Supporting Information: Universal patterns of purifying selection at non-coding positions in bacteria

Nacho Molina and Erik van Nimwegen
Biozentrum, the University of Basel, and Swiss Institute of Bioinformatics
Klingelbergstrasse 50/70, 4056-CH, Basel, Switzerland,
email: j.molina@unibas.ch and erik.vannimwegen@unibas.ch

September 5, 2007

Contents

1 Lists of organisms in each clade

For our analyses we used 105 bacterial genomes divided into 22 different clades with 4 or 5 genomes in each clade. All genomes where downloaded from Genbank at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. Table 1 lists the names of the organisms in each clade, with the reference organism appearing on top.

Clade Name	Organism Name
Rhizobiales	Agrobacterium tumefaciens C58 UWash Rhizobium etli CFN 42 Mesorhizobium loti Sinorhizobium meliloti Brucella suis 1330
Bacillus	Bacillus subtilis Bacillus anthracis Ames Bacillus clausii KSM-K16 Bacillus halodurans

	Bacillus licheniformis ATCC 14580
Burkholderia	Burkholderia 383
	Burkholderia cenocepacia AU 1054
	Burkholderia mallei ATCC 23344
	Burkholderia thailandensis E264
	Burkholderia xenovorans LB400
Chlamydiales	Chlamydophila caviae
	Chlamydophila felis Fe C-56
	Chlamydophila abortus S26 3
	Chlamydophila pneumoniae AR39
	Chlamydia trachomatis
Clostridum	Clostridium acetobutylicum
	Clostridium perfringens
	Clostridium perfringens ATCC 13124
	Clostridium perfringens SM101
	Clostridium tetani E88
Corynebacterium	Corynebacterium glutamicum
	Corynebacterium efficiens YS-314
	Corynebacterium diphtheriae
	Corynebacterium jeikeium K411
Enterobacteria	Escherichia coli K12
	Salmonella typhi
	Yersinia pestis KIM
	Photorhabdus luminescens
	Erwinia carotovora atroseptica SCRI1043
Ehrlichia	Ehrlichia canis Jake
	Ehrlichia chaffeensis Arkansas
	Ehrlichia ruminantium Gardel
	Ehrlichia ruminantium Welgevonden
Pasteurellales	Haemophilus influenzae
	Haemophilus ducreyi 35000HP
	Pasteurella multocida
	Mannheimia succiniciproducens MBEL55E
Helicobacter	Helicobacter acinonychis Sheeba
	Helicobacter pylori 26695
	Helicobacter pylori HPAG1
	Helicobacter pylori J99
Lactobacillus	Lactobacillus acidophilus NCFM
	Lactobacillus delbrueckii bulgaricus
	Lactobacillus johnsonii NCC 533
	Lactobacillus plantarum
	Lactobacillus salivarius UCC118
Mycobacterium	Mycobacterium tuberculosis CDC1551
	Mycobacterium leprae
	Mycobacterium avium paratuberculosis
	Mycobacterium MCS
	Nocardia farcinica IFM10152
Prochloroccus	Prochlorococcus marinus MIT9313
	Prochlorococcus marinus CCMP1375
	Prochlorococcus marinus MED4
	Prochlorococcus marinus NATL2A
	Synechococcus sp WH8102
Pseudomonas	Pseudomonas syringae
	Pseudomonas fluorescens PfO-1
	Pseudomonas entomophila L48
	Pseudomonas putida KT2440

	Pseudomonas aeruginosa
Ralstonia	Ralstonia eutropha JMP134 Ralstonia metallidurans CH34 Ralstonia solanacearum Bordetella parapertussis Bordetella pertussis
Rhodopseudomonas	Rhodopseudomonas palustris BisB18 Rhodopseudomonas palustris BisB5 Rhodopseudomonas palustris CGA009 Rhodopseudomonas palustris HaA2 Bradyrhizobium japonicum
Rickettsia	Rickettsia typhi wilmington Rickettsia prowazekii Rickettsia felis URRWXCal2 Rickettsia conorii Rickettsia bellii RML369-C
Staphylococcus	Staphylococcus aureus N315 Staphylococcus aureus MW2 Staphylococcus epidermidis RP62A Staphylococcus haemolyticus Staphylococcus saprophyticus
Streptococcus	Streptococcus pneumoniae TIGR4 Streptococcus agalactiae 2603 Streptococcus pyogenes Streptococcus mutans Streptococcus thermophilus LMG 18311
Synechoccus	Synechococcus CC9605 Synechococcus CC9902 Synechococcus sp WH8102 Synechococcus elongatus PCC 6301 Synechocystis PCC6803
Vibrio	Vibrio cholerae Vibrio vulnificus CMCP6 Vibrio vulnificus YJ016 Vibrio parahaemolyticus Vibrio fischeri ES114
Xanthomonas	Xanthomonas campestris Xanthomonas campestris vesicatoria 85-10 Xanthomonas citri Xanthomonas oryzae MAFF 311018

Table 1: Names of all the organisms used in our study. Different clades are separated by horizontal lines. The names by which we refer to each clade are shown on the left. The reference species in each clade appears at the top of the list of species for each clade

2 Number of operons as a function of the total number of genes

The recent operon prediction algorithm of [1] uses a Bayesian method to estimate the probability that two contiguous genes belong to the same operon. We used the operon predictions that are available for 416 bacterial genomes from http://www.microbesonline.org/operons/, and which use a probability of $1/2$ as a cut-off. Figure 1 shows the total number of operons as a function of the total number of genes in the genome for all 416 genomes. We fitted the operon numbers to a power-law as a function of the total number of genes

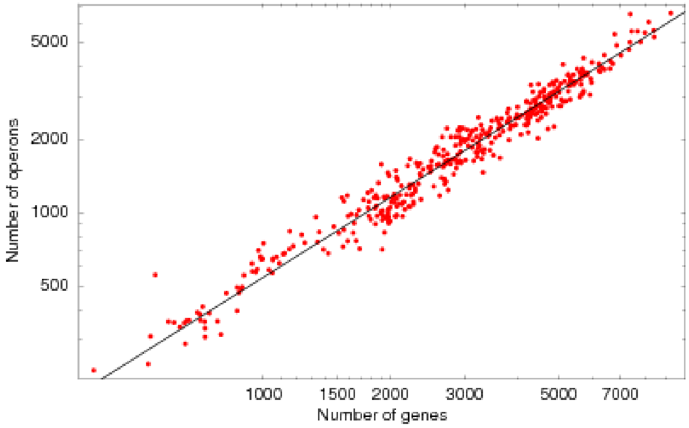


Figure 1: The estimated number of operons (vertical axis) as a function of the total number of genes (horizontal axis) for all 416 currently fully-sequenced genes in the NCBI database. Each red dot corresponds to one genome. The black line shows a power-law fit.

in the genome (black line). The fitted line has exponent 1.09.

3 Phylogenetic trees

The inferred phylogenetic trees for all 22 clades are as a supplementary file in Newick format.

4 R values versus total branch length in the phylogenetic tree

For each clade we calculate the total branch length T in its phylogenetic tree by summing the negative logarithms of the proximities q_b over all branches in the tree, i.e.

$$T = -\sum_b \log(q_b). \tag{1}$$

In figure 2 we show how the average values of R in intergenic and coding regions depend on this total branch length T. As the figure shows, there is a clear correlation between the average value of R and the total tree length, both in intergenic (red dots) and in coding regions (purple dots). For intergenic regions there seems to be an approximately linear relationship whereas for coding regions R seems to increase even faster

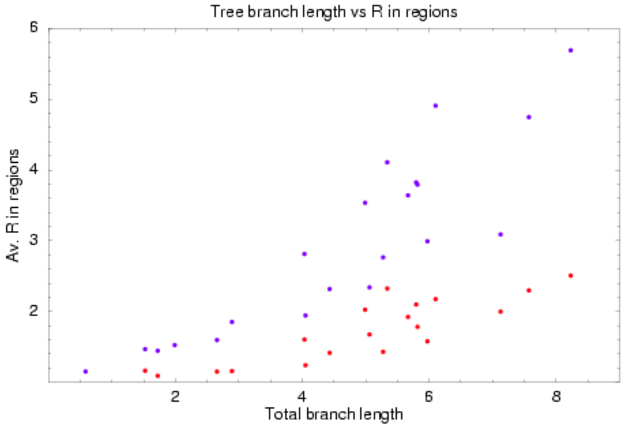


Figure 2: Average R values (vertical axis) as a function of total branch length T of the phylogenetic tree of the clade (horizontal axis). Each dot represents one of the 22 clades. Red dots show the average R value in intergenic regions (averaged over DR, SR, and NR regions). The purple dots shows the average R values in coding regions.

than linearly. The reason there is this general correlation between R and the length of the branches in the tree is that for longer branches the evidence of selection is easier to detect than for short branches, i.e. for very close species most bases are already conserved due to evolutionary proximity.

5 Evidence of selection as a function of genome size

In this section we use 12 different statistics to investigate if a correlation between genome size and the amount of evidence for selection in intergenic regions can be detected. In figure 3 we show 4 different R value statistics as a function of the number of genes in the genome. In the top left panel of Fig. 3 we show the average value of R, averaged over all SR and DR regions, directly against the number of genes in the genome. There is no significant correlation (p-value 0.23). We observed in the main paper that R values in DR and SR (upstream) regions are substantially higher than those in NR (downstream) regions, which is most likely the result of regulatory elements being more abundant upstream of genes than in regions downstream of genes. Therefore, one might argue that a more ‘accurate’ assessment of the density of regulatory sites can be made by comparing the R values in SR and DR regions with those in NR regions. In the upper-right panel we show the ratio of the average R values in SR and DR regions and the average R in NR regions as a function of the number of genes in the genome. Again there is no significant correlation (p-value 0.21). The difference between R values in SR and DR regions and R values in NR regions also shows no correlation with genome size (data not shown). Another issue that might complicate observation of a correlation with genome size is that the rate of turnover of regulatory sites may be significantly different in different clades. Of course, given that we do

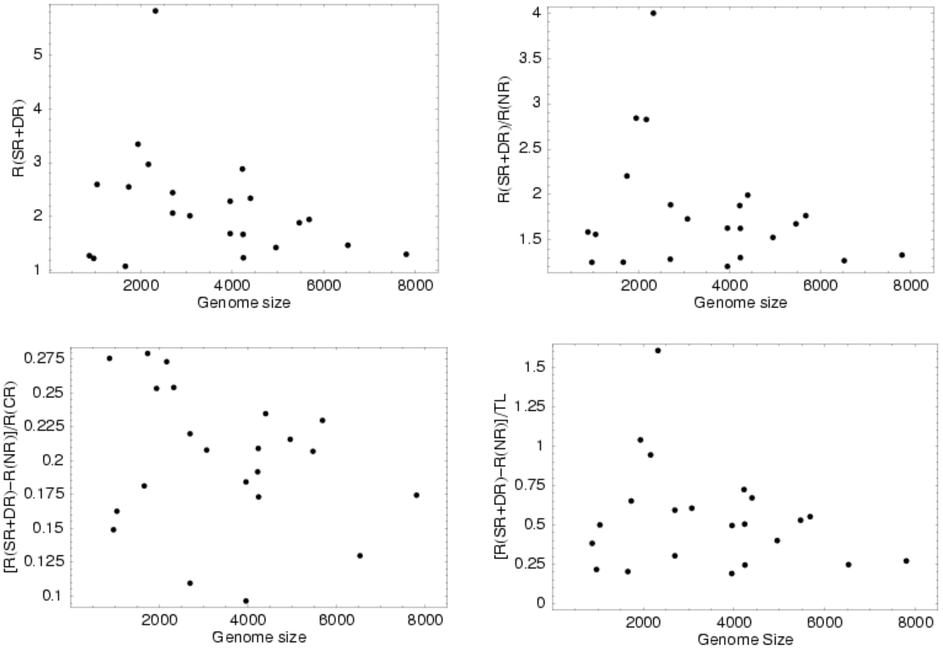


Figure 3: Various average R statistics as a function of genome size. In each panel each dot represents one clade. The horizontal axis in each panel shows the total number of genes in the reference species of the clade. The vertical axes show respectively **Top left:** The average value of R in SR and DR regions. **Top right:** The ratio between the average value of R in SR and DR regions and the average value of R in NR regions. **Bottom-left:** The difference between R in SR and DR regions and the average R in NR regions, relative to the average R value in coding positions. **Bottom-right:** The difference between the average R in SR and DR regions and the average R in NR regions relative to the total branch length in the tree.

not know what the TFs in almost all of these genomes bind, it is hard to estimate the rate of regulatory site turnover directly. However, we would generally expect the rate of turnover to be smallest if the organisms in the clade occupy very similar niches. To some extent we can estimate this from the rate of protein evolution. That is, the amount of conservation at the amino acid level will be higher for organisms living in a similar niche, compared to those that occupy different niches. In the lower-left panel of Fig. 3 we show the relative *difference* $[R(\mathrm{SR}+\mathrm{DR})-R(\mathrm{NR})]\,/R(\mathrm{CR})$ between R in SR and DR regions and R in NR regions, *relative* to the average R in coding positions $R(\mathrm{CR})$. That is, we have normalized the difference between R in upstream and downstream regions to the R values at coding positions. We again see that there is no significant correlation (p-value 0.24). Finally, we also saw that R values generally correlate positively with the sum of the branch lengths in the phylogenetic tree of the clade. Therefore, one might argue that to obtain properly 'normalized' R values we should divide the R values by the total branch length in the tree. In the bottom-right panel of Fig. 3 we show the relative difference $[R(\mathrm{SR}+\mathrm{DR})-R(\mathrm{NR})]\,/\mathrm{TL}$ relative to the tree length TL. Here too there is no correlation (p-value 0.29). We also tried other combinations such as non-normalized differences, or normalized versions of $R(\mathrm{SR}+\mathrm{DR})$ but none gave significant correlations (data not shown).

Finally, note that the R values are calculated compared to what would be expected based on the phylogenetic tree of the species, which was calculated from the silent positions in genes. If intergenic regions are subject to different mutational mechanisms than coding regions than the tree inferred from silent positions may not be appropriate for intergenic regions. To control for this possibility we also build phylogenetic trees from the NR regions in the clade and then calculated R values in intergenic regions using this phylogenetic tree. The results again showed no signs of correlations between the R values in upstream regions and the genome size (data not shown).

5.1 Substitution rate reduction

In the supplementary methods we detail how we calculate, for each alignment column C, a statistic $Q(C)$ that quantifies the extent to which the effective substitution rate in this column is *reduced* compared to what would be expected from the background model. This Q statistic is thus an alternative measure for the strength of selection in an alignment column that doesn't intrinsically scale with the length of the branches in the phylogenetic tree. We calculated the values $Q(C)$ for all alignment columns in all clades and investigated if the average Q values in different regions correlate with genome size. The results are shown in Fig. 4

In the top-left panel we show the average Q in SR and DR (upstream) regions as a function of the number of genes in the reference species of the clade. Although by eye there may appear to be some negative correlation, this correlation is not significant (p-value 0.38). Note though that even if the correlation was significant it would go in the *wrong direction*, i.e. larger genomes would show less evidence of selection. In the top-right panel we show the average Q in SR and DR (upstream) regions relative to the average Q in NR (downstream) regions. As discussed before, we generally find more evidence of selection in upstream (SR and DR) regions than in downstream (NR) regions and we interpret this as the result of a higher density of regulatory sites

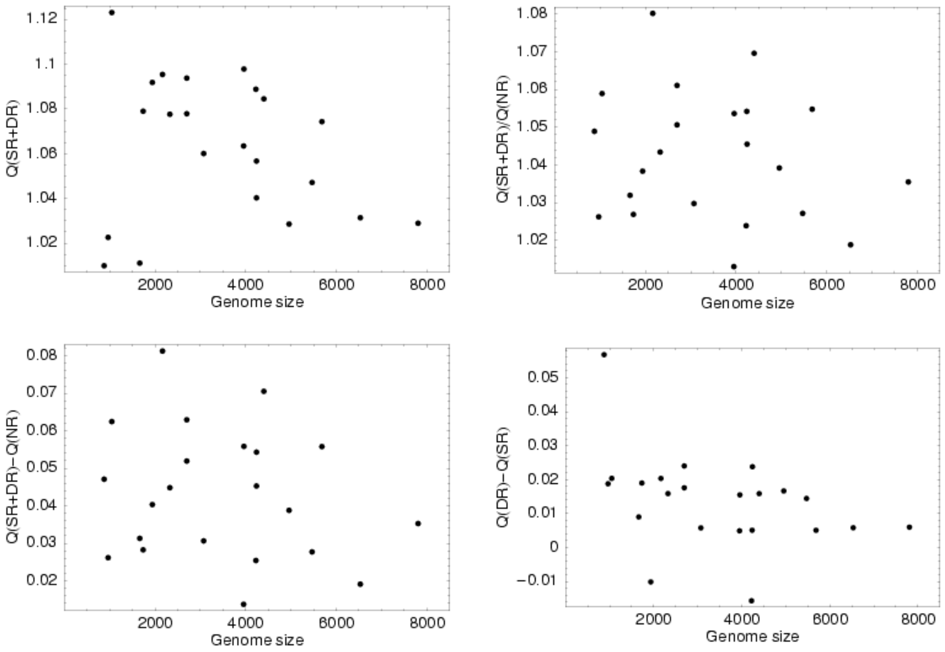


Figure 4: Estimated average substitution rate statistics as a function of the number of genes in the genome. In each panel each dot represents one clade. The horizontal axis in each panel shows the total number of genes in the reference species of the clade. The vertical axis in each panel shows: **Top-left:** The average Q in SR and DR regions **Top-right:** The ratio between the average Q in SR and DR regions and the average Q in NR regions **Bottom-left:** The difference between the average Q in SR and DR regions and the average Q in NR regions. **Bottom-right:** The difference between the average Q in DR regions and the average Q in SR regions.

in upstream than in downstream regions. Thus, it can be argued that the abundance of regulatory sites should be reflected in the relative sizes of Q in upstream and downstream regions. However, we see in the upper-right panel that there is no correlation whatsoever between this relative substitution rate and genome size (p-value 0.40). Instead of the *ratio* of Q values in upstream and downstream regions we can also consider their *difference* and this is shown in the bottom-left panel of Fig. 4. Again there is no evidence of correlation with genome size (p-value 0.38). Finally, we have also observed that DR regions often show more evidence of selection than SR regions. In the bottom-right panel we show the difference between the average Q value in DR regions and the average Q value in DR regions as a function of genome size. Here there is a marginally significant correlation (p-value 0.07), but again this correlation goes in the wrong direction, i.e. the difference in Q value between DR and SR regions is less in larger genomes.

