsupervised way. Six models were proposed, with 10, 20, 30, 40, 50 or 60 different profiles and named C10 to C60.

All these site-homogeneous and site-heterogeneous models were run on the LeuB data to evaluate their ability to efficiently fit the data (see Material & Methods in the main text and below)

Data fitting

To determine the best model in terms of data fitting to perform ASR on the LeuB data, we used the AIC criterion (Akaike, 1973). This criterion allows the evaluation of non-nested models by penalizing the number of parameters influencing the likelihood. The AIC criterion is computed as follows:

$$AIC = -2 \times lnL + 2 \times K,$$

with lnL the final likelihood and K the total number of parameters. For a site-homogeneous model, only the α parameter of the Γ distribution is involved, so that $K = 1$. For site-heterogeneous models, the sum of all component-specific weights equals 1, so that $K = (n-1)+1$, with n the number of components of the mixture model.

Differences between LeuB$_{S-unaw}$ and LeuB$_{S-aw}$ enzymes

LeuB$_{S-unaw}$ and LeuB$_{S-aw}$ differ by approximately 10% in terms of amino acid sequence. In an attempt to rationalise the biochemical differences between LeuB$_{S-unaw}$ and LeuB$_{S-aw}$, we modelled the structure of LeuB$_{S-aw}$ with SWISS-MODEL (Arnold et al., 2006) (Supplementary Figure 4) and mapped on this structure the amino acid differences with LeuB$_{S-unaw}$. The active site, as well as the co-factor binding site and residues involved in the interaction between LeuB monomers are unchanged between the two enzymes. The large majority of the amino acid differences are located at the surface of the protein. One of these differences is located near the active site, although it does not involve a change in physicochemical properties: this residue is VAL in LeuB$_{S-aw}$ and ALA in LeuB$_{S-unaw}$. In conclusion, deciphering the exact reason(s) of the biochemical differences between the two enzymes would require additional experiments that are beyond the scope of this manuscript.

Supplementary Table

Supplementary Table 1: **Patterns of Grantham scores for reconstruction errors obtained with the LG S-unaware trees.** Number of reconstruction errors were sorted increasingly along the Grantham score. All 190 possible pairwise substitutions are represented. To a given Grantham score may correspond several amino acid pairs. Categories of Grantham scores with a percentage of errors higher than 1% are in bold lines. The patterns of Grantham scores for the S-unaware trees reconstructed with the C60 model or for the S-aware trees were similar and are not represented here.

Grantham Score	AA substitution	Number of errors	% of errors
5	**L--I**	**3178**	**3.96**
10	M--I	668	0.83
15	**M--L**	**1612**	**2.01**
21	**F--I** **V--M**	**903**	**1.12**
22	**F--L** **Y--F**	**3791**	**4.72**
23	**D--N**	**1317**	**1.64**
24	H--Q	382	0.48
26	**K--R**	**2053**	**2.56**
27	**P--A**	**1191**	**1.48**
28	F--M	218	0.27
29	**E--Q** **H--R** **V--I**	**6011**	**7.49**
32	**K--H** **V--L**	**2576**	**3.21**
33	Y--I	179	0.22
36	Y--L Y--M	619	0.77
37	Y--W	486	0.61
38	T--P	314	0.39
40	H--E W--F	697	0.87
42	**E--N** **P--G** **T--Q**	**1188**	**1.48**
43	Q--R	682	0.85
45	**E--D**	**2256**	**2.81**
46	**Q--N** **S--N**	**1290**	**1.61**
47	T--H	156	0.19
50	V--F	374	0.47
53	**K--Q**	**801**	**1.00**
54	E--R	480	0.60
55	V--Y	238	0.30
56	**K--E** **S--G**	**2459**	**3.06**
58	**T--A** **T--S**	**3267**	**4.07**
59	T--G	393	0.49
60	**G--A**	**2678**	**3.33**

