

ORIGINAL ARTICLE

The slow:fast substitution ratio reveals changing patterns of natural selection in γ -proteobacterial genomes

B Jesse Shapiro¹ and Eric Alm^{1,2,3,4,5}

¹Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA, USA; ²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; ³Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; ⁴The Virtual Institute of Microbial Stress and Survival, Berkeley, CA, USA and ⁵The Broad Institute of MIT and Harvard, Cambridge, MA, USA

Different microbial species are thought to occupy distinct ecological niches, subjecting each species to unique selective constraints, which may leave a recognizable signal in their genomes. Thus, it may be possible to extract insight into the genetic basis of ecological differences among lineages by identifying unusual patterns of substitutions in orthologous gene or protein sequences. We used the ratio of substitutions in slow- versus fast-evolving sites (nucleotides in DNA, or amino acids in protein sequence) to quantify deviations from the typical pattern of selective constraint observed across bacterial lineages. We propose