5.2 Branch lengths inferred by PAML

Instead of using our methods to estimate the strength of selection we performed an analogous analysis using PAML. In particular, we looked for correlations between the number of genes in the genome and the branch lengths inferred by the PAML algorithm for alignment columns from different regions. The results are shown in Fig. 5

For each clade, we let PAML infer 12 different phylogenetic trees and calculated the total branch length in each of the trees. One tree was inferred from all alignment columns in NR regions. We denote its branch length by BL(NR). The second tree was inferred from all alignment columns in SR regions, and we denote its total branch length by BL(SR). The third tree was inferred from all alignment columns in DR regions and we denote its total branch length by BL(DR). We denote by BL(SR+DR) the average of BL(SR) and BL(DR). Finally, 8 different trees were inferred from the silent positions of each of the 8 fourfold degenerate codons. We denote by BL(syn) the median of the total branch lengths of these 8 trees.

In the top-left panel we show BL(SR+DR)/BL(syn) for each clade as a function of the total number of genes in the reference species of that clade. That is, we compare the branch lengths in upstream regions with those at silent positions. Selection conserving regulatory elements in upstream regions would lead to lowered branch lengths in upstream regions relative to silent positions. As the density of regulatory sites increase the ratio BL(SR+DR)/BL(syn) should thus decrease. However, as the figure shows, even though the ratio is less than 1 in all clades, there is no observable correlation between these branch lengths and genome size (p-value 0.20). In the top-right panel we show the ratio BL(SR+DR)/BL(NR), that is the total branch length in upstream regions relative to the total branch length in downstream regions. Upstream regions are expected to contain much more regulatory elements than downstream regions so that it can be argued that the ratio BL(SR+DR)/BL(NR) quantifies the density of regulatory sites in upstream regions. However, we again observe no correlation with genome size (p-value 0.41). In the bottom-left panel we look at the relative difference [BL(NR)-BL(SR+DR)]/BL(syn). Here we look at the difference in branch lengths between upstream and downstream regions and normalize this using the branch lengths of the silent positions. Again, no correlation with genome size is observed (p-value

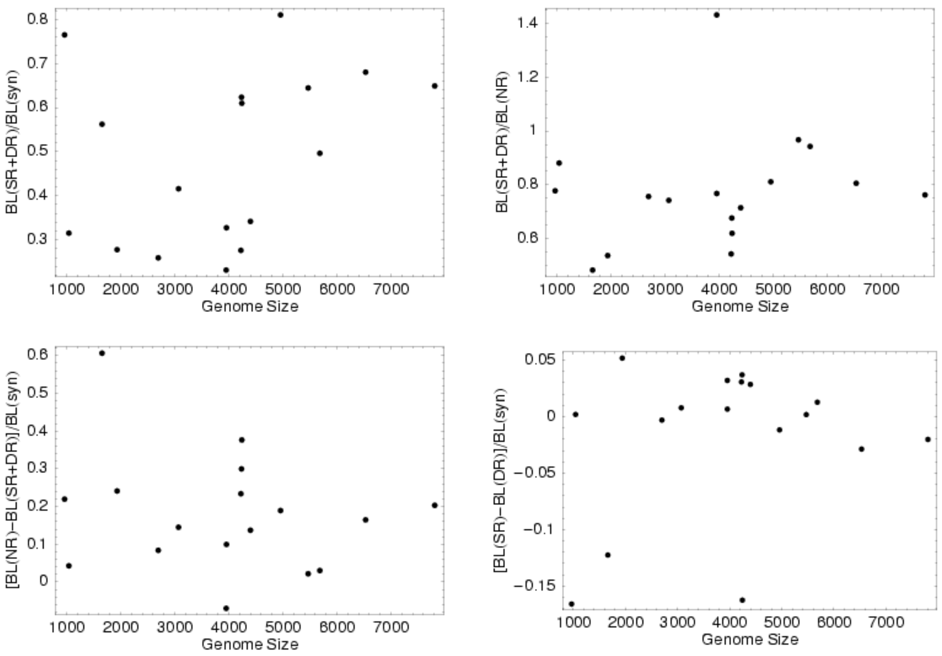


Figure 5: Statistics of total branch lengths in the phylogenetic trees of different clades, as inferred by PAML from alignment columns of different regions. Each dot in each panel represents one clade. The horizontal axis in each panel shows the total number of genes in the reference species of the clade. The vertical axes in the four panels show: **Top-left:** The average total branch lengths BL(SR+DR) in the phylogenetic trees inferred from the alignment columns of SR and DR regions relative to the total branch length BL(syn) inferred from alignment columns of silent positions. **Top-right:** The average total branch length BL(SR+DR) inferred from SR and DR regions relative to the total branch length BL(NR) inferred from alignment columns in NR regions. **Bottom-left:** The difference of the total branch length for NR regions and the total branch length for SR and DR regions relative to the total branch length for silent positions, i.e. [BL(NR)-BL(SR+DR)]/BL(syn). **Bottom-right:** The difference between the total branch length for SR regions and for DR regions relative to the total branch length for silent positions, i.e. [BL(SR)-BL(DR)]/BL(syn).

0.34). Finally, we look at the difference in total branch length for SR and DR regions (normalized again by BL(syn)). DR regions generally should have more regulatory sites than SR regions and their difference can again be argued to reflect the average density of regulatory elements per gene, but again no correlation with genome size is observed (p-value 0.40).

In summary, in spite of of using three different methods (R values, reduction in substitution rates, and branch lengths inferred by PAML), using both silent positions and NR regions to infer the phylogenetic trees, and using a number of different statistics for each of these methods, we did not find any indication that the density of regulatory sites increases with genome size (the total number of genes in the genome). Although it could be argued that more sophisticated models than the ones we employed might be able to uncover a subtle correlation it seems highly unlikely that the density of regulatory sites in intergenic regions changes substantially between the smallest and largest genomes. For example, the fraction of genes in the genome that are regulatory genes increases by about a factor of 20 between the smallest and largest genomes. If the density of regulatory sites would have increased by a similar factor then our methods would have detected such a increase. Note that our methods do infer more evidence of regulatory sites upstream then dowstream of genes, they detect the elevated selection at silent sites immediately downstream of translation start, and they correctly infer the strong selection on the Shine-Dalgarno sequence immediately upstream of translation start. It thus seems highly unlikely that a significant increase in the density of regulatory sites would have gone undetected.

6 R value profiles for all clades

Figures 6 and 7 show the R value profiles for all 22 clades we analyzed. Each figure shows 12 panels with 11 panels corresponding to the R value profiles in different clades and one panel corresponding to the profile averaged over all clades.

Note that although individual clades show differences in the details of the R value profiles, there are a number of features shared by essentially all clades. Selection is strongest at coding positions, in the order: second positions in codons, first positions in codons, and third positions in codons. That is, in order of the frequency with which substitutions at these positions effect the amino acid. Selection at coding positions drops at the starts and ends of genes. Silent positions away from the starts and ends of genes evolve according to the background model ($R = 1$). Selection in intergenic regions is almost always higher than at silent positions and is higher in upstream than in downstream regions. Generally selection in intergenic regions is highest immediately upstream of genes and lowest immediately downstream. There is almost always a sharp peak in selection a few bases upstream of selection start. This peak corresponds to conserved Shine-Dalgarno sequences. Finally, in all clades there is heightened selection at silent positions immediately downstream of translation start.

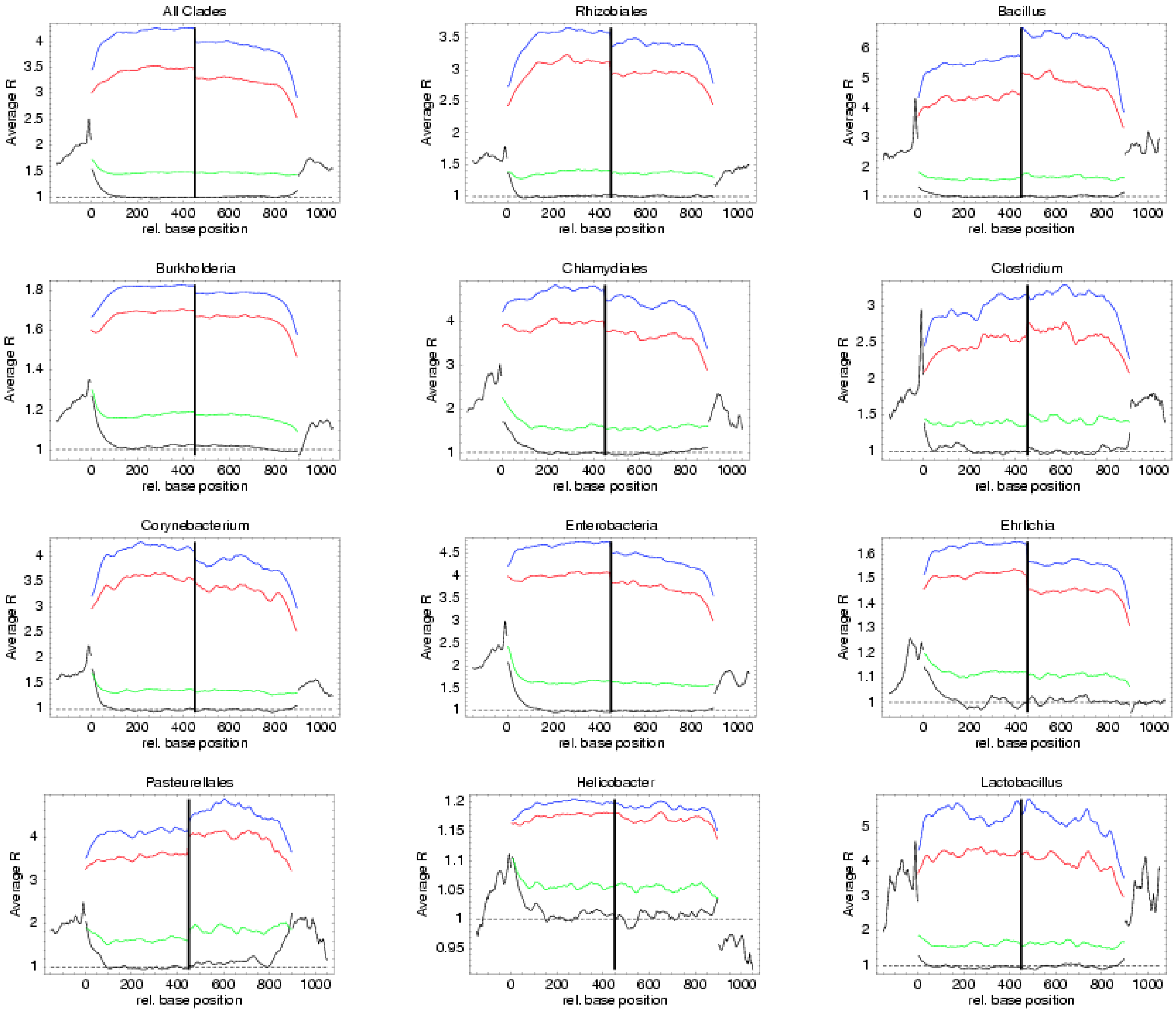


Figure 6: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows R values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

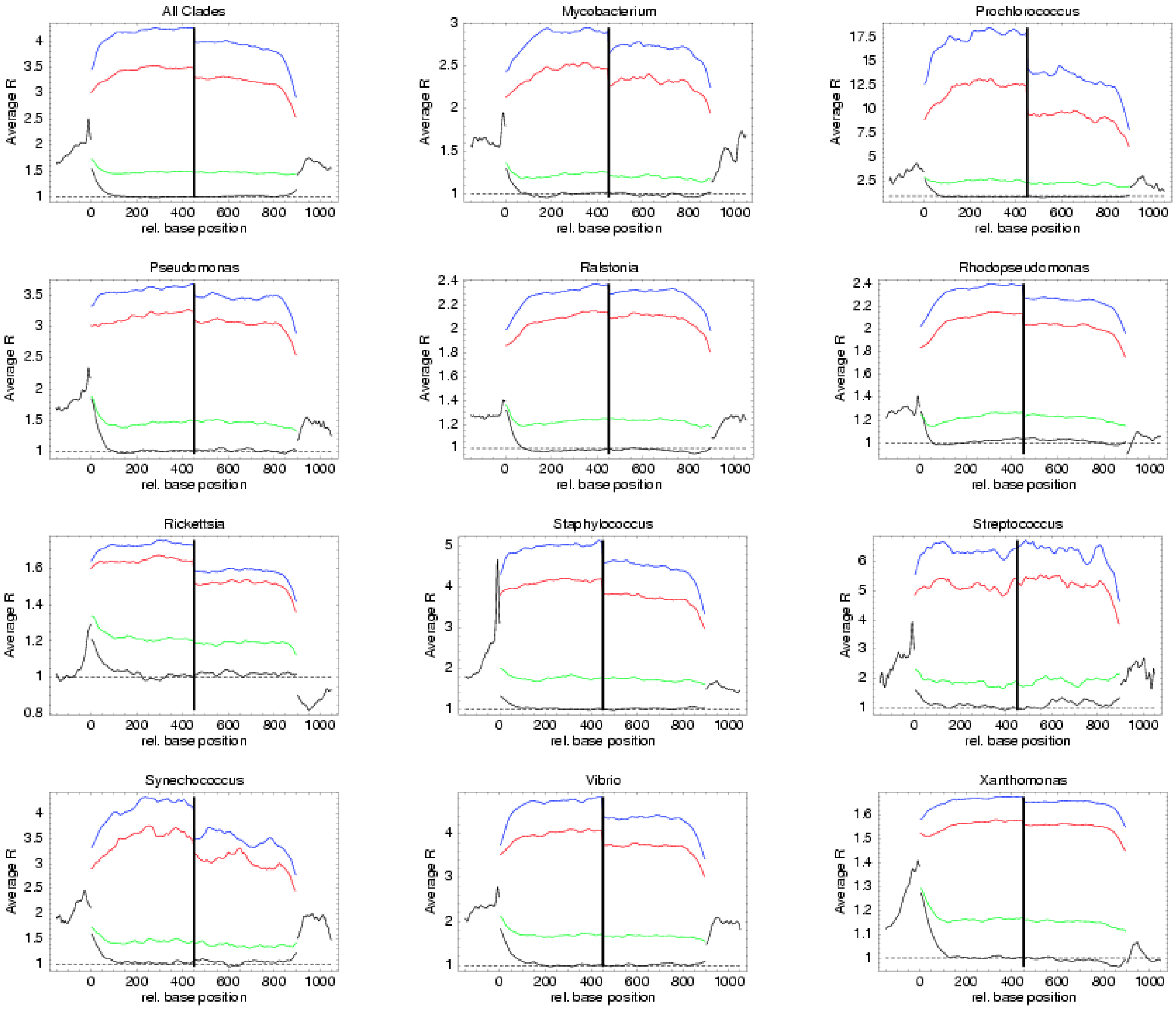


Figure 7: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows R values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

7 R value profiles at intra-operonic regions

Figures 8 and 9 show the R value profiles upstream and in genes that are not the first in their operon. The profiles in particular show the R values in intra-operonic regions and at the starts of genes that are not the first in the operon. Each figure shows 12 panels with 11 panels corresponding to the R value profiles in different clades and one panel corresponding to the profile averaged over all clades.

8 Q value profiles

Apart from the R values we also estimated, for each alignment column, the effective substitution rate statistic Q, i.e. the observed reduction in substitution rate reduction Q at this column relative to the substitution rate reduction expected from the background model (see supporting methods). These profiles are shown in figures 10 and 11.

Comparing figures 10 and 11 with the R value profiles Figs. 6 and 7 we see that all main characteristics of the R value profiles are reproduced in the Q value profiles. In fact, the two pairs of figures look very similar. The evidence of selection is highest at coding positions in the order second positions, first positions, and than third positions in codons. In most clades substitution rate reduction is lowest at silent positions in the middle of genes. As in the R profiles substitution rate reduction is higher in upstream regions than in downstream regions, generally is highest immediately upstream of translation start, and lowest immediately downstream of the stop codon. We also again see the sharp peak a few bases upstream of translation start, corresponding to the Shine-Dalgarno sequences, in most clades. Finally, the increase in selection at silent positions immediately downstream of translation start is again observed in essentially all clades.