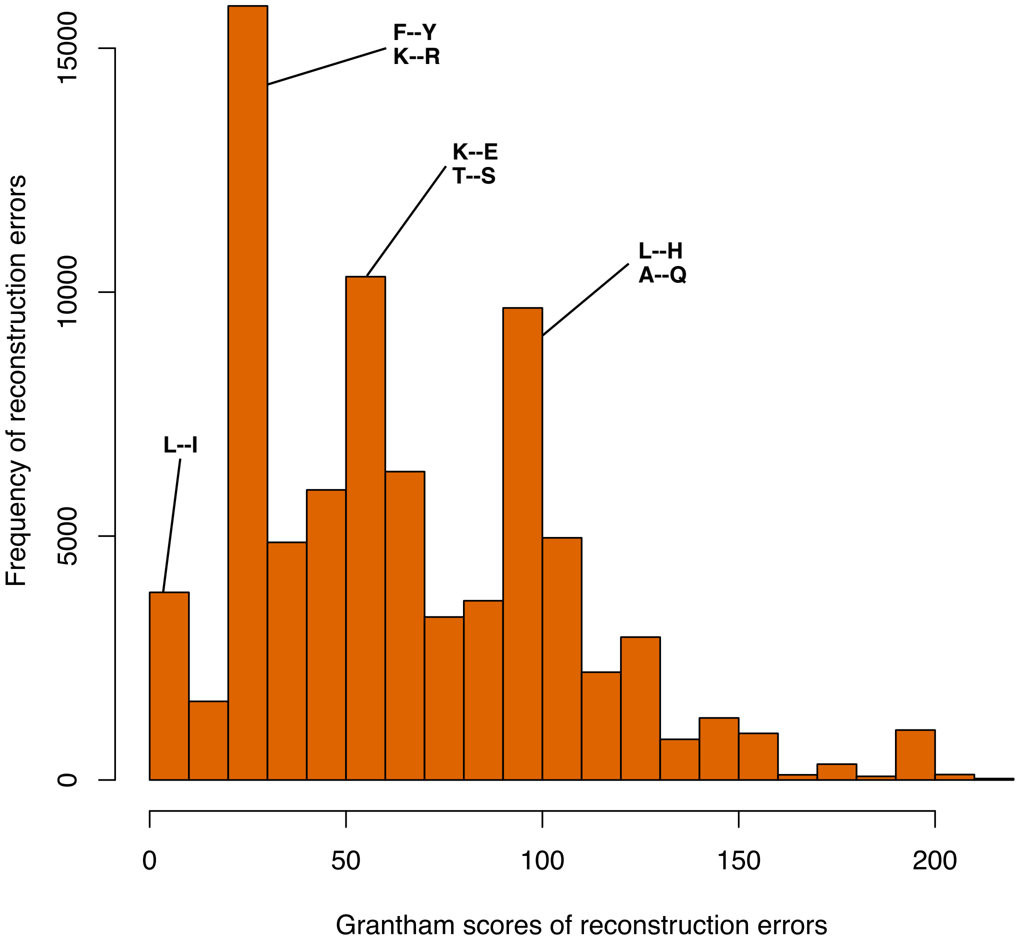
61	Q--D W--I W--L	744	0.93
64	**V--A**	**1762**	**2.19**
65	**S--D** **T--N** **T--E**	**1739**	**2.17**
67	W--M	68	0.08
68	**H--N** **S--Q** **V--P**	**1097**	**1.37**
69	**V--T**	**911**	**1.13**
71	T--R	384	0.48
74	S--P	644	0.80
76	P--Q	215	0.27
77	P--H Y--R	222	0.28
78	T--K	506	0.63
80	**G--N** **S--E**	**1370**	**1.71**
81	H--D T--M	414	0.52
83	Y--H	555	0.69
84	M--A V--H	356	0.44
85	T--D Y--K	506	0.63
86	N--R H--A	573	0.71
87	G--Q M--H P--M	383	0.48
88	V--W	48	0.06
89	**S--H** **T--I**	**839**	**1.04**
91	**Q--A** **M--R** **P--N**	**822**	**1.02**
92	T--L Y--T	560	0.70
93	P--E	387	0.48
94	**G--D** **I--A** **I--H** **K--N**	**1874**	**2.33**

95	M--K P--I	204	0.25
96	**D--R** **L--A** **V--R** **V--Q**	**1334**	**1.66**
97	I--R F--R V--K	472	0.59
98	**G--E** **H--G** **P--L**	**951**	**1.18**
99	**L--H** **S--A** **Y--Q**	**2857**	**3.56**
100	F--H	215	0.27
101	K--D M--Q W--R	683	0.85
102	L--R K--I F--K	576	0.72
103	P--R P--K T--F	581	0.72
106	K--A	611	0.76
107	**E--A** **K--L**	**1394**	**1.74**
108	P--D	239	0.30
109	I--Q V--G	301	0.37
110	S--R W--K Y--P	579	0.72
111	N--A	414	0.52
112	**R--A** **S--C** **Y--A**	**1068**	**1.33**
113	L--Q F--A	548	0.68
114	P--F	57	0.07
115	W--H	54	0.07
116	F--Q	69	0.09
121	S--K V--E	795	0.99
122	Y--E	70	0.09

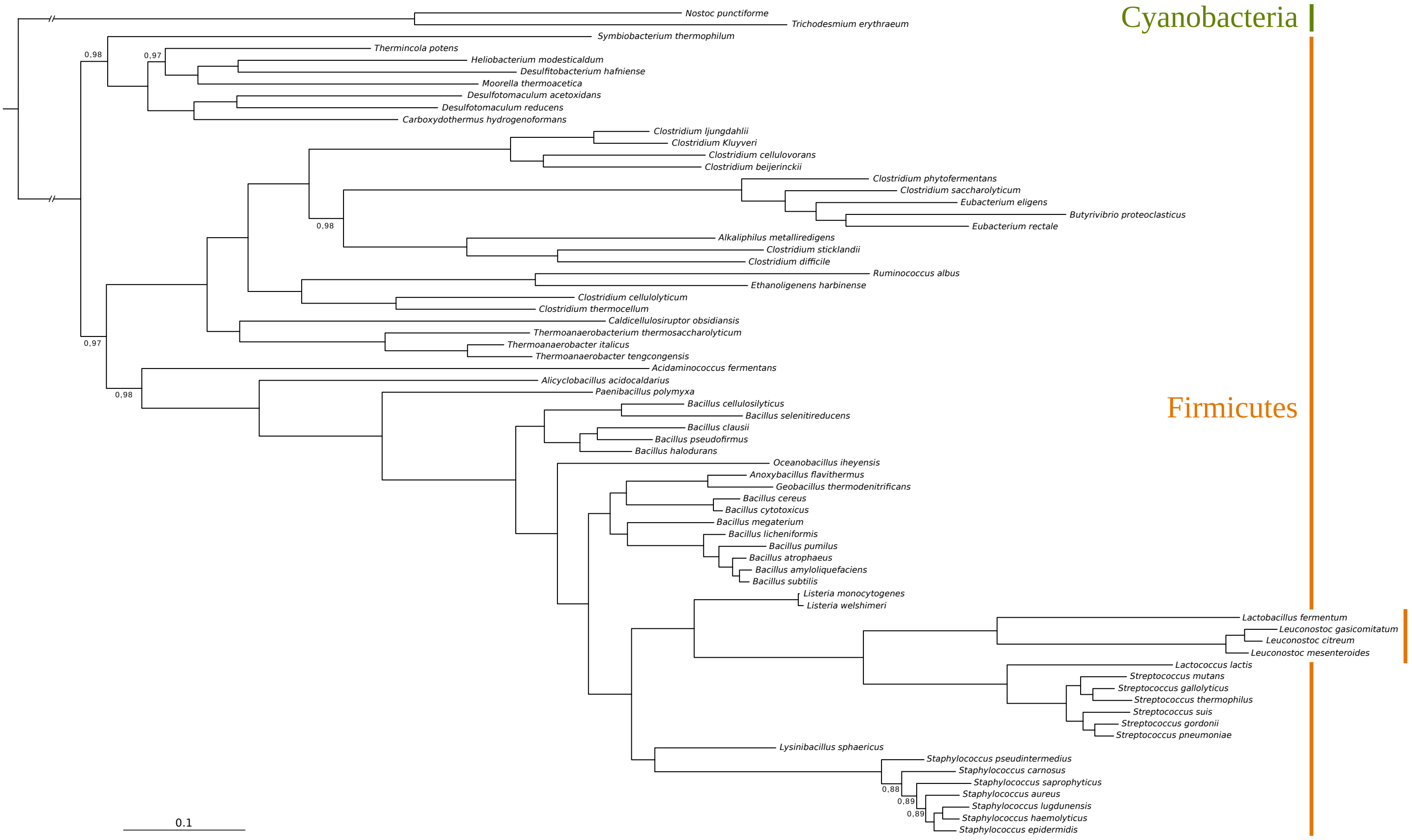
124	V--S	434	0.54
125	G--R	405	0.50
126	D--A M--E	729	0.91
127	K--G M--G	439	0.55
128	W--T	24	0.03
130	W--Q	33	0.04
133	V--N	117	0.15
134	I--E	133	0.17
135	I--G S--M	198	0.25
138	L--E L--G	295	0.37
139	C--N	55	0.07
140	F--E	36	0.04
142	M--N S--I	248	0.31
143	Y--N	171	0.21
144	Y--S	135	0.17
145	S--L	320	0.40
147	W--P Y--G	84	0.10
148	W--A	47	0.06
149	I--N T--C	267	0.33
152	W--E V--D	179	0.22
153	L--N F--G	204	0.25
154	C--D Q--C	64	0.08
155	S--F	148	0.18
158	F--N	69	0.09
159	G--C	172	0.21
160	M--D Y--D	120	0.15
168	I--D	55	0.07
169	P--C	17	0.02
170	E--C	34	0.04
172	L--D	101	0.13
174	H--C W--N	62	0.08
177	F--D W--S	93	0.12
180	C--R	69	0.09

181	W--D	22	0.03
184	W--G	53	0.07
192	V--C	250	0.31
194	Y--C	69	0.09
195	C--A	413	0.51
196	M--C	40	0.05
198	I--C L--C	252	0.31
202	K--C	27	0.03
205	F--C	85	0.11
215	W--C	29	0.04