8.1 Q value profiles for small, medium-sized, and large genomes

In complete analogy with the relative R value profiles upstream and downstream of genes shown in Fig. 5 in the main paper we calculated relative Q value profiles in the upstream regions of small, medium-size and large genomes. These profiles are shown in Fig. 12

As with the R profiles of Fig. 5 of the main paper the profiles have very similar shape and there is no clear trend that correlates with genome size.

9 Nucleotide composition profiles

We determined the average base composition at positions from 150 bps upstream of translation start to 100 bps downstream of translation start in all 22 clades of bacteria. These nucleotide composition profiles are shown in Figures 13 and 14.

Although there are significant differences in the base composition profiles between different clades, there are again several features that are universal. For example, in all clades (except for the two cyanobacteria clades Prochlorococcus and Synechococcus) there is a peak in the frequency of A nucleotides around the translation start. In

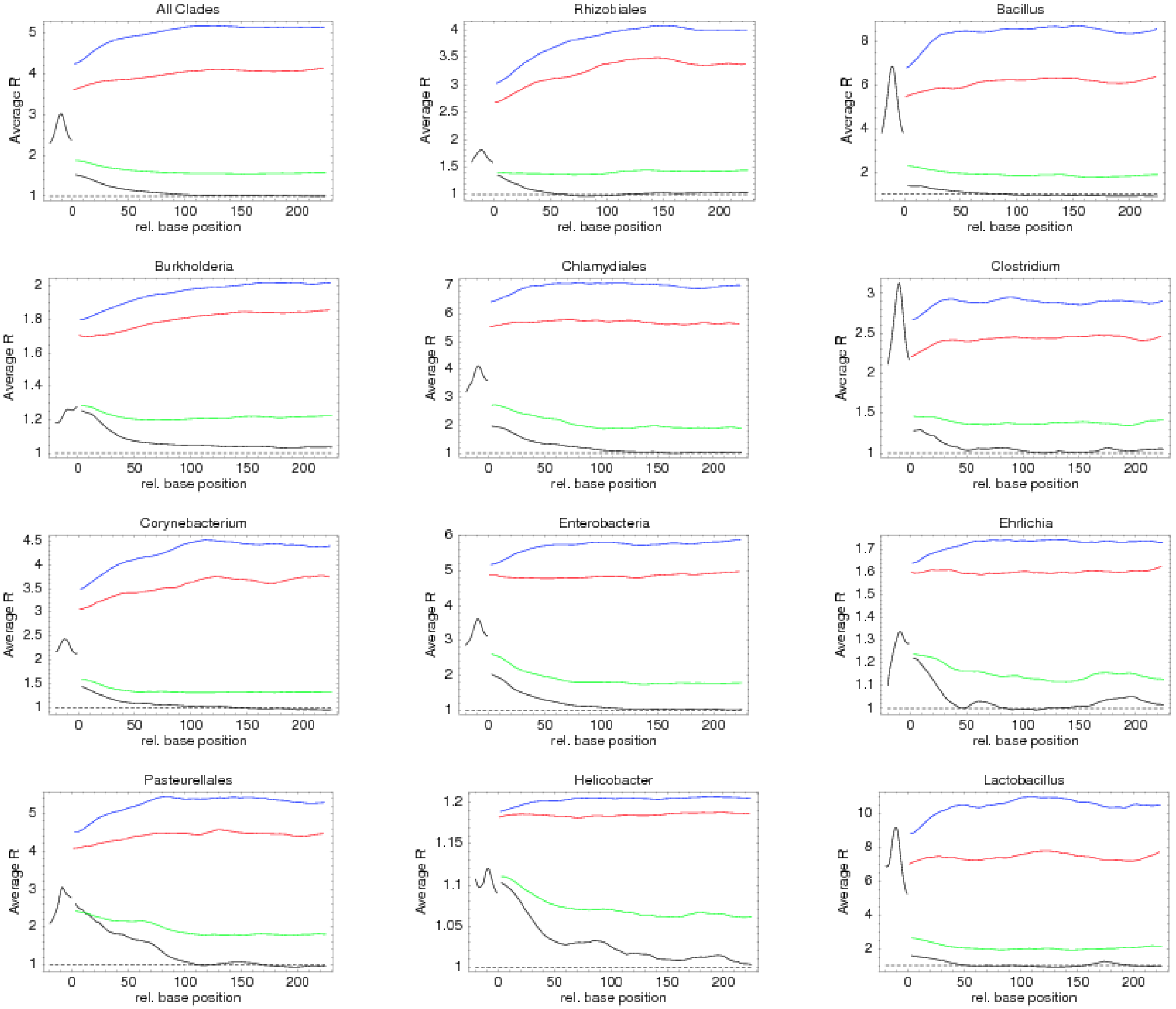
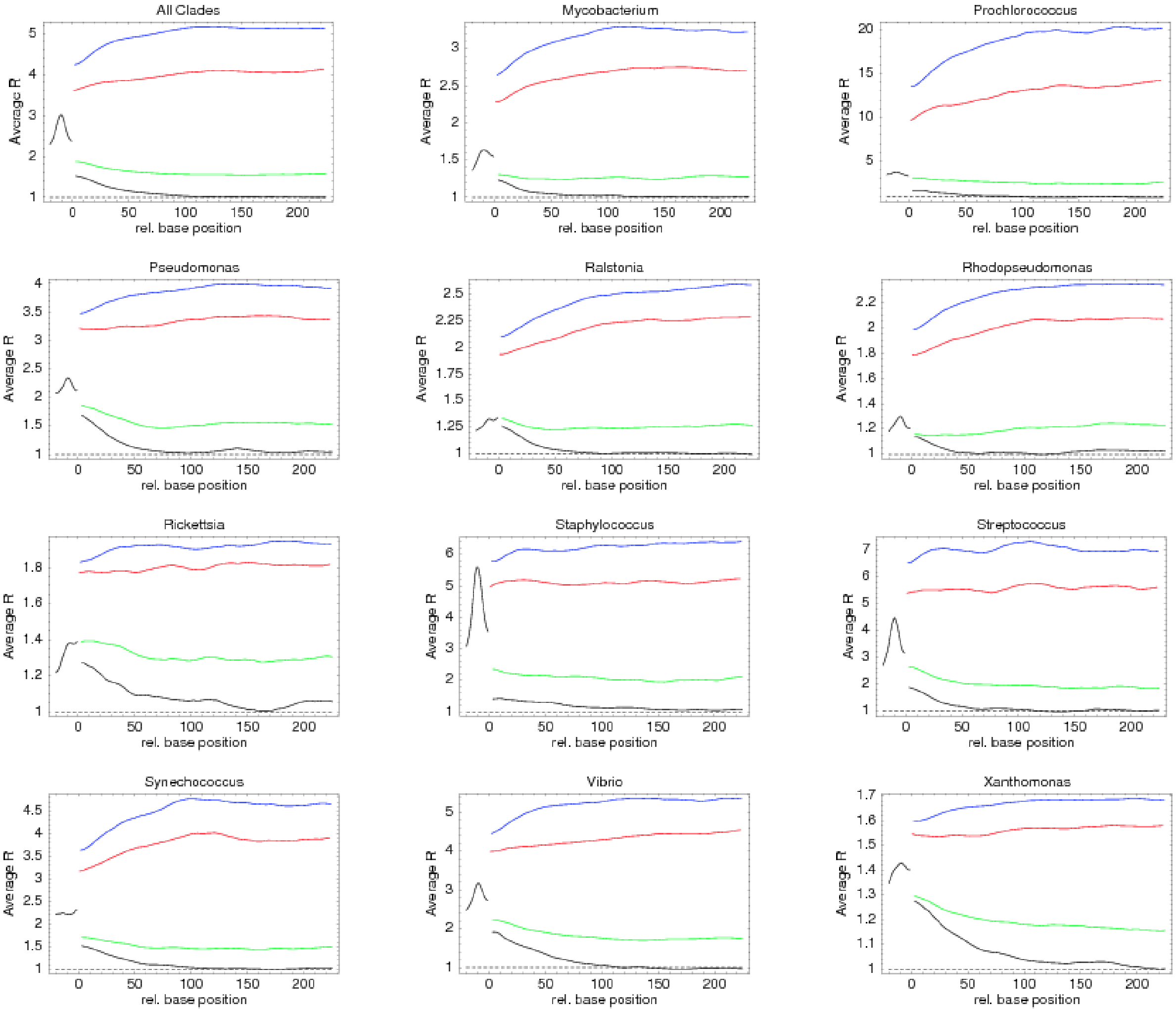


Figure 8: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) of genes that are not the first in their operon in 11 different clades of species, and averaged over all clades. Each panel shows R values in 50 bps upstream regions and the initial 250 bps of genes. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.



All Clades

Mycobacterium

Prochlorococcus

Pseudomonas

Ralstonia

Rhodopseudomonas

Rickettsia

Staphylococcus

Streptococcus

Synechococcus

Vibrio

Xanthomonas

Figure 9: Evidence of selection (average R value) as a function of position with respect to translation start (position 0) of genes that are not the first in their operon in 11 different clades of species, and averaged over all clades. Each panel shows R values in 50 bps upstream regions and the initial 250 bps of genes. Average R values at first (red), second (blue), and third positions of codons within genes are shown, as well as average R values within intergenic regions and at silent positions (black). The dotted horizontal line shows $R = 1$ in each panel, which corresponds to evolution according to the background model.

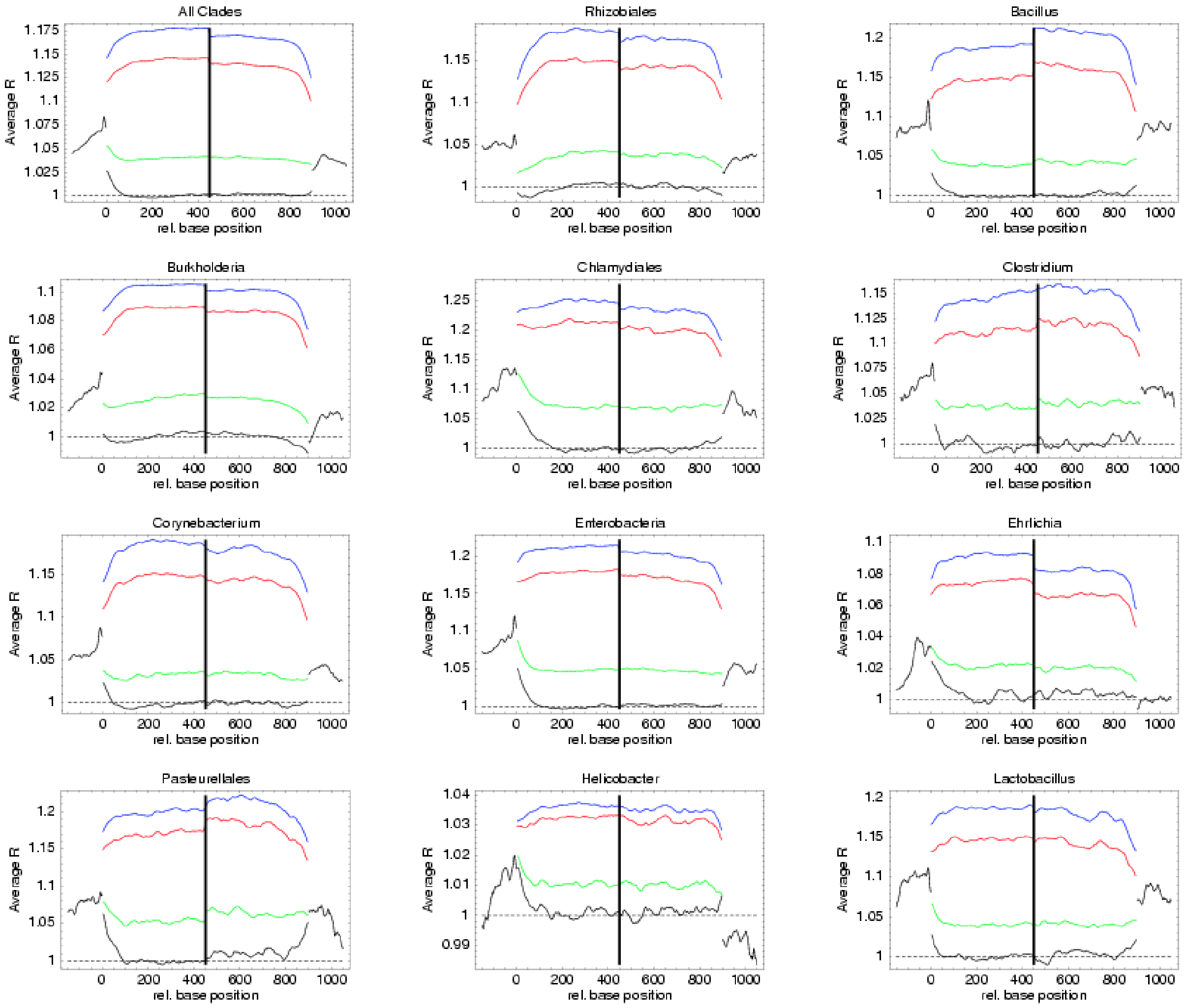


Figure 10: Evidence of selection as measured by reduction in effective substitution rate (Q values, see supporting methods) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows Q values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average values at first (red), second (blue), and third positions of codons within genes are shown, as well as average values within intergenic regions and at silent positions (black). The dashed lines show $Q = 1$, corresponding to the substitution rate expected from the background model.

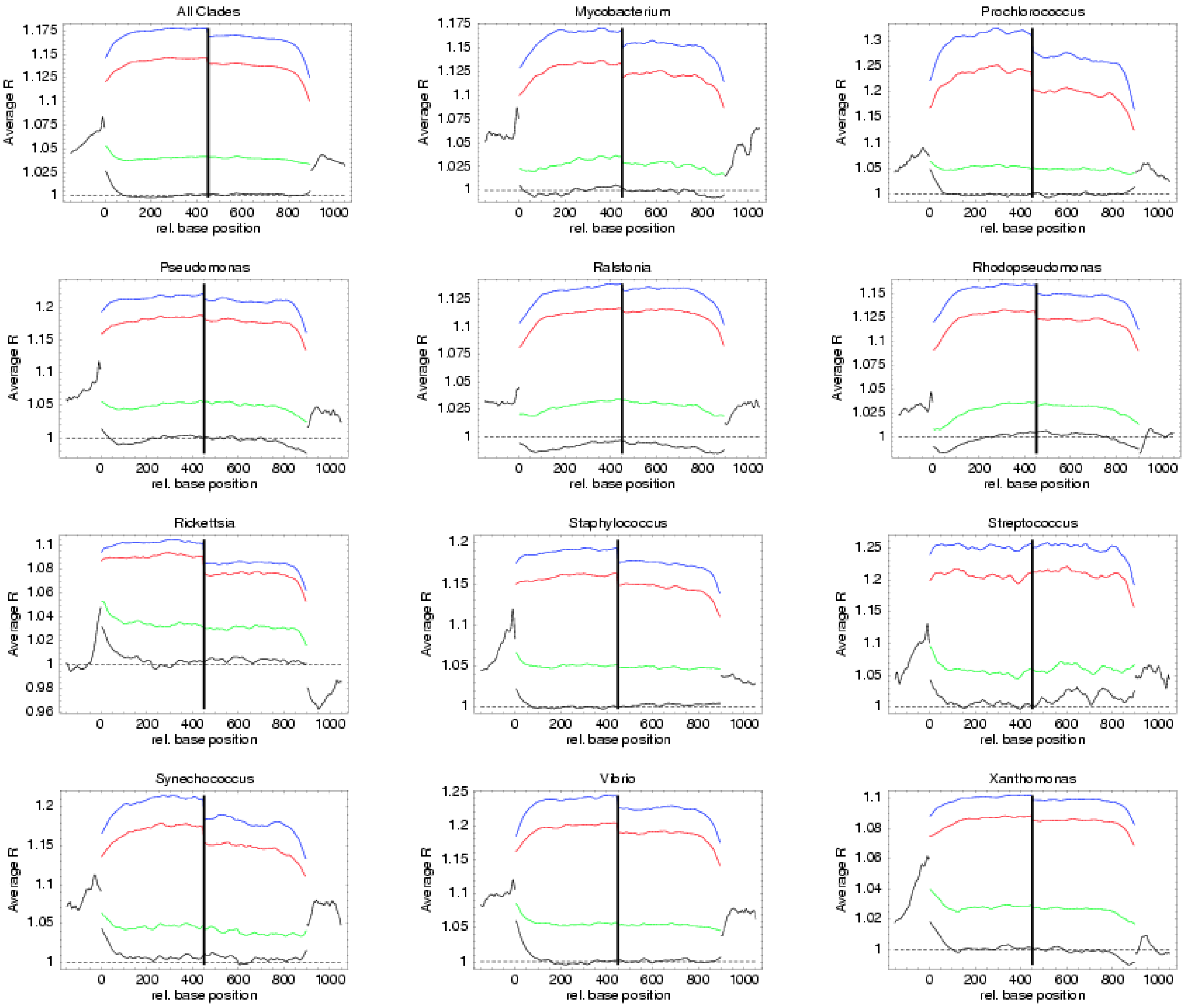


Figure 11: Evidence of selection as measured by reduction in effective substitution rate (Q values, see supporting methods) as a function of position with respect to translation start (position 0) and end (position 900) in 11 different clades of species, and averaged over all clades. The left half of each panel shows Q values in 150 bps upstream regions and the initial 300 bps of genes. The right half shows the last 300 bps of genes plus 150 bps downstream. Average values at first (red), second (blue), and third positions of codons within genes are shown, as well as average values within intergenic regions and at silent positions (black). The dashed lines show $Q = 1$, corresponding to the substitution rate expected from the background model.