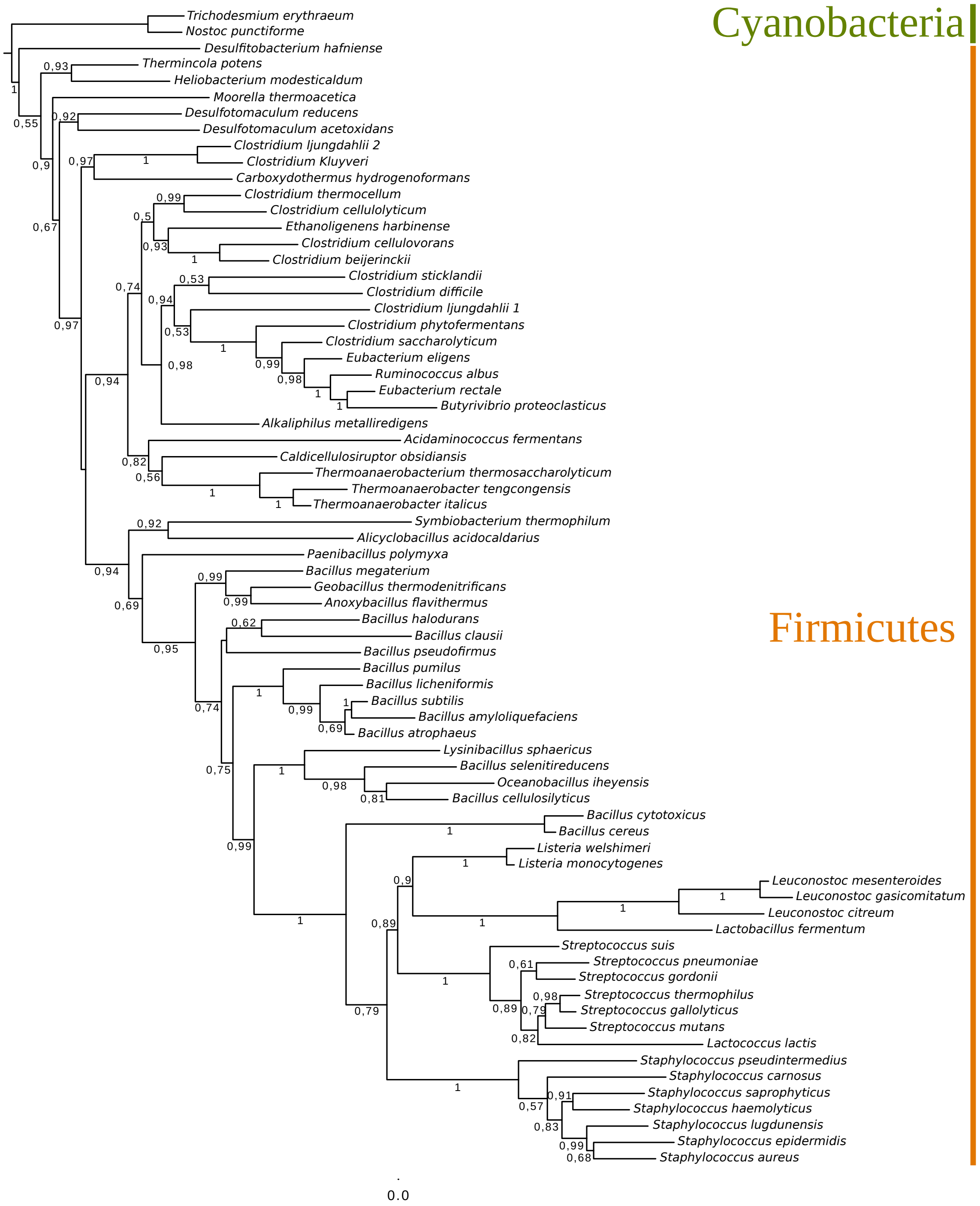
Supplementary Figures



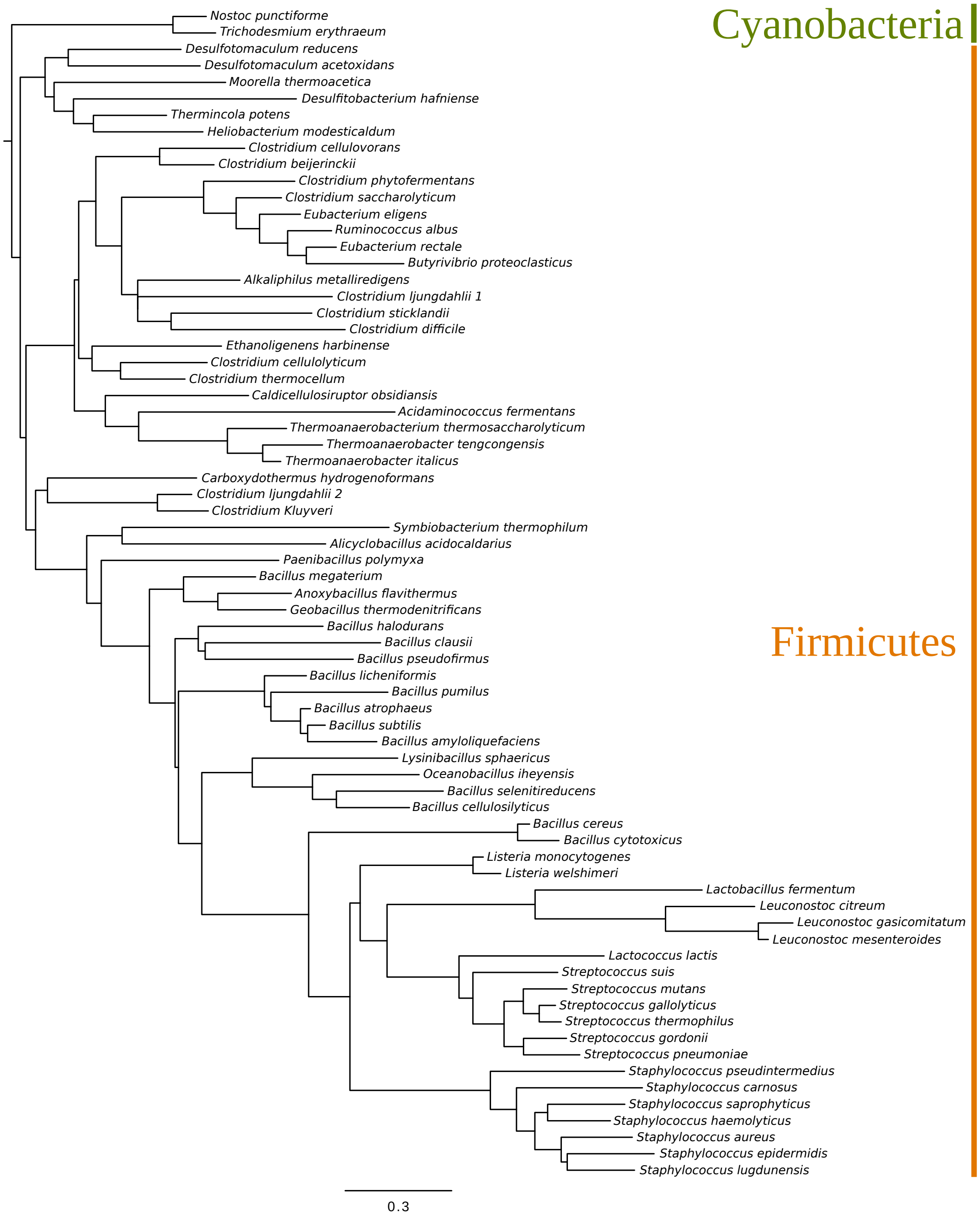
Supplementary Figure 1: **Distribution of the Grantham scores for reconstruction errors obtained with the LG S-unaware trees.** The distribution of Grantham scores for the S-unaware trees reconstructed with the C60 model or for the S-aware trees were similar and are not represented here. For each reconstruction error, the biochemical distance between the two different amino acids was determined with the Grantham matrix, which accounts for volume, polarity and composition of the amino acids. For certain categories of Grantham scores, examples of pairs of amino acids are shown.


Nostoc punctiforme
Trichodesmium erythraeum
Cyanobacteria
Symbiobacterium thermophilum
0,98
0,97
Thermincola potens
Heliobacterium modesticaldum
Desulfitobacterium hafniense
Moorella thermoacetica
Desulfotomaculum acetoxidans
Desulfotomaculum reducens
Carboxydothermus hydrogenoformans
Clostridium ljungdahlii
Clostridium Kluyveri
Clostridium cellulovorans
Clostridium beijerinckii
Clostridium phytofermentans
Clostridium saccharolyticum
Eubacterium eligens
Butyrivibrio proteoclasticus
0,98
Eubacterium rectale
Alkaliphilus metalliredigens
Clostridium sticklandii
Clostridium difficile
Ruminococcus albus
Ethanoligenens harbinense
Clostridium cellulolyticum
Clostridium thermocellum
Caldicellulosiruptor obsidiansis
Thermoanaerobacterium thermosaccharolyticum
0,97
Thermoanaerobacter italicus
Thermoanaerobacter tengcongensis
Acidaminococcus fermentans
Alicyclobacillus acidocaldarius
0,98
Paenibacillus polymyxa
Bacillus cellulosilyticus
Firmicutes
Bacillus selenitireducens
Bacillus clausii
Bacillus pseudofirmus
Bacillus halodurans
Oceanobacillus iheyensis
Anoxybacillus flavithermus
Geobacillus thermodenitrificans
Bacillus cereus
Bacillus cytotoxicus
Bacillus megaterium
Bacillus licheniformis
Bacillus pumilus
Bacillus atrophaeus
Bacillus amyloliquefaciens
Bacillus subtilis
Listeria monocytogenes
Listeria welshimeri
Lactobacillus fermentum
Leuconostoc gasicomitatum
Leuconostoc citreum
Leuconostoc mesenteroides
Lactococcus lactis
Streptococcus mutans
Streptococcus gallolyticus
Streptococcus thermophilus
Streptococcus suis
Streptococcus gordonii
Streptococcus pneumoniae
Lysinibacillus sphaericus
Staphylococcus pseudintermedius
Staphylococcus carnosus
0,88
Staphylococcus saprophyticus
0,89
Staphylococcus aureus
0,89
Staphylococcus lugdunensis
Staphylococcus haemolyticus
0.1
Staphylococcus epidermidis