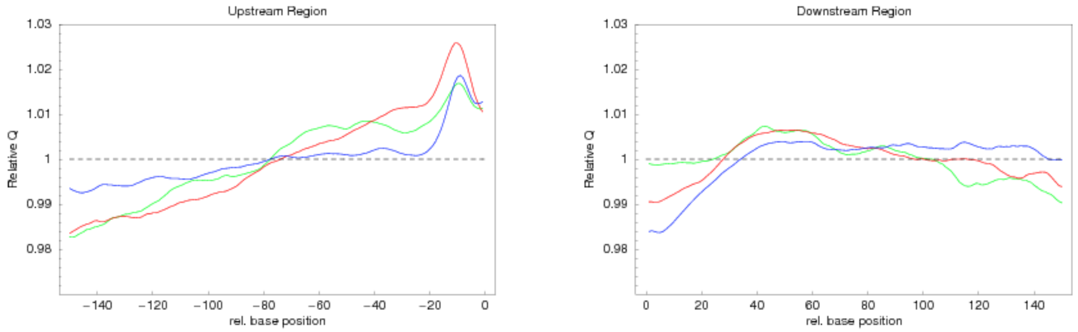


Figure 12: Relative average Q values upstream and downstream of genes averaged separately over genomes with less than 2000 genes (green), genomes with between 2000 and 4500 genes (red), and genomes with more than 4500 genes. In order to compare the shapes of the Q value profiles the values on the vertical axis are scaled to have a mean of 1 when averaged over the 150 bps upstream and when averaged over the 150 bps downstream.

particular, within genes the frequency of A nucleotides is maximal at the start of the gene and decreases over the first 20 nucleotides. G nucleotides have a minimum at the start of the gene and increase over the first 20 nucleotides. The peak in A nucleotide frequency extends into the upstream region. A few bps upstream of translation start a sharp peak in G nucleotides is observed which corresponds to the Shine-Dalgarno sequence. As mentioned, cyanobacteria are the only clades that do not show these patterns. Instead of a peak in the frequency of A nucleotides around translation start the cyanobacteria show a peak in C nucleotides. The cyanobacteria also do not show the peak in G nucleotide frequency immediately upstream of start. These observations suggest cyanobacteria use another mechanism for translation initiation than all other clades. Note that in many clades there seems to be a small but significant minimum in the frequency of A nucleotides between 10 and 20 codons downstream of translation start. We currently have no idea what the meaning or the role of this minimum might be but it seems plausible that it is also related to translation initiation.

In [2] it was shown that, in almost all bacteria DR regions have the highest AT content followed by SR regions, and then NR regions. As demonstrated in Figs. 13 and 14, we addition find that in all clades the AT content upstream of translation start is higher than the AT content downstream of translation start.

10 Selection at silent sites immediately downstream of the start codon

We performed a number of controls to check if the observed elevated selection at silent sites immediately downstream of translation start can be an artefact of another bias. Some of these controls are presented below.

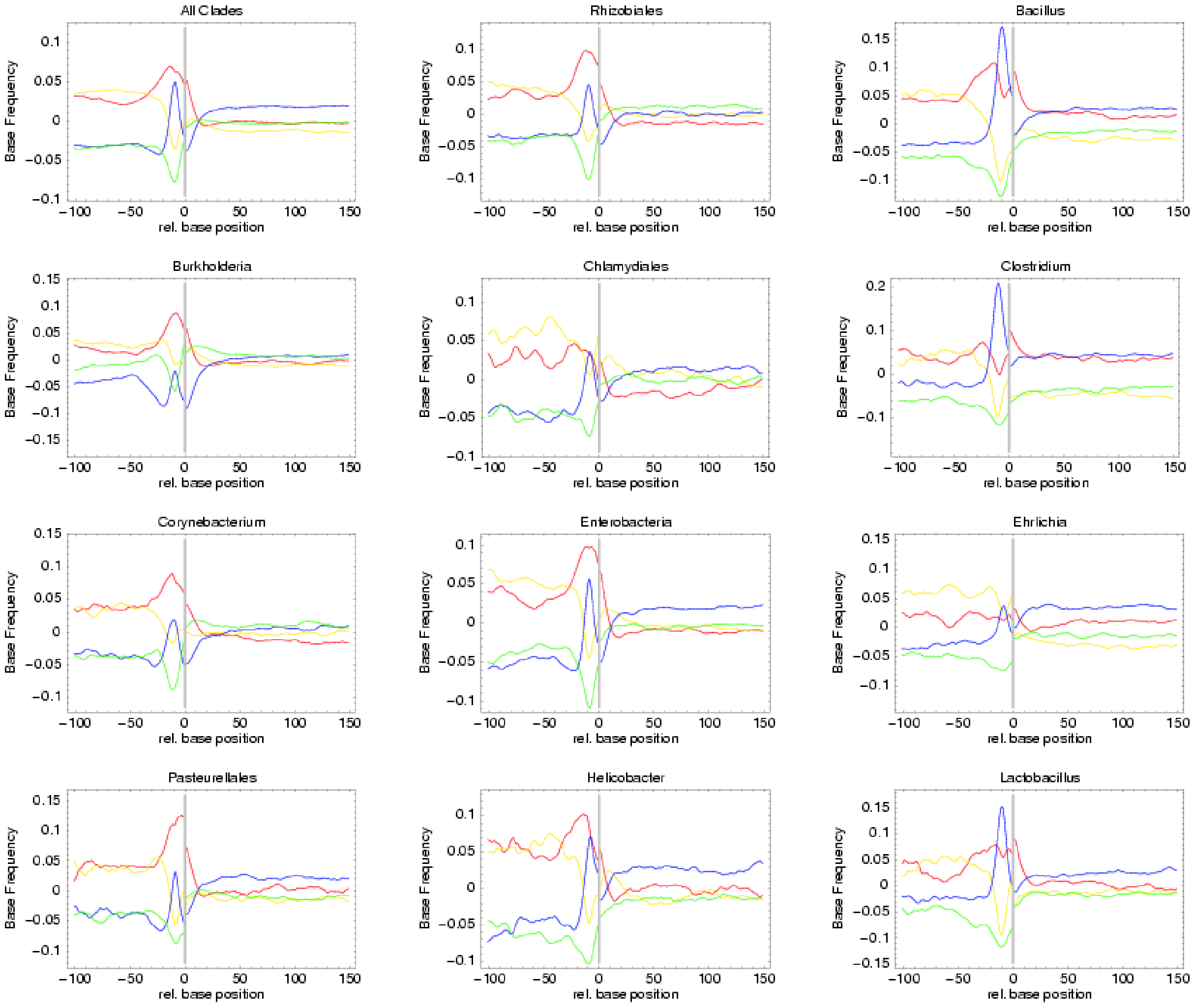


Figure 13: Nucleotide composition profiles from 100 bps upstream of translation start to 150 bps downstream of translation start for 11 different clades and the average profiles over all clades. The vertical axis shows the difference between the frequency of A (red), C (green), G (blue), and T (yellow) nucleotides at each position and the average frequency of the corresponding nucleotides in the entire genome.

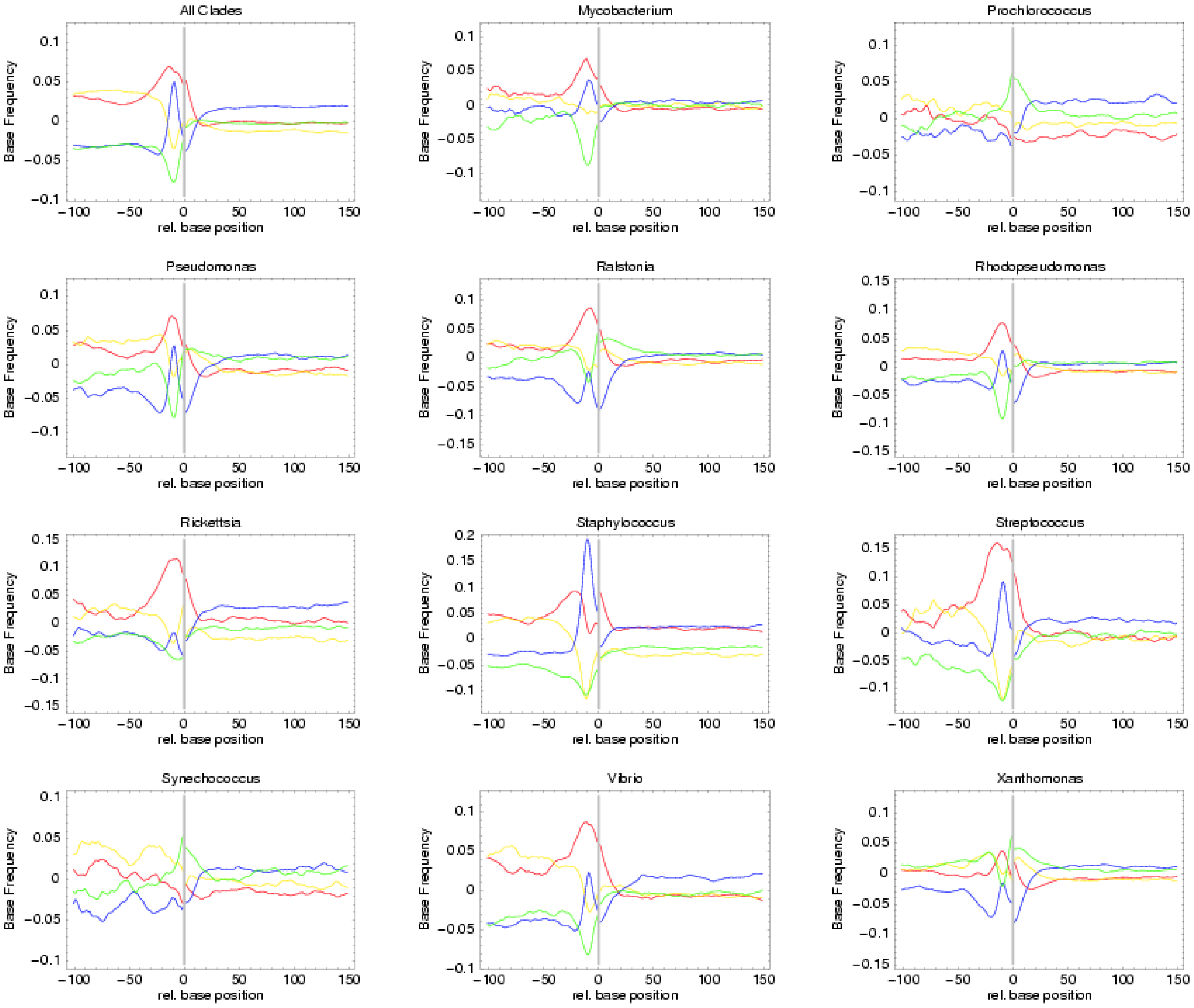


Figure 14: Nucleotide composition profiles from 100 bps upstream of translation start to 150 bps downstream of translation start for 11 different clades and the average profiles over all clades. The vertical axis shows the difference between the frequency of A (red), C (green), G (blue), and T (yellow) nucleotides at each position and the average frequency of the corresponding nucleotides in the entire genome.

10.1 Position-dependent codon adaptation index

One hypothesis for the apparent increase of selection immediately downstream of translation start is that it is caused by an increase in codon bias in this region. For example, highly expressed genes such as ribosomal genes generally show elevated codon bias which is likely the result of selection for translation efficiency. It is conceivable that the initial positions of genes are generally under a stronger selection for efficient translation than positions further downstream in the genes, which would lead to higher codon bias and the elevated selection would be the result of an elevated codon bias only.

To test this hypothesis we have computed position-dependent codon adaptation index $\mathrm{CAI}(d)$ [3] profiles as a function of the position relative to the start of the gene. These profiles are shown in figures 15 and 16. The profiles show that, for almost all clades, the CAI values go *down* rather than up near the start and end of the genes. For the remaining clades an approximately flat CAI profile is observed. These clades have a codon bias that prefers A nucleotides at the third positions of codons such that the elevated frequency of A nucleotides immediately downstream of translation start matches the overall codon bias. In summary, it is clear that the selection immediately downstream of translation start generally reflects a selection for A nucleotides at these positions and not a selection to match the codon bias in the species.

10.2 Reannotation of gene starts

Another hypothesis is that the apparent increase in selection immediately downstream of translation start, and the corresponding lower selection at first and second positions, is an artifact of the incorrect annotation of gene starts in a subset of the genes. That is, if the 'true starts' of a significant fraction of the genes were downstream of the annotated ones, then what we consider to be the initial coding positions of these genes are in fact intergenic positions. Given that the amount of selection at intergenic positions is higher than at silent positions and lower than at coding positions this would produce the pattern of a lowered selection at coding positions and an increase in selection at silent positions, i.e. similar to what we observe.

We implemented a simple procedure, using conservation information, in order to identify gene starts that have potentially been placed too far upstream. First, we search, in the multiple alignment, for an alternative start codon (ATG, GTG or TTG) which is conserved across all species. If such an alternative start exists, we compute the fraction of conserved amino acids $f(i)$ for all columns i of the alignment, the average $\bar{f}_s$ over the positions from the first to the second start codon, and the average $\bar{f}_r$ over the reset of the protein. In this context the 'conservation' fraction $f(i)$ at a position i is the fraction of amino acids in the other species that match the amino acid of the reference species. We then calculate the z-statistic

$$Z = \frac{\bar{f}_r - \bar{f}_s}{\sqrt{\sigma_r^2 + \sigma_s^2}} \tag{2}$$

where σ_r and σ_s are the standard errors of the fraction of conserved amino acids in the region between the first and second codon, and in the rest of the protein respectively. Whenever $Z \geq 5$ and the length of the region between the first and second start is

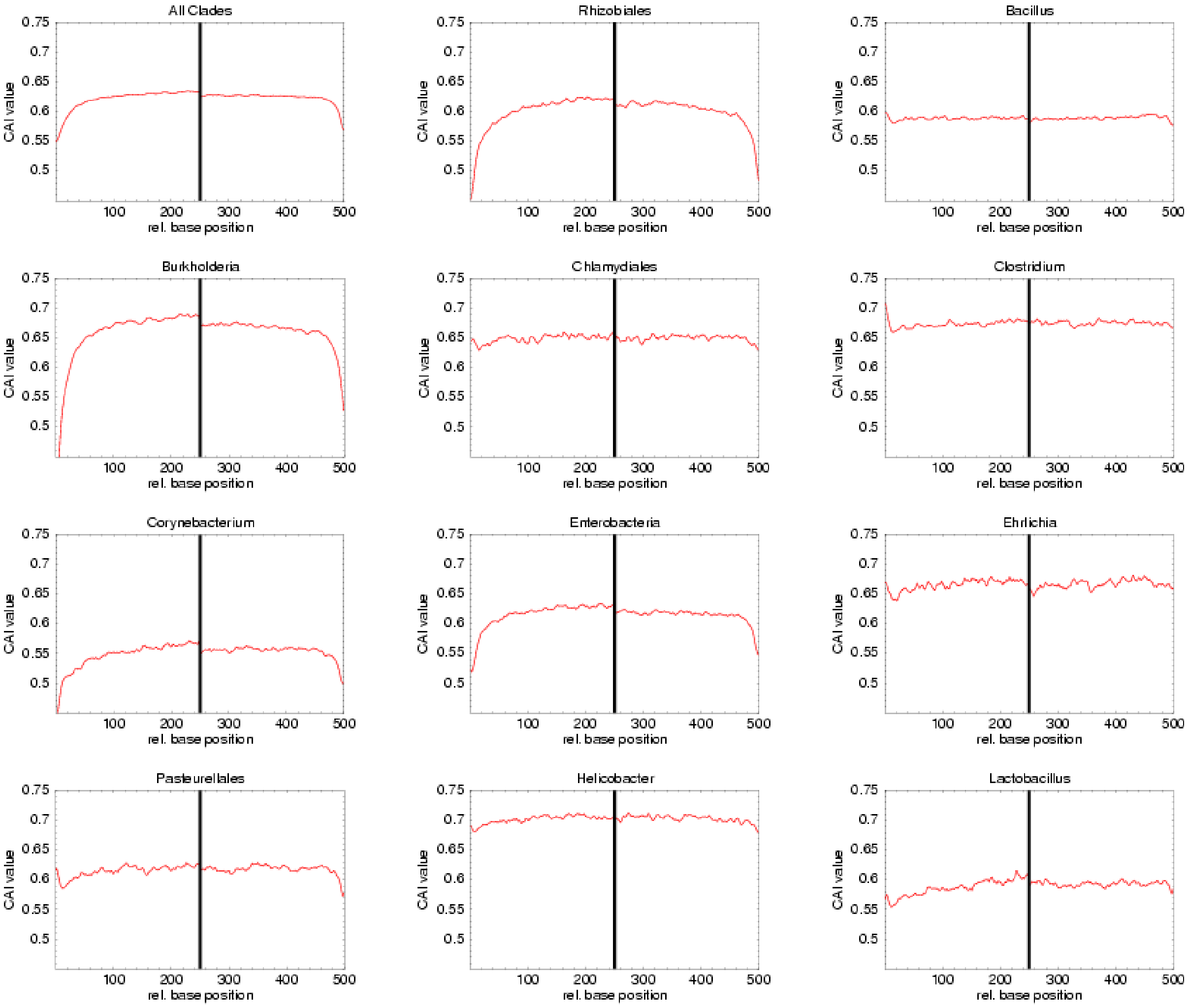


Figure 15: Codon adaptation index (CAI) profiles as a function of position relative to translation start (position 0) and translation end (position 500) for the reference species of 11 clades and averaged over all reference species. Each panel corresponds to one reference species. The left half of each panel corresponds to the first 250 codons downstream of translation start, and the right half to the last 250 codons before the stop codon.