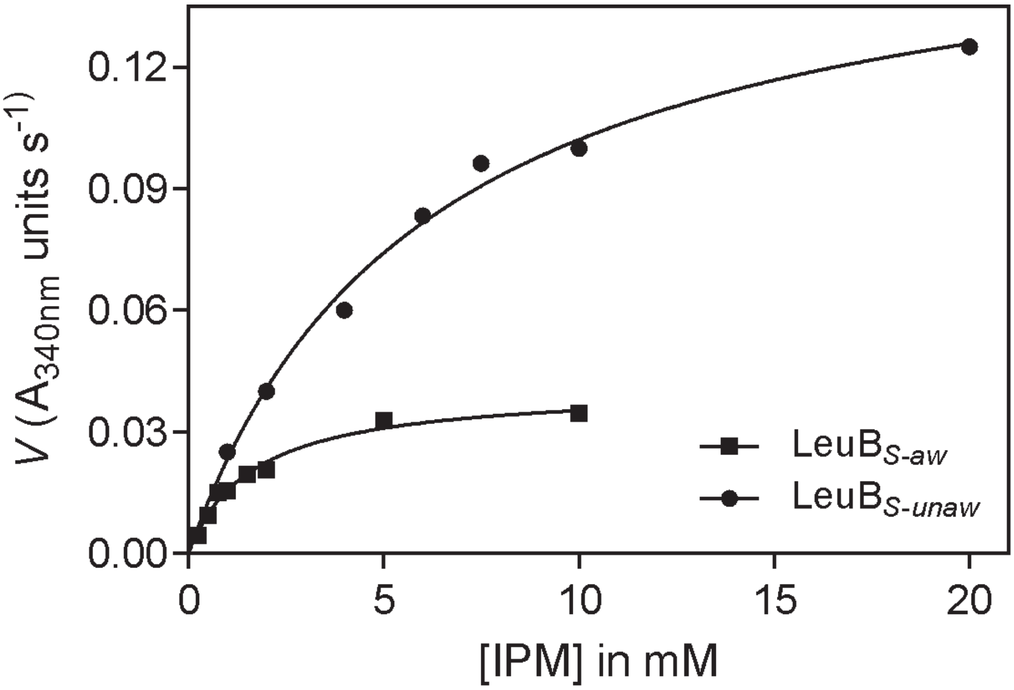
Supplementary Figure 2: **Species tree of Firmicutes**. This consensus posterior tree was reconstructed with the CAT model in Phylobayes. All branches with no support information have a PP of 1.


Cyanobacteria
Firmicutes
Trichodesmium erythraeum
Nostoc punctiforme
Desulfitobacterium hafniense
Thermincola potens
Heliobacterium modesticaldum
Moorella thermoacetica
Desulfotomaculum reducens
Desulfotomaculum acetoxidans
Clostridium ljungdahlii 2
Clostridium Kluyveri
Carboxydothermus hydrogenoformans
Clostridium thermocellum
Clostridium cellulolyticum
Ethanoligenens harbinense
Clostridium cellulovorans
Clostridium beijerinckii
Clostridium sticklandii
Clostridium difficile
Clostridium ljungdahlii 1
Clostridium phytofermentans
Clostridium saccharolyticum
Eubacterium eligens
Ruminococcus albus
Eubacterium rectale
Butyrivibrio proteoclasticus
Alkaliphilus metalliredigens
Acidaminococcus fermentans
Caldicellulosiruptor obsidiansis
Thermoanaerobacterium thermosaccharolyticum
Thermoanaerobacter tengcongensis
Thermoanaerobacter italicus
Symbiobacterium thermophilum
Alicyclobacillus acidocaldarius
Paenibacillus polymyxa
Bacillus megaterium
Geobacillus thermodenitrificans
Anoxybacillus flavithermus
Bacillus halodurans
Bacillus clausii
Bacillus pseudofirmus
Bacillus pumilus
Bacillus licheniformis
Bacillus subtilis
Bacillus amyloliquefaciens
Bacillus atrophaeus
Lysinibacillus sphaericus
Bacillus selenitireducens
Oceanobacillus iheyensis
Bacillus cellulosilyticus
Bacillus cytotoxicus
Bacillus cereus
Listeria welshimeri
Listeria monocytogenes
Leuconostoc mesenteroides
Leuconostoc gasicomitatum
Leuconostoc citreum
Lactobacillus fermentum
Streptococcus suis
Streptococcus pneumoniae
Streptococcus gordonii
Streptococcus thermophilus
Streptococcus gallolyticus
Streptococcus mutans
Lactococcus lactis
Staphylococcus pseudintermedius
Staphylococcus carnosus
Staphylococcus saprophyticus
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Staphylococcus epidermidis
Staphylococcus aureus
0.0