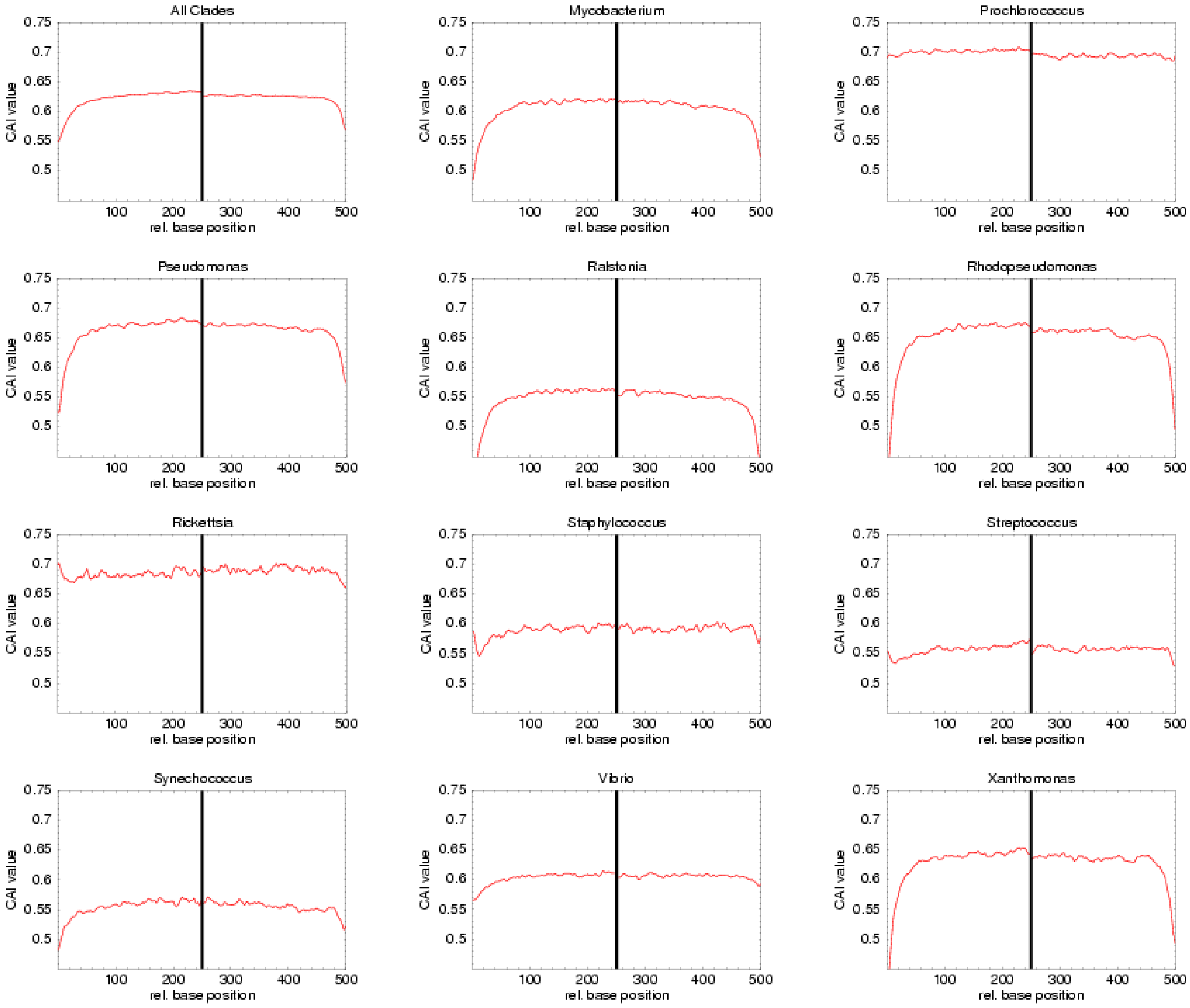


Figure 16: Codon adaptation index (CAI) profiles as a function of position relative to translation start (position 0) and translation end (position 500) for the reference species of 11 clades and averaged over all reference species. Each panel corresponds to one reference species. The left half of each panel corresponds to the first 250 codons downstream of translation start, and the right half to the last 250 codons before the stop codon.

Clade Name	Genes	Starts	Stops	Orths. start	Orths. stop	Reannot.	frac.
Rhizobiales	5469	849	522	2.26	2.16	171	20%
Bacillus	4224	818	492	1.93	2.12	112	14%
Burkholderia	7805	1916	1551	2.04	1.96	429	22%
Chlamydiales	1046	325	224	2.90	2.80	27	8%
Clostridium	3954	196	152	2.88	2.89	14	7%
Corynebacterium	3072	589	432	1.89	1.79	123	21%
Enerobacteria	4400	1127	706	2.19	2.08	169	15%
Ehrlichia	967	368	301	2.54	2.27	64	17%
Pasteurealles	1735	236	100	1.83	1.96	38	16%
Helicobacter	1660	238	136	2.74	2.60	69	29%
Lactobacillus	1938	224	146	1.89	1.88	34	15%
Mycobacterium	4237	479	302	2.29	2.08	134	28%
Prochlorococcus	2324	294	236	2.81	2.52	57	19%
Pseudomonas	5684	454	341	2.90	2.77	70	15%
Ralstonia	6532	1172	735	1.80	1.81	346	30%
Rhodopseudomonas	4958	1230	893	3.02	2.85	359	29%
Rickettsia	877	308	301	3.12	2.78	44	14%
Staphylococcus	2698	1076	1034	2.63	2.50	55	5%
Streptcoccus	2164	126	72	2.62	2.62	13	10%
Synechococcus	2697	289	166	2.08	1.84	66	23%
Vibrio	3958	713	539	3.09	2.90	213	30%
Xanthomonas	4242	1333	1201	2.45	2.37	460	35%

Table 2: Number of regions used in R value profiles, and number of reannotated regions. For each clade the columns show (from left to right): the total number of genes in the reference species, the number of regions around gene starts used for building the R value profiles, the number of regions around gene ends used for building the R value profiles, the average number of orthologs per gene start region, the average number of orthologs per gene end region, the number of reannotated gene starts, and the fraction of gene starts that were reannotated.

less than half of the protein, we reannotate the start of the gene, i.e. move it to the downstream start position.

Table 2 shows a number of general statistics on our clades such as the total number of genes in the reference species, the number of regions that were useed for constructing the R value profiles and the average number of orthologs per region. It also shows the number of gene starts that were reannotated in our reannotation control. The fraction of reannotated gene starts varies from 5% in Staphylococcus to 35% in Xanthomonas.

Figures 17 and 18 show the original R value profiles together with the R value profiles using the reannotated gene starts. As the figure shows, although the reannotation decreases the amount of selection immediately downstream of translation start, and increases the conservation at second and first positions in this region, the changes are small and significant evidence of selection immediately downstream of translation

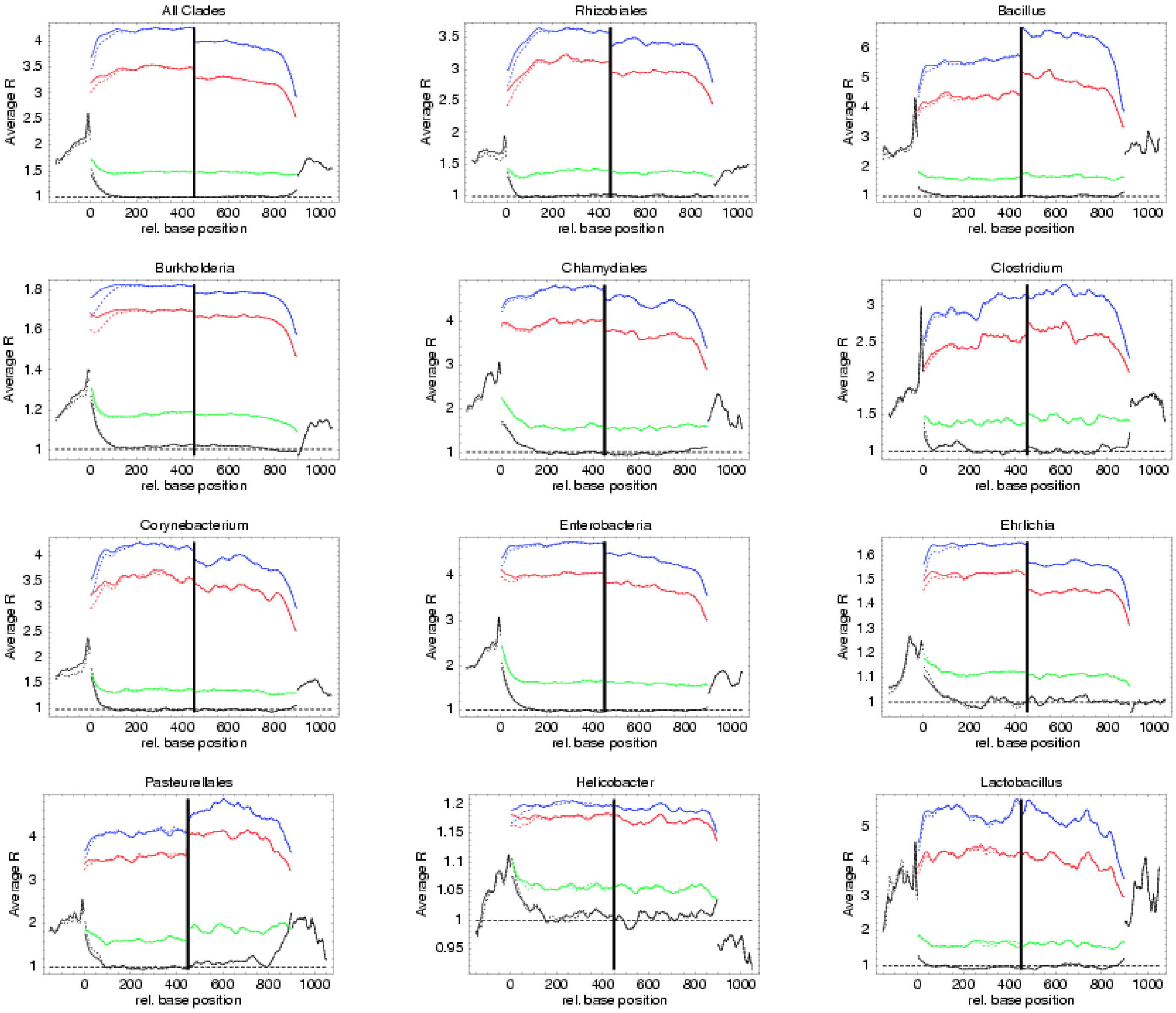


Figure 17: Comparison of R value profile with the original and reannotated gene starts. The R profiles with the original gene starts are shown as dotted lines whereas the R profiles with the reannotated gene starts are shown as solid lines. See the caption of figure 6 for a description of the data shown.

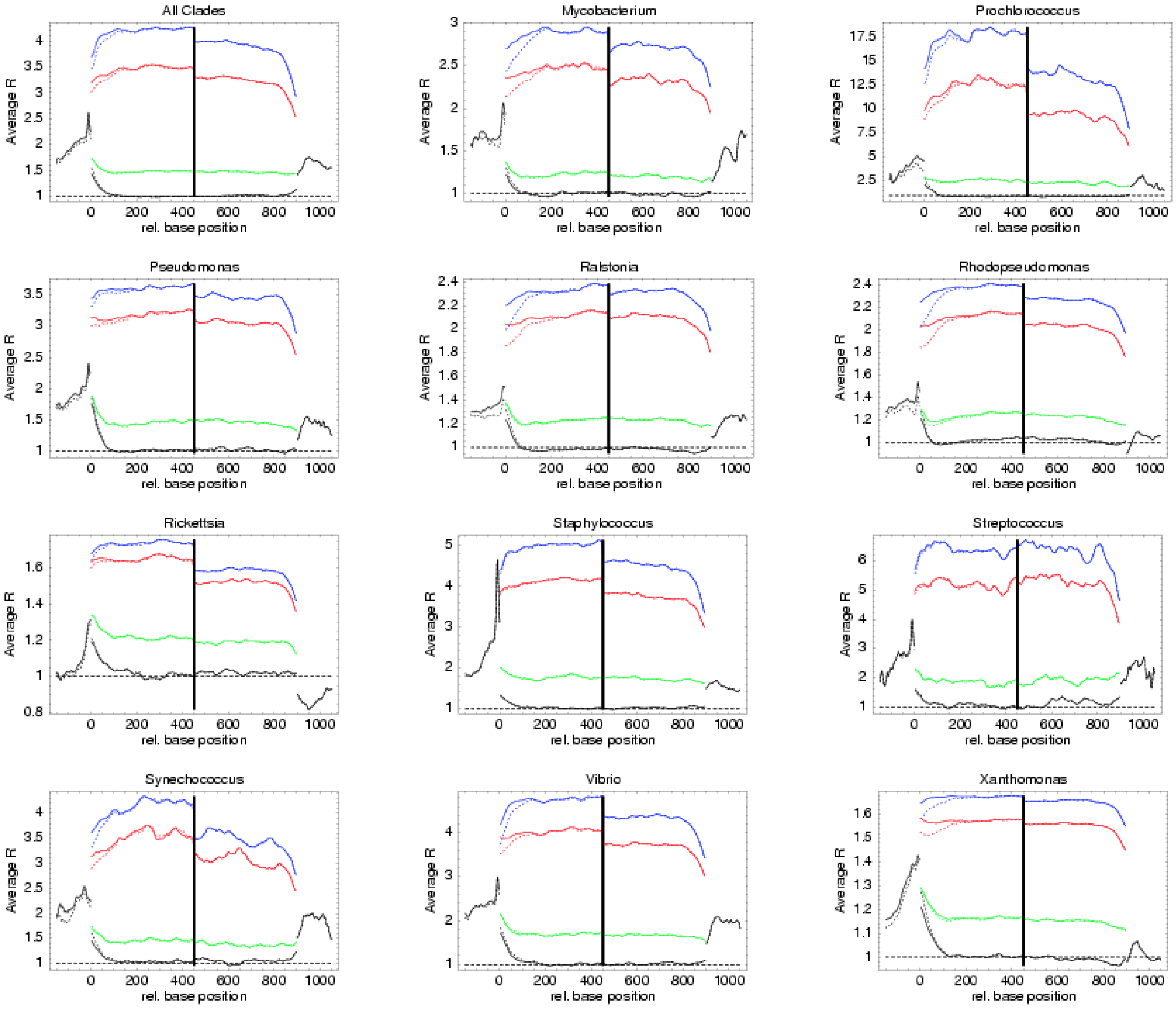


Figure 18: Comparison of R value profile with the original and reannotated gene starts. The R profiles with the original gene starts are shown as dotted lines whereas the R profiles with the reannotated gene starts are shown as solid lines. See the caption of figure 7 for a description of the data shown.

start remains. We also observed that, using the reannotated gene starts, intergenic regions now exhibit more evidence of selection at second and first positions then at third positions (relative to the start codon), which was not the case with the original annotations (data not shown). This suggests that our reannotation has already misclassified a significant number of coding regions as intergenic.

An alternative way of refuting that the selection immediately downstream of translation start is a result of misannotated gene starts is to calculate R values for a set of proteins with well-known amino acid sequences, i.e. with known starts. We built such a set by collecting all E. coli K12 proteins for which the function has been experimentally determined. Again we observed that the R profiles of this set are very similar to the R profiles of all genes. In summary, we believe we can exclude the hypothesis that the observed selection downstream of translation start is an artefact of misannotated gene starts.

10.3 Shine-Dalgarno peak and downstream selection signal

If the avoidance of secondary structure around gene starts were related to transcription initiation we would expect to observe this pattern only in genes that are the first in their operon. In the left panel of Fig. 19 we compare the selection at the first 20 silent positions immediately downstream of ATG (the 'downstream signal') in genes with small and large upstream regions. Although the downstream signal is often largest in genes with large upstream regions there is clear evidence of downstream signal in genes with small upstream regions, which in some cases is even larger than in genes with large upstream regions.