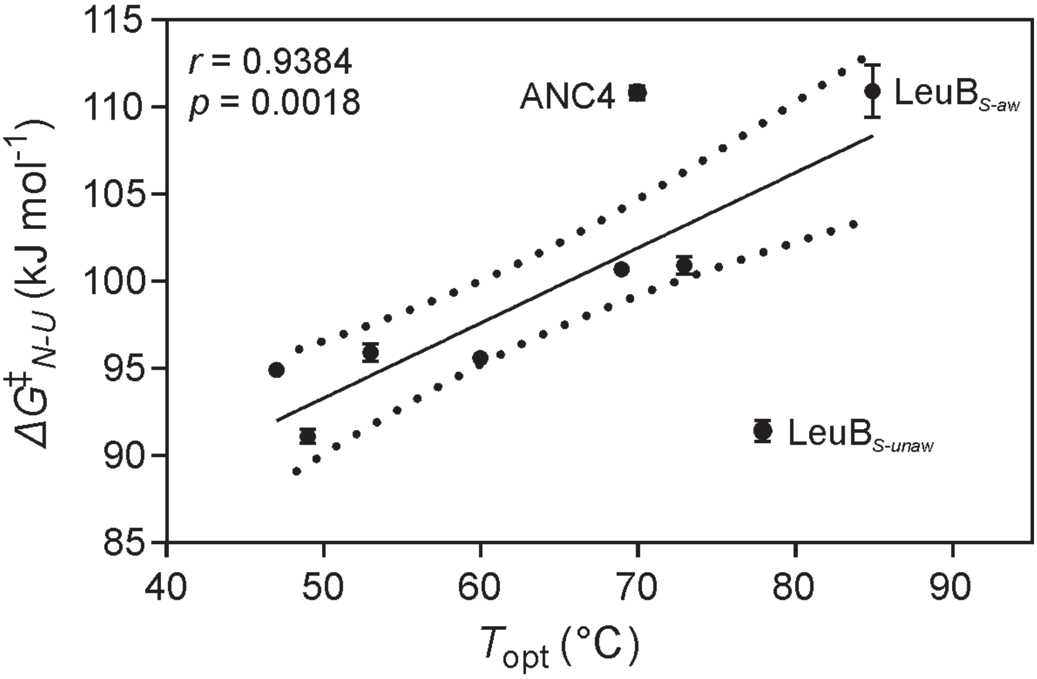
Supplementary Figure 3: **Species tree-unaware Gene tree of LeuB sequences**. This tree represents the consensus posterior tree of LeuB sequences reconstructed with PhyloBayes, using the LG+$\Gamma(4)$ model.


Cyanobacteria
Firmicutes
Nostoc punctiforme
Trichodesmium erythraeum
Desulfotomaculum reducens
Desulfotomaculum acetoxidans
Moorella thermoacetica
Desulfitobacterium hafniense
Thermincola potens
Heliobacterium modesticaldum
Clostridium cellulovorans
Clostridium beijerinckii
Clostridium phytofermentans
Clostridium saccharolyticum
Eubacterium eligens
Ruminococcus albus
Eubacterium rectale
Butyrivibrio proteoclasticus
Alkaliphilus metalliredigens
Clostridium ljungdahlii 1
Clostridium sticklandii
Clostridium difficile
Ethanoligenens harbinense
Clostridium cellulolyticum
Clostridium thermocellum
Caldicellulosiruptor obsidiansis
Acidaminococcus fermentans
Thermoanaerobacterium thermosaccharolyticum
Thermoanaerobacter tengcongensis
Thermoanaerobacter italicus
Carboxydothermus hydrogenoformans
Clostridium ljungdahlii 2
Clostridium Kluyveri
Symbiobacterium thermophilum
Alicyclobacillus acidocaldarius
Paenibacillus polymyxa
Bacillus megaterium
Anoxybacillus flavithermus
Geobacillus thermodenitrificans
Bacillus halodurans
Bacillus clausii
Bacillus pseudofirmus
Bacillus licheniformis
Bacillus pumilus
Bacillus atrophaeus
Bacillus subtilis
Bacillus amyloliquefaciens
Lysinibacillus sphaericus
Oceanobacillus iheyensis
Bacillus selenitireducens
Bacillus cellulosilyticus
Bacillus cereus
Bacillus cytotoxicus
Listeria monocytogenes
Listeria welshimeri
Lactobacillus fermentum
Leuconostoc citreum
Leuconostoc gasicomitatum
Leuconostoc mesenteroides
Lactococcus lactis
Streptococcus suis
Streptococcus mutans
Streptococcus gallolyticus
Streptococcus thermophilus
Streptococcus gordonii
Streptococcus pneumoniae
Staphylococcus pseudintermedius
Staphylococcus carnosus
Staphylococcus saprophyticus
Staphylococcus haemolyticus
Staphylococcus aureus
Staphylococcus epidermidis
Staphylococcus lugdunensis
0.3