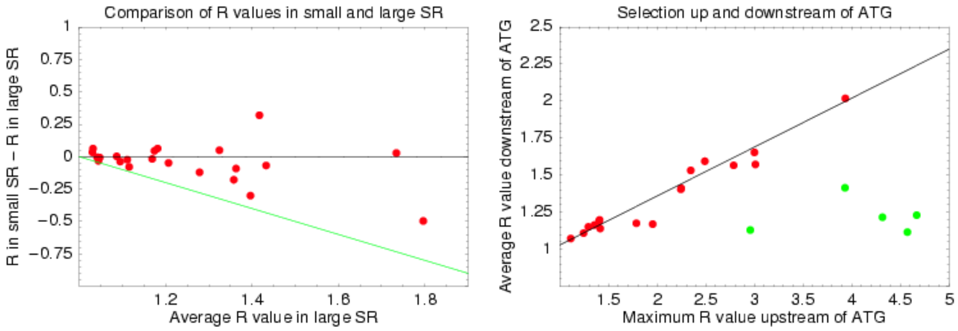


Figure 19: Left Panel: Difference of the 'downstream signal' (average value of R in the first 20 silent positions downstream of translation start) between genes with small ($<$ 50 bp) upstream regions and genes with large ($>$ 150 bp) upstream regions as a function of the downstream signal in genes with large upstream regions. The green line corresponds to a value of $R = 1$ in genes with small upstream regions. Right panel: The downstream signal (vertical axis) as a function of the height (in R value) of the peak corresponding to the Shine-Dalgarno signal. The green dots correspond to firmicutes clades. Each dot corresponds to one of the 22 clades in both panels.

If the downstream signal is associated with translation initiation we might expect a correlation of this signal with the strength of selection at the Shine-Dalgarno sequences. As shown in the right panel of Fig. 19, there is in general a linear correlation between the height of the Shine-Dalgarno peak and the downstream signal. Interestingly, the firmicutes clades (green dots) deviate from this pattern and show relatively little downstream signal and very strongly conserved Shine-Dalgarno sequences. The results in Fig. 19 strongly suggest that the avoidance of secondary structure around translation start is the result of a selection pressure for ensuring efficient translation initiation.

11 RNA secondary structure profiles

Figures 20 and 21 show position dependent z-statistics for the average probability of bases at that position to be unpaired in the RNA secondary structure of the mRNA around translation start, both compared to the average probability of of being unpaired in the flanking regions $(-50, -31)$ and $(31, 80)$, and compared to the average probability of being unpaired in random sequences with the same position-dependent base composition (see supplementary methods).

We see that for all clades there are peaks in ‘openess’ immediately upstream and downstream of translation start compared to the flanking regions more to the left and right (red curves). Note that the G nucleotides of the Shine-Dalgarno sequence and the start codon itself tend to lead to minima in openess at these positions. In addition, the blue curves show that the region immediately around translation start shows even more ‘openess’ then random sequences with the exact same base composition. This strongly suggests that base composition in these regions is the result of a selection for avoiding secondary structure in essentially all clades.

11.1 Z-values for the region immediately around translation start

We calculated z-values for the average openess in the region $(-20, 20)$ immediately around translation start, compared with the average openess in the flanking regions (regions $(-50, -31)$ and $(31, 80)$) and z-values for the average openess in the region $(-20, 20)$ compared with the average openess of the same region in random sequences with the same position-dependent base composition. The results are shown in figure 22.

The figure shows that in all clades there is significantly more openess, i.e. $z > 2$, in the region immediately around translation start than in the flanking regions. In addition, for all but two clades (Mycobacterium and Synechococcus) there is significantly more openess in the region $(-20, 20)$ then in random sequences with the same position-dependent base composition.

11.2 5’ UTR lengths in E. coli

For folding the region around translation start we assumed that transcription start occurs 60 bp upstream of translation start, i.e. we include 60 bp upstream of translation

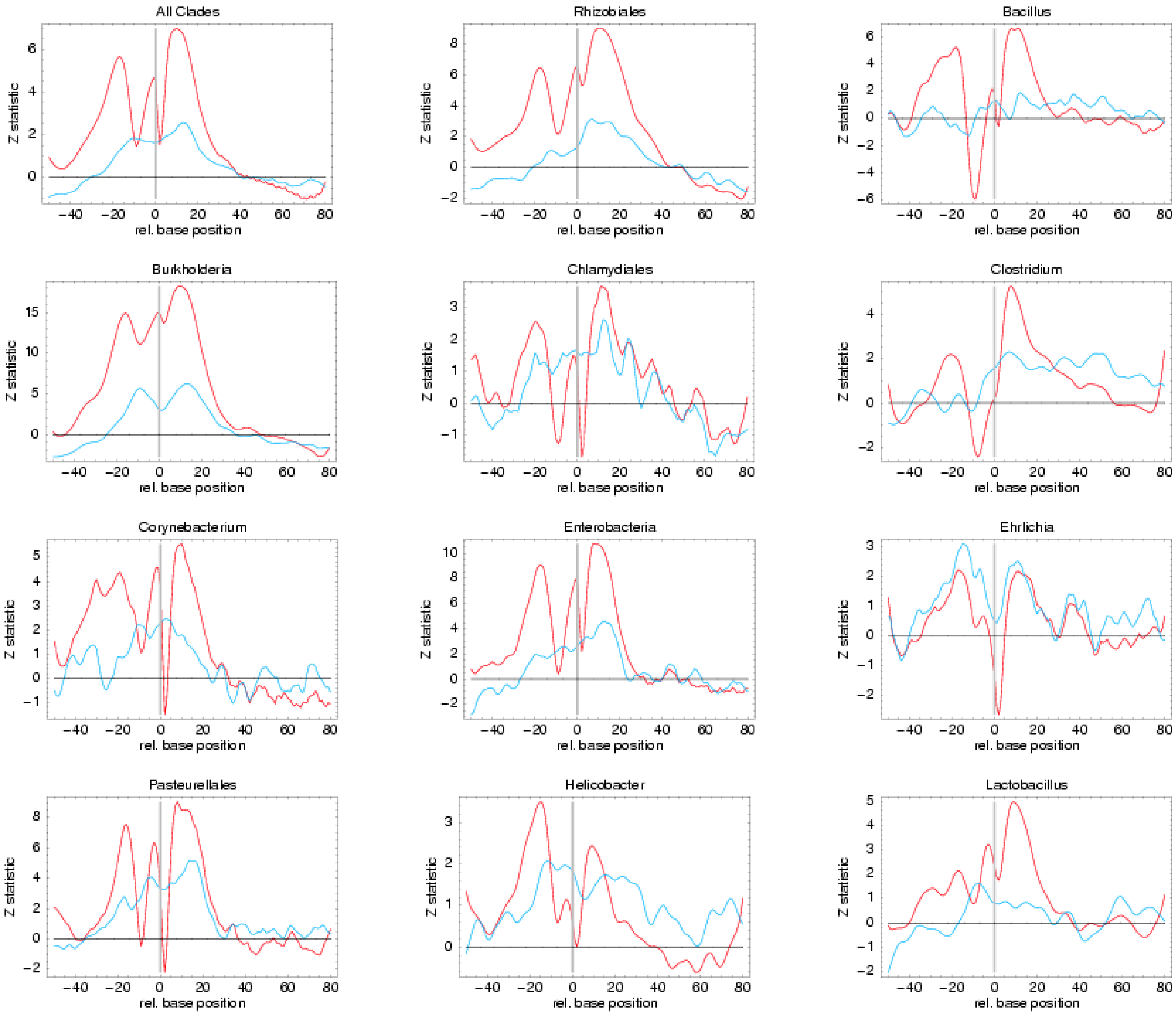


Figure 20: RNA secondary structure profiles for 11 clades and averaged over all clades. The horizontal axis in each panel shows the position relative to translation start, from 50 bp upstream to 80 bp downstream. The vertical axes show two z-statistics for the probability of the nucleotide at that position to be *unpaired*. The red lines show the z-statistic of the probability for the position to be unpaired relative to the average probability over the flanking segments $(-50, -31)$ and $(31, 80)$. The blue lines show the z-statistics for the position to be unpaired relative to the average probability of the same position being unpaired in random sequences with the same position-dependent base composition as observed in the clade.

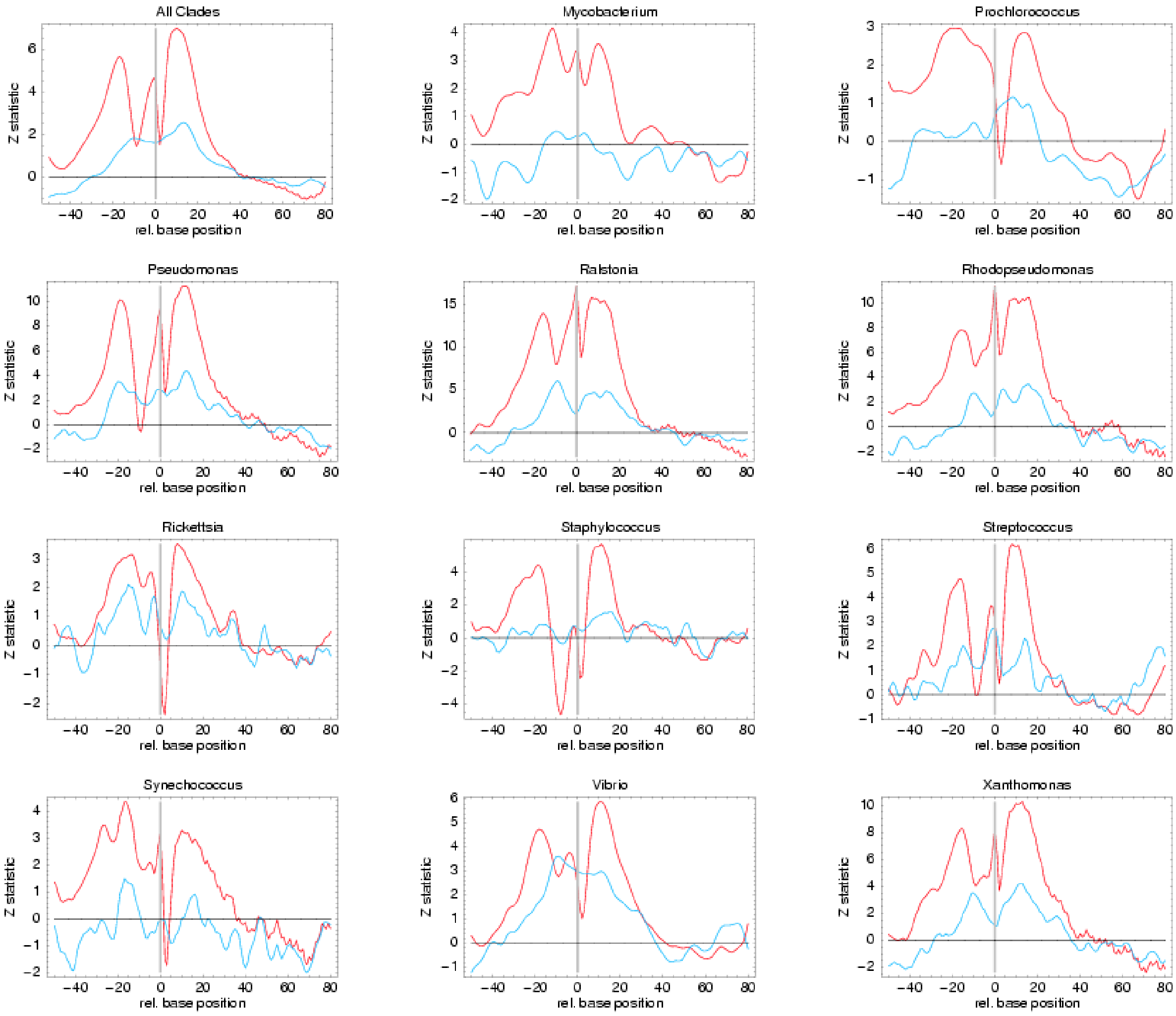


Figure 21: RNA secondary structure profiles for 11 clades and averaged over all clades. The horizontal axis in each panel shows the position relative to translation start, from 50 bp upstream to 80 bp downstream. The vertical axes show two z-statistics for the probability of the nucleotide at that position to be *unpaired*. The red lines show the z-statistic of the probability for the position to be unpaired relative to the average probability over the flanking segments $(-50, -31)$ and $(31, 80)$. The blue lines show the z-statistics for the position to be unpaired relative to the average probability of the same position being unpaired in random sequences with the same position-dependent base composition as observed in the clade.

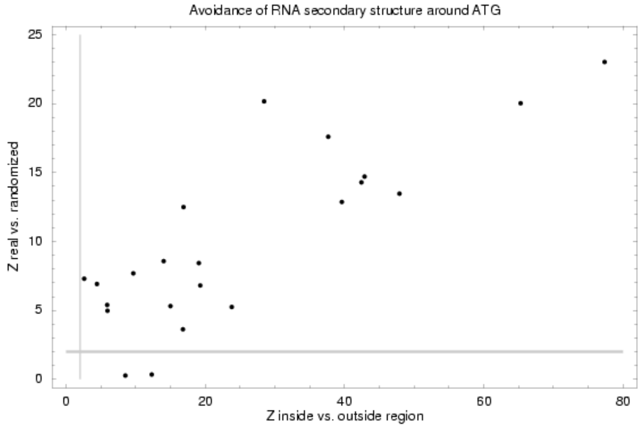


Figure 22: Z-statistics of the RNA secondary structure for the region $(-20, 20)$ immediately around translation start. Each dot in the plot corresponds to one clade. On the horizontal axis is the z-statistic of the average openess in the region $(-20, 20)$ compared to the average openess in the flanking regions $(-50, -31)$ and $(31, 80)$. On the vertical axis is the z-statistic of the average openess in the region $(-20, 20)$ compared to the average openess in random sequences with the same position-dependent base composition. The grey lines show the value $z = 2$.

start in the sequence to be folded. The estimate of 60 bp is based on analysis of the distribution of 5' UTR lengths in E. coli.

RegulonDB [4] contains a collection of experimentally determined transcription start sites in E. coli. For each TSS we calculated the distance to the start of the downstream ORF and determined the distribution of 5' UTR lengths. Fig 23 shows the distribution of 5' UTR lengths that we observed. Note that the majority of 5' UTRs is less than 50 bp long but that there is a fairly long tail including apparent very long 5' UTRs. Some of these very long 5' UTRs are possibly due to misidentification of the downstream gene or other sources of error. If one excludes all 5' UTRs longer than 150 bps the average length is 60 bp. If one excludes 5' UTRs longer than 250 bp the average length is 77 bps.

12 Clusters of TF DNA binding domains

Figure 24 shows the number of different clusters of paralogous transcription factors as a function of the total number of genes in the genome at different similarity cut-offs.

The figure shows that, at all similarity cut-offs the function can be reasonably well-fitted by a power law. The fitted intersepts and exponents are

- Intercept -9.656, Exponent 1.845 at a cut-off of 100% similarity.
- Intercept -9.699, Exponent 1.847 at a cut-off of 85% similarity.
- Intercept -9.568, Exponent 1.827 at a cut-off of 65% similarity.

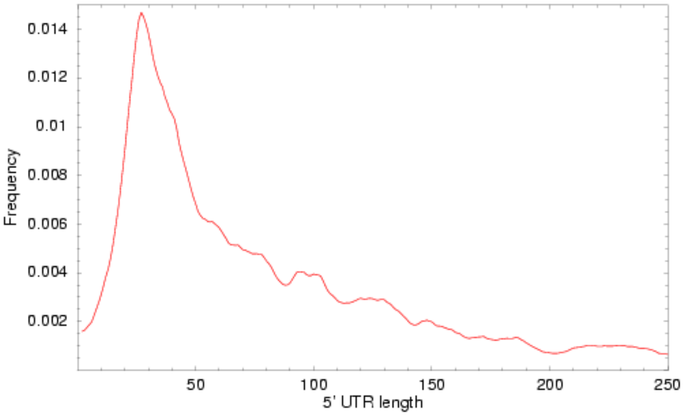


Figure 23: Distribution of 5’ UTR lengths in E. coli as estimated from the collection of transcription start sites in RegulonDB [4]. The horizontal axis shows the length of the 5’ UTR and the vertical axis shows the the frequency of 5’ UTRs of the corresponding length. The distribution was smoothed with an exponential kernel (see supporting methods).

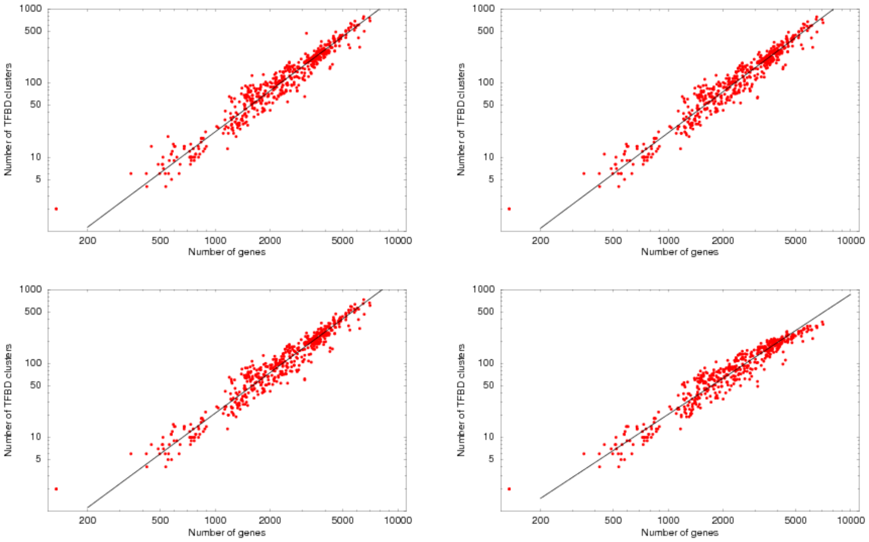


Figure 24: Number of clusters of transcription factors with similar DNA binding domains at cut-offs of 100% amino acid identity (top-left), 85% amino acid identity (top-right), 65% amino acid identify (bottom-left) and 45% amino acid identity (bottom-right) as a function of the total number of genes in the genome. Both axes are shown on logarithmic scales. The black lines are power-law fits.

- Intercept -8.209, Exponent 1.625 at a cut-off of 45% similarity.

We thus find that at all three higher cut-offs there is very little evidence of TF clustering, and the amount of clustering does not increase with genome size. At 45% identity there is some clustering but the exponent is still as high as 1.6 (i.e. far from linear) and at this low similarity there is little guarantee that the TFs will bind similar motifs.

13 Word Count Ratios

For each clade we produced a list of all 4^7 sevenmers ordered by the evidence for each of the sevenmers to be under purifying selection. We then determined the number of unique sevenmers n_t from the top of the list that together account for 5% of all sequence segments of length 7 in intergenic regions, and the number of unique segments n_b from the bottom of the list that together account for 5% of all windows in intergenic regions. Figure 7 in the main paper showed the ratio n_t/n_b as a function of the total number of TFs in the genome.

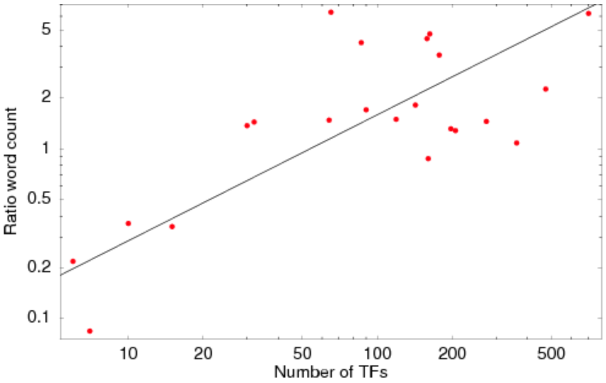


Figure 25: Sequence diversity of the most and least conserved sevenmers as a function of the number of TFs in the genome. For each genome we ordered all sevenmers by their evidence for being under purifying selection and collected the most and least conserved unique sevenmers such that the sevenmers of both sets each account for 5% of all sequence segments in the genome. The vertical axis shows the ratio between the number of most conserved and least conserved sevenmers in the corresponding set as a function of the total number of TFs in the genome (horizontal axis). Both axes are shown on logarithmic scale. The black line shows a linear fit.

Here figures 25 and 26 show the word count ratios that are obtained when we collect sevenmers to account for 10% (Fig. 25) or 20% (Fig. 26) of all windows of length 7 in the intergenic regions. The figure shows that also with these larger numbers of sevenmers the word-count ratio increases significantly with the number of TFs in the genome.

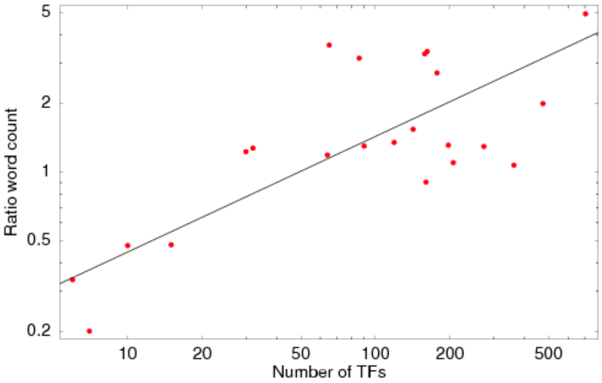


Figure 26: Sequence diversity of the most and least conserved sevenmers as a function of the number of TFs in the genome. For each genome we ordered all sevenmers by their evidence for being under purifying selection and collected the most and least conserved unique sevenmers such that the sevenmers of both sets each account for 5% of all sequence segments in the genome. The vertical axis shows the ratio between the number of most conserved and least conserved sevenmers in the corresponding set as a function of the total number of TFs in the genome (horizontal axis). Both axes are shown on logarithmic scale. The black line shows a linear fit.

14 Methods

14.1 Power-law fitting

In several places we fit a power-law to a scatter of points, i.e. the number of operons as a function of the number of genes, the number of TFs as a function of the number of genes, and the number of TF clusters as a function of the number of genes. To perform this fit we first log-transform all the data points. Let (x_i, y_i) denote the log-transformed data points. The Bayesian straight-line fitting assumes Gaussian noise of unknown size in both horizontal and vertical components and assumes a rotationally invariant prior for the slope of the line [5]. In this model the posterior probability $P(a|D)$ that the slope of the line is a given the data is given by [6]

$$P(a|D) \propto \frac{\left(a^2+1\right)^{(n-3)/2}}{\left(\mathrm{var}(y) + a^2\mathrm{var}(x) - 2a\mathrm{cov}(x,y)\right)^{(n-1)/2}}, \tag{3}$$

where n is the number of data points, $\mathrm{var}(x)$ is the variance of the x-values

$$\mathrm{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x\rangle)^2, \tag{4}$$

$\mathrm{var}(y)$ is the variance of the y-values

$$\mathrm{var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y\rangle)^2, \tag{5}$$

and $\mathrm{cov}(x, y)$ is the covariance

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \langle x\rangle\langle y\rangle). \tag{6}$$

Note that the optimal line in this procedure corresponds roughly to the line that minimizes the sum of the squared orthogonal distances of the data points to the line, which also corresponds to the first principal component of the data.

14.2 Mapping Orthologs

First we collect the list of all (predicted) protein sequences for each genome from the corresponding genbank file. A list of putative orthologs for each pair of genomes is obtained by running WU-BLAST [7]. As shown in [8], ortholog identification becomes more accurate if evolutionary distances, estimated by maximum likelihood, are used instead of BLAST scores. Thus, for each reported hit, we globally align the corresponding pair of proteins using CLUSTALW [9]. To avoid mistaking single domain matches for orthologs we only retain alignments that cover at least 50% of both proteins. We estimate the evolutionary distance d of the pair using PAML [10] and assign a score $H = -\log(d)$ to the pair.

We number all the proteins in both genomes according to their position on the chromosome and then identify orthologs by the following iterative procedure:

1. A pair of genes (α, β) are considered orthologs, which we write as $\alpha \backsim \beta$, if they are best reciprocal hits, and there is no other hit with a score larger than a fraction f of the score of the pair. That is $\alpha \backsim \beta$ if $H_{\alpha j} < fH_{\alpha\beta}$ for all $j \neq \beta$ and $H_{i\beta} < fH_{\alpha\beta}$ for all $i \neq \alpha$. We search for all pairs satisfying these conditions. After that, for each identified orthologous pair $\alpha \backsim \beta$ we set all scores $H_{\alpha j}$ and $H_{i\beta}$, i.e. hits to other proteins, to zero. We then repeat the search for orthologs until no more new orthologs are found.

2. We construct diagonals of consecutive or "anti-consecutive" pairs, i.e. runs of syntenic orthologous pairs of the form $\{\alpha \backsim \beta,\, (\alpha{+}1) \backsim (\beta{+}1), \ldots, (\alpha{+}n) \backsim (\beta + n)\}$ or $\{\alpha \backsim \beta,\, (\alpha + 1) \backsim (\beta - 1), ..., (\alpha + n) \backsim (\beta - n)\}$.

3. We now collect the set of pairs of proteins (i, j) that lie at the start or end of any of the syntenic runs of orthologs. Note that this includes all pairs of genes that lie in "gaps" between consecutive syntenic runs. We then perform the ortholog search on only this subset of pairs.

4. When no more orthologs are found we identify the remaining set of best reciprocal pairs, i.e. no longer demanding that all other scores are less than a fraction f of the best reciprocal pair score.

In the paper we used $f = 0.5$, i.e. the score of the best pair should be twice as high as the next best pair. We find that, even for the sets of genomes of relatively closely related species that we work with, this procedure increases the number of orthologous pairs found by 10% or more over just using best reciprocal hits. However, the results shown in the paper are not affected if a simple best-reciprocal-hit procedure is used for identifying orthologs instead of our procedure.

14.3 Finding Orthologous Cliques

Having determined all the pairwise orthology relations for all pairs of genomes in a clade, it is straitforward to find orthologues cliques. An orthologous clique is a set of genes, one from each organism, such that all genes in the set are orthologs of each other. First, we assign to each gene an n-dimensional vector (with n the number of genomes in the clade) where the ith entry in the vector is the identity of the ortholog in the ith genome (if i is the genome from which the gene itself stems, then the entry is the identity of the gene itself). For each genome we produce a list of such vectors. Cliques are identified as those vectors that occur in the list of vectors of all genomes.

14.4 Removing the least and most conserved cliques

We align the DNA sequences of all orthologous cliques using T-coffee [11]. For each alignment we identify all third positions in codons of the sequence of the reference species and check what fraction of the aligned bases from the other species in the clade is conserved. In this way a conservation statistic is assigned to each multiple alignment. For each clade we sort all multiple alignments by this conservation statistic and remove the top 10% and bottom 10% of the multiple alignments. These ‘outliers’ will not be used for our parameter estimation.

14.5 Reconstructing the phylogenetic tree topology

For each clade we concatenated the remaining 80% of the multiple alignments and let the TREE-puzzle algorithm [12] determine a phylogenetic tree from this concatenated alignment.

14.6 Determining base composition and Codon Bias

We use 12 different background models, for 12 different classes of positions: intergenic positions, first, second, and third positions in codons, and silent positions in each of 8 fourfold degenerate codons. For each background model c we need to determine the vector of equilibrium frequencies w^c, with w^c_α the frequency of base α in class c. To estimate the equilibrium frequencies we average over all organisms in the clade. Base frequencies in intergenic regions are determined from all intergenic regions in all genomes in the clade. For the coding positions and the silent positions for the 8 fourfold degenerate codons we used all the remaining orthologous cliques, but have excluded the first and last 20 amino acids in each clique. The latter is done because, as our results show, there are significant deviations in base composition at the starts and ends of genes.

14.7 Estimating pairwise species distances

For each clade, and each pair of species in this clade, we start by collecting the data-set D of all pairwise aligned third positions in fourfold degenerate codons from the filtered set of cliques (excluding the first and last 20 amino acids in each protein). We use only those third positions for which the amino acid is conserved and count the number of times $n^c_{\alpha\beta}$ that base α occurs in the first species and base β in the other in codons of type c. Further let w^c_α and $\tilde{w}^c_\alpha$ denote the frequency of base α in codons of type c in the first species and second species respectively. We will approximate the probability to observe the pair of bases $\alpha\beta$ at a codon of type c by the average of the probabilities (under the F81 model) to start with base α in the first species and evolve to base β in the second and the probability to start with base β in the second species and evolve base α in the first. That is, the probability $P(\alpha|\beta, t, w^c)$ that the third position of a codon of type c will evolve from base β in the second genome to base α in the first genome assuming distance t between the genomes, is given by

$$P(\alpha|\beta, t, w^c) = \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) w^c_\alpha \tag{7}$$

where w^c_α is the fraction of all codons of type c in the first genome that have base α at the third position. Analogously, the probability to evolve from α in the first genome to β in the second genome is given by

$$P(\beta|\alpha, t, \tilde{w}^c) = \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) \tilde{w}^c_\beta. \tag{8}$$

We now approximate the probability to find the pair of bases $\alpha\beta$ at the third positions of a codon of type c in the alignment of two orthologous proteins from the two genomes as the average of $P(\alpha|\beta, t, w^c)\tilde{w}^c_\beta$ and $P(\beta|\alpha, t, \tilde{w}^c)w^c_\alpha$:

$$P(\alpha\beta|t, w^c, \tilde{w}^c) = \frac{(w^c_\alpha + \tilde{w}^c_\beta)}{2} \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) w^c_\alpha \tilde{w}^c_\beta. \tag{9}$$

Using this expression, the probability $P(D|t, w, \tilde{w})$ of the observed dataset D of counts $n^c_{\alpha\beta}$ is then given by

$$P(D|t, w, \tilde{w}) = \prod_{c,\alpha,\beta} \left[\frac{(w^c_\alpha + \tilde{w}^c_\alpha)}{2} \delta_{\alpha\beta} e^{-t} + (1 - e^{-t}) w^c_\alpha \tilde{w}^c_\beta \right]^{n^c_{\alpha\beta}}, \tag{10}$$

where the product is over all 8 fourfold degenerate codons c and 16 base combinations $\alpha\beta$. We determine the distance t of the pair of species by maximizing this expression with respect to t. We take the derivative of the logarithm of the expression (10) and set the result equal to zero. This leads to the following algebraic equation

$$N^{\text{diff}} = \sum_c \sum_\alpha \frac{N^{\text{cons}}_{c,\alpha} (W^c_\alpha - w^c_\alpha \tilde{w}^c_\alpha) e^{-t}}{e^{-t} W^c_\alpha + (1 - e^{-t}) w^c_\alpha \tilde{w}^c_\alpha} \tag{11}$$

where $N^{\text{cons}}_{c,\alpha}$ is the number of occurrences of a conserved pair $\alpha\alpha$ among codons of type c. N^{diff} is the total number of pairs with different bases in the two species, and

$$W^c_\alpha = \frac{w^c_\alpha + \tilde{w}^c_\alpha}{2}. \tag{12}$$

The equation above can be solved by standard numerical techniques since the expression on the right is a monotonically increasing function of t on the positive real axis.

14.8 Fitting the tree from the pairwise distances

As described above, we have already determined the topology of the phylogenetic tree for each clade. In addition, we have determined all pairwise distances t_{ij} between each pair of species (ij) for each clade. To determine the distance t_b on each of the branches b in each phylogenetic tree we use the standard least-square fitting with a fixed tree topology [13]. For completeness we describe the procedure here.

We find the set of distances t_b such that, for all pairs ij, the distances t_{ij} are best approximated by the total distance along the branches connecting i and j. That is, we minimize

$$F = \sum_{ij} \left(t_{ij} - \sum_{b \in \Pi_{ij}} t_b \right)^2 \tag{13}$$

where Π_{ij} is defined as the set of branches connecting nodes i and j of the tree. Taking the derivative with respect to t_b and setting it zero we obtain

$$0 = \sum_{ij} \delta(b \in \Pi_{ij}) \left[t_{ij} - \sum_{b' \in \Pi_{ij}} t_{b'} \right], \tag{14}$$

where $\delta(b \in \Pi_{ij})$ is 1 if branch b is an element of Π_{ij} and 0 otherwise. If we define the vector $\sum_{ij} \delta(b \in \Pi_{ij}) t_{ij} = A_b$, and the matrix $V_{bb'} = \sum_{ij} \delta(b \in \Pi_{ij}) \delta(b' \in \Pi_{ij})$, then equation (13) becomes

$$0 = A_b - \sum_{b'} V_{bb'} d_{b'} \Leftrightarrow d_b = \sum_{b'} \left(V^{-1}\right)_{bb'} A_{b'}. \tag{15}$$

Thus, the optimal set of branch lengths can be determined by a simple matrix inversion. Notice also that A_b is the sum of all pairwise distances between species that are connected through b and $V_{bb'}$ is the number of pairs of species ij for which the path that connects them passes through both b and b'.

14.9 Calculating R values

Let c generally denote a class of positions. The background evolutionary model for positions within class c is given in terms of the base frequencies w^c_α for this class, and the branch lengths t_b for each branch b of the phylogenetic tree of the clade. Along a single branch of the tree, the probability to evolve from ancestral base β to descendant base α is given by

$$P(\alpha|\beta, t_b, w^c) = \delta_{\alpha\beta} e^{-t_b} + (1 - e^{-t_b}) w^c_\alpha. \tag{16}$$

Let C denote an aligment column for the species of the clade. The probability $P(C|\mathrm{bg}, c)$ to observe alignment column C under the background evolution model of class c is given by taking the product of (16) over all branches in the phylogenetic tree, and summing over the bases at all internal nodes:

$$P(C|\text{bg}, c) = \sum_{\beta_i|i\in I} w_{\beta_r} \prod_{n\neq r} P(\beta_n|\beta_{a(n)}, t_b, w^c), \tag{17}$$

where β_i is the base at node i, I is the set of internal nodes of the tree, $a(n)$ is the ancestral node of node n, r is the root, and the product is over all nodes except for the root. The sum over the bases at the internal nodes is calculated using the standard recursive method introduced by Felsenstein [14]. That is, let C^n_α denote the probability of the subtree rooted at node n, assuming that the base at node n was α. We then have the recursion relation

$$C^n_\alpha = \prod_{m\in c(n)} \left[\sum_\beta P(\beta|\alpha, t_m, w^c) C^m_\beta\right], \tag{18}$$

where the product is over all nodes m that are in the set of children $c(n)$ of node n, and t_m is the length of the branch leading from n to child m. Note that for leafs n we have

$$C^n_\alpha = \delta_{\alpha\alpha_n}, \tag{19}$$

with α_n the base at leaf n. Starting from the leafs we can determine the C^n_α at all nodes recursively. Once we have determined C^r_α of the root r we finally have

$$P(C|\text{bg}, c) = \sum_\alpha w^c_\alpha C^r_\alpha. \tag{20}$$

For each class c the foreground model is calculated by assuming that the nucleotide frequencies w are not given but *unknown*. That is, we integrate over all possible vectors of nucleotide frequencies:

$$P(C|\text{fg}, c) = \int P(C|w)P(w|c)dw, \tag{21}$$

where $P(C|w)$ is the exact same expression as (17) but with the class specific vector of frequencies w^c replaced by the unknown vector of frequences column w, and the integral is over the simplex $w_A + w_C + w_G + w_T = 1$. Finally, $P(w|c)$ gives the prior probability distribution that a foreground column in class c will have frequency vectors w. We choose for $P(w|c)$ a Dirichlet prior, and we set the parameters of this prior to match the base composition in this class:

$$P(w|c) = \prod_\alpha \frac{(w_\alpha)^{w^c_\alpha - 1}}{\Gamma(w^c_\alpha)}. \tag{22}$$

Note that to calculate this integral we again have to sum over the bases at all internal nodes of the tree. Whereas each term in this sum can be integrated analytically using the general expression

$$\int \prod_\alpha (w_\alpha)^{n_\alpha - 1} dw = \frac{\prod_\alpha \Gamma(n_\alpha)}{\Gamma(\sum_\alpha n_\alpha)}, \tag{23}$$

there is no simple recursive way to calculate the sum and we are forced to sum all terms explicitly. However, we only need to do this once for each clade.