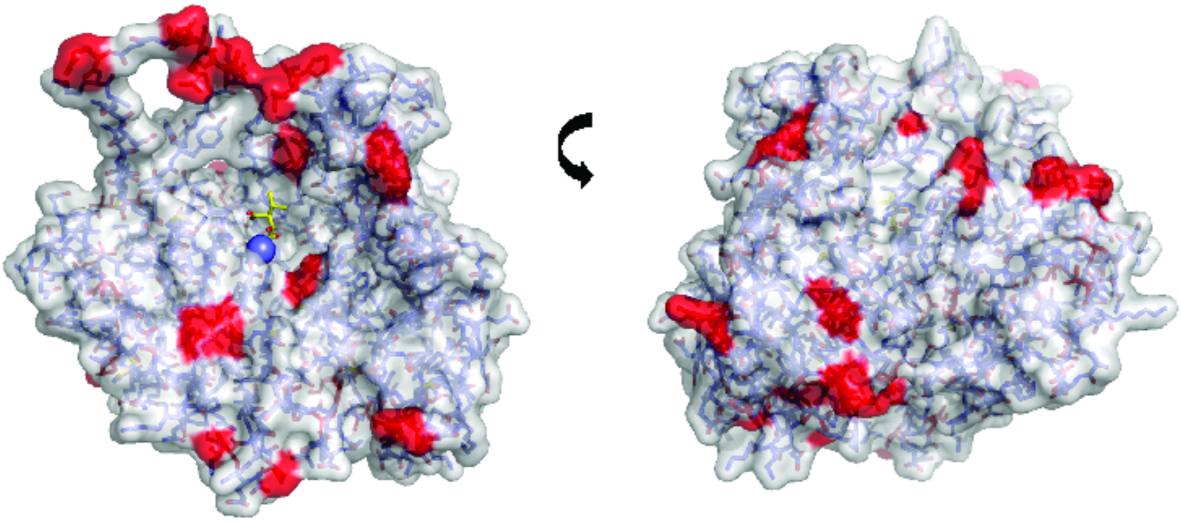
Supplementary Figure 4: **Species tree-aware (reconciled) Gene tree of LeuB sequences**. This ML reconciled joint tree was reconstructed with ALE and represents the tree that maximises the joint sequence-reconciliation likelihood.



Supplementary Figure 5: **Difference in substrate affinity between LeuB$_{S-unaw}$ and LeuB$_{S-aw}$.** K_M values were measured for the IPM substrate (see Material & Methods). K_M (IPM) for LeuB$_{S-unaw}$ is 4-fold higher than LeuB$_{S-aw}$, indicating its poorer affinity for this substrate.



Supplementary Figure 6: **Positive correlation between T_{opt} and $\Delta G^{\ddagger}_{N-U}$.** T_{opt} and $\Delta G^{\ddagger}_{N-U}$ values for each LeuB enzyme mentioned in Table 1 are plotted. T_{opt} and $\Delta G^{\ddagger}_{N-U}$ are positively and significantly correlated. The $\Delta G^{\ddagger}_{N-U}$ value of LeuB$_{S-unaw}$ falls far below the regression line.



Supplementary Figure 7: **Amino acid differences between LeuB$_{S-unaw}$ and LeuB$_{S-aw}$ mapped onto the LeuB$_{S-aw}$ structure.** A homology model of the LeuB$_{S-aw}$ enzyme was generated based on the crystal structure of the ANC4 LeuB (PDB 3U1H, see Hobbs et al. (2012)) using SWISS-MODEL (Arnold et al., 2006). The structure is depicted as a transparent surface with a stick representation beneath. Variable positions are shown in red and surface exposed residues also show a red area at the surface. The active site is shown by placing isopropylmalate (yellow sticks) and Mg^{2+} (purple sphere) from the *Acidithiobacillus ferrooxidans* structure (PDB 1A05).

References

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* pages 267–281. Petrov BN, Csaki F, editors Budapest (Hungary).

Arnold K, Bordoli L, Kopp J, and Schwede T. 2006. The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201.

Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, and Arcus VL. 2012. On the Origin and Evolution of Thermophily: Reconstruction of Functional Precambrian Enzymes from Ancestors of Bacillus. *Mol Biol Evol* 29:825–835.

Le SQ and Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25:1307–1320.

Le SQ and Gascuel O. 2010. Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst Biol* 59:277–287.

Le SQ, Gascuel O, and Lartillot N. 2008a. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Le SQ, Lartillot N, and Gascuel O. 2008b. Phylogenetic mixture models for proteins. *Phil. Trans. R. Soc. Lond. B* 363:3965–3976.

Schneider R, de Daruvar A, and Sander C. 1997. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* 25:226–230.

Whelan S and 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.