For each class c and each possible alignment column C we calculate the ratio $R(C|c)$ between the foreground and background model for this class,

$$R(C|c) = \frac{P(C|\text{fg}, c)}{P(C|\text{bg}, c)}, \tag{24}$$

which quantifies the amount of evidence that column C is evolving according to a selection pressure different from the background model for this class. Finally, we analyze the evidence of selection in different groups of non-coding positions by calculating the average value of $R(C|c)$ for different groups of positions. In particular, we determine the average value of R in different types of intergenic regions, the average value of R within different classes of positions within genes, and the average value of R at a given locations relative to the start and stop codons of genes.

14.10 R values for positions evolving according to the background model

For the bulk of silent positions in proteins we observe an average R value of $R = 1$. Here we show that this suggests that these positions evolve according to the background model. Assume that, for a certain class of positions, a fraction $f(C)$ show alignment column C. The average R value in these positions is then given by

$$\langle R \rangle = \sum_C f(C) \frac{P(C|\text{fg})}{P(C|\text{bg})}. \tag{25}$$

If the positions in this set are evolving according to the background model we have

$$f(C) = P(C|\text{bg}). \tag{26}$$

Therefore, we have

$$\langle R \rangle = \sum_C f(C) \frac{P(C|\text{fg})}{P(C|\text{bg})} = \sum_C P(C|\text{bg}) \frac{P(C|\text{fg})}{P(C|\text{bg})} = 1, \tag{27}$$

where the last equality follows because the foreground distribution $P(C|\text{fg})$ is of course also normalized. In summary, the fact that $R = 1$ on average at silent positions suggests that the fractions $f(C)$ at these positions are close to the background model frequencies $P(C|\text{bg})$.

14.11 R values in the limit of $t \to 0$

Imagine an alignment column for only two species that are so closely-related that their phylogenetic distance is essentially zero, i.e. $t \approx 0$, and that all nucleotides are conserved. Obviously there is no useful conservation information whatsoever in this alignment and as a consequence our R statistic should be equal at all alignment columns and not indicate any evidence of the foreground over the background model, i.e. we should have $R = 1$ for all alignment columns.

Let's assume a given alignment column of class c has α in both sequences. Under the background evolution model the probability of this data is just given by the frequency w^c_α of nucleotide α in this class. Assume that for the foreground evolution model we integrate over w with Dirichlet prior

$$P(w) = \Gamma(\lambda) \prod_{\alpha} \frac{(w_\alpha)^{\lambda_\alpha - 1}}{\Gamma(\lambda_\alpha)}, \tag{28}$$

where the λ_α are the pseudocounts of the prior and $\lambda = \sum_\alpha \lambda_\alpha$. A simple calculation shows that the probability of an alignment column with nucleotide α in both species (at distance $t = 0$) under this foreground model is given by λ_α/λ. Therefore we find that $R = 1$ if and only if $\lambda_\alpha \propto w^c_\alpha$. That is, the pseudocounts of the prior should be proportional to the overall frequencies w^c_α of the background model. This leaves the overall scale λ free to determine. The overall scale λ sets the expected bias for columns in the foreground model with $\lambda = 4$ corresponding roughly to a uniform prior. We set $\lambda = 1$ which corresponds roughly to the bias observed in known regulatory sites in *E. coli*.

14.12 Calculating substitution rates

The $R(C)$ statistic of a column C calculates the likelihood ratio of the column under the foreground and background evolutionary model. As the branch lengths in the phylogenetic tree grow it generally becomes easier to distinguish if a column is evolving under the foreground or the background model and R values thus typically grow with the total branch length of the phylogenetic tree. As the scaling of R with the branch lengths of the phylogenetic tree may complicate the detection of a correlation between genome size and selection in intergenic regions we calculated an alternative measure of the amount of selection on an alignment column which does not scale with branch length. Instead of assuming that a column evolves either according to a background model, or according to some unknown WM column, we will instead assume that each alignment column evolves according to a WM column and infer the effective substitution rate at this position.

Note that, if a given position evolves according to WM column w, then the overall rate of substitution at this position depends on the WM column w. That is, given a total mutation rate of μ, the rate $s_{\alpha\beta}$ of substitution from base α to base β is

$$s_{\alpha\beta} = \mu w_\beta, \tag{29}$$

and total rate of substitution away from α is given by

$$s_\alpha = \sum_{\beta \neq \alpha} s_{\alpha\beta} = \mu(1 - w_\alpha). \tag{30}$$

Since w_α also gives the equilibrium frequency of occurrence of base α at this position, the probability to find nucleotide α at this position at a given point in time is w_α.

Therefore, the average rate of substitution in this column is given by

$$s(w) = \sum_{\alpha} \mu w_{\alpha}(1 - w_{\alpha}) = \mu \left(1 - \sum_{\alpha} (w_{\alpha})^2 \right). \tag{31}$$

The prefactor μ just gives the overall rate at which mutations are introduced, independent of w, and the second factor encodes the efect on substitution rate by selection (and mutational bias) as encoded by WM column w. The stronger this bias in the WM column w, the lower the mutation rate. We can thus quantify the strength of selection and mutational bias at this column by a *substitution rate reduction* $\mathrm{SSR}(w)$, which we define as 1 minus the relative substitution rate $s(w)/\mu$:

$$\mathrm{SRR}(w) = 1 - \frac{s(w)}{\mu} = \sum_{\alpha} (w_{\alpha})^2. \tag{32}$$

Given an alignment column C, we can thus calculate an expected overall substitution rate reduction $\mathrm{SRR}(C)$ at this position:

$$\mathrm{SRR}(C) = \int \sum_{\alpha} (w_{\alpha})^2 P(w|C) dw = \frac{\int \sum_{\alpha} (w_{\alpha})^2 P(C|w) P(w) dw}{\int P(C|w) P(w) dw}. \tag{33}$$

That is, just as we calculated an R value for each alignment column C, we now calculate an expected overall substitution rate at this column $\mathrm{SRR}(C)$. Finally, to quantify the evidence of selection in a column C we defined the Q-statistic $Q(C)$ by normalizing $\mathrm{SRR}(C)$ to the expected $\mathrm{SRR}(C)$ given the background model:

$$Q(C) = \frac{\mathrm{SRR}(C)}{\sum_C \mathrm{SRR}(C) P(C|bg)}. \tag{34}$$

and we again calculate average Q values over classes of positions.

14.13 Estimating branch lengths with PAML

We construct pseudo-alignments by concatenating all alignment columns of fourfold degenerate codons from clique alignments. Since we want to take into account codon bias, we separate silent sites according to codon type. Similarly, we extract the intergenic alignment columns from gene-intergenic-gene alignments separately for each type of intergenic region (NR, SR, and DR).

We use these pseudo-alignments as input for PAML, together with the topology of the phylogenetic tree of the clade which we estimated as described above. We then run PAML with the following set of options:

1. Evolutionary model: HKY85 which treats transition and transversion events separately.

2. Kappa: we allow the program to estimate the ratio between transitions and transversions kappa.

3. Clock: we do not assume a molecular clock.

4. Alpha: we set parameter alpha to zero, which means that the rate of mutation is assumed constant across sites.

PAML uses maximum likelihood under the HKY85 model to estimate the branch lengths in the tree. We take the sum of branch lengths as a measure of the total rate of substitutions within regions of each type.

14.14 Multiple alignments of syntenic regions

When calculating average R values across intergenic regions and when calculating R values as a function of their position with respect to the starts and ends of genes, we want to make sure not to erroneously align intergenic regions that have undergone rearrangement since the species diverged from their common ancestor. To do this we consider an intergenic region in a given species orthologous to an intergenic region in the reference species only if the genes at both ends are orthologous and their orientation is conserved.

Let X denote an intergenic region in the reference species, let g_l and g_r denote the genes in the reference species at the left and right end of the intergenic region X, and let o_l and o_r be the orientations of the genes g_l and g_r, i.e. $o_l = 1$ means gene g_l is on the plus strand and $o_l = -1$ means it is on the negative strand. Similarly, let $\tilde{X}$ denote an intergenic region in another species of the clade with $\tilde{g}_l$ and $\tilde{g}_r$ the genes on the right and left of $\tilde{X}$, and $\tilde{o}_l$ and $\tilde{o}_r$ their orientations. The regions X and $\tilde{X}$ are considered orthologous if one of the following two sets of conditions holds

1. g_l is orthologous to $\tilde{g}_l$, g_r is orthologous to $\tilde{g}_r$, $o_l = \tilde{o}_l$, and $o_r = \tilde{o}_r$.

2. g_l is orthologous to $\tilde{g}_r$, g_r is orthologous to $\tilde{g}_l$, $o_l = -\tilde{o}_r$, and $o_r = -\tilde{o}_l$.

For each intergenic region X in the reference species we collect all orthologous intergenic regions in the other species of the clade. We then extracted, from each species, the DNA sequences of the intergenic region *plus* the two flanking genes. This set of sequences was then aligned with the T-Coffee algorithm [11] using default parameters.

For the calculation of average R values across intergenic regions of different type we only considered intergenic regions that were at least 50 bp wide.

14.15 R value profiles

The R value profiles, as shown in figures 6 and 7 were calculated from the gene-intergenic-gene multiple alignments just described. To obtain the profiles we need to calculate the average R value of alignment columns at a given positions relative to the start codon, and alignment columns at a given position relative to the stop codon. To do this we calculate, for each alignment column in each multiple alignment, the relative position r_l of the nucleotide in the reference species to the start/stop codon of the gene on the left and relative position r_r to the start/stop codon of the gene on the right (whether these are start or stop codons depends on the orientations of the flanking genes for the intergenic region under study). The R value of the column in question is then added to both averages at positions r_l and r_r.

In this way two average profiles were created, average R values around the start codons of genes, and average R values around stop codons of genes. We concatenated these profiles into a single profile by taking from each profile 150 bps in the intergenic region and 300 bps in the coding region.

14.16 RNA secondary structure statistics

For each gene in each clade we wanted to determine the secondary structure in the mRNA immediately around the start codon. It is hard to do this accurately for two main reasons: First, the secondary structure will depend on the precise transcription start site, i.e. where the mRNA starts, and this is generally unknown. Second, folding algorithms can reasonably accurately determine RNA secondary structure in thermodynamic *equilibrium* but it is likely that the true RNA secondary structure at the start of the mRNA is determined by an essentially kinetic process in which the RNA starts folding as the nascent transcript emerges from the RNA polymerase. That is, we are likely to get a more accurate approximation of the true RNA secondary structure by folding only an initial piece of the mRNA. As an approximation we chose to position the hypothesized 'start' of each transcript 60 bps upstream of its translation start and to fold a region of 150 bps long, i.e. up to 90 bps downstream of translation start.

We thus extracted, for each gene in each clade, the region from 60 bps upstream of the start codon to 90 bps downstream of the start codon and folded it using the Vienna RNA package [15]. Among the statistics that the Vienna package provides is the probability p_i (or fraction of time in equilibrium) for each nucleotide i to *not* be paired to another nucleotide. For each position i, with i running from -60 to $+90$, we calculated the average probability $\langle p_i \rangle$ by averaging p_i over all genes,

$$\langle p_i \rangle = \frac{1}{G} \sum_{g=1}^{G} p_i(g), \tag{35}$$

with $p_i(g)$ the probability that position i is open in gene g and G is the total number of genes in the genome. We also calculated the variance v_i as

$$v_i = \frac{1}{G} \sum_{g=1}^{G} p_i(g) \left(1 - p_i(g)\right), \tag{36}$$

The standard error e_i in the estimated $\langle p_i \rangle$ is then given by

$$e_i = \sqrt{\frac{v_i}{G}}. \tag{37}$$

We now want to compare the probabilities $\langle p_i \rangle$ at different locations relative to translation start. To do this we compare $\langle p_i \rangle$ at each position with the average value of $\langle p_i \rangle$ in the regions away from translation start. That is, we define 'flanking' regions running from $[-50, -31]$ and $[31, 80]$ and calculate the average openess in these areas as

$$\langle p_{\text{flanking}} \rangle = \frac{1}{70} \left[\sum_{i=-50}^{-31} \langle p_i \rangle + \sum_{i=31}^{80} \langle p_i \rangle \right]. \tag{38}$$

Note that we have excluded the regions $[-60, -51]$ and $[81, 90]$ at the ends of the sequence that we fold to avoid boundary effects, i.e. the regions at the ends of the sequence show less base pairing in general.

The standard error e_{flanking} in the estimate of $\langle p_{\text{flanking}} \rangle$ is

$$e_{\text{flanming}} = \frac{1}{72} \sqrt{\left[\sum_{i=-50}^{-31} (e_i)^2 + \sum_{i=31}^{80} (e_i)^2 \right]}. \tag{39}$$

Finally, we calculate the Z-statistic at each position i as

$$z_i = \frac{\langle p_i \rangle - \langle p_{\text{flanking}} \rangle}{\sqrt{(e_i)^2 + (e_{\text{flanking}})^2}}. \tag{40}$$

Positive z_i values indicate that position i tends to be more open than the flanking regions, and negative values indicate that the position tends to be more closed than the flanking regions.

The openess value p_i at different positions are to a large extent driven by base composition, e.g. the elevated frequency of A nucleotides immediately upstream and immediately downstream of translation start leads to less secondary structure in these areas. It is a priori not clear if the base composition is driving the RNA secondary structure in this area or that a selection for avoiding RNA secondary structure in this area is driving base composition. To test this we compared the observed openess values $\langle p_i \rangle$ with those observed at this position in G randomly generated sequences with the exact same base composition. That is, if the selection is on the RNA secondary structure then one might expect that the openess in the regions around translation start is larger even than the openess in random sequences with the same base composition.

For each clade we created G random sequences where, at each position i, the probability to put A, C, G, or T match the observed base frequencies at that position. We then fold all G sequences and calculate $\langle p_i(\text{rand}) \rangle$ for this random data-set as well as the standard errors $e_i(\text{rand})$. Finally, we calculate the Z-statistics

$$z'_i = \frac{\langle p_i \rangle - \langle p_i(\text{rand}) \rangle}{\sqrt{(e_i)^2 + (e_i(\text{rand}))^2}}. \tag{41}$$

Positive z'_i values indicate positions at which the openess is larger than in random sequences with the same base composition.

14.17 Smoothed profiles

All the position dependent profiles that are shown in the main paper and in the supporting figures were smoothed to reduce fluctuations on short distance scales. We produced the smoothed profiles $\overline{S}(x)$ of a statistic S using a double-exponential kernel:

$$\overline{S}(x) = \frac{1}{N} \sum_{y} S(x-y) e^{-\frac{|x-y|}{\alpha}} \tag{42}$$

where N is a normalization factor,

$$N = \sum_{y} e^{-\frac{|x-y|}{\alpha}} \tag{43}$$

and α is a length-scale which, for this study, was set to 3. In order to avoid the mixture of the statistics in intergenic or coding regions, special boundaries were take into account for summing in (42) and (43). That is, for calculating the smoothed statistic $\overline{S}(x)$ at a position x that lies within intergenic, the sum on the right runs only over positions y that are in intergenic as well. Similarly, for calculating the smoothed statistic $\overline{S}(x)$ at a position x within the coding region, the sum on the right runs only over positions y that are in the coding region as well.

References

[1] Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucl Acids Res 33:880–892.

[2] Mitchison G (2005) The regional rule for bacterial base composition. Trends Genet 21:440–443.

[3] Sharp PM, Li WH (1987) The Codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acids Res 15:1281–1295.

[4] Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Blattner F, et al. (2000) RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. Nucl Acids Res 28:65–7.

[5] Jaynes ET (1991) Straight-line fitting - A Bayesian solution. Http://bayes.wustl.edu/etj/node2.html.

[6] van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends in Genet 19:479–484.

[7] (1996-2004). Wu-blast. Http://blast.wustl.edu.

[8] Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19:1710–1711.

[9] Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

[10] Yang Z (1997) Paml: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in BioSciences 13:555–556.

[11] Notredame C, Higgins D, Heringa J (2000) T-Coffee: A novel method for multiple sequence alignments. J Mol Biol 302:205–217.

[12] Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

[13] Cavalli-Sforza L, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257.

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

[15] Hofacker I, Fontana W, Stadler P, Bonhoeffer L, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie/Chemical Monthly 125:167–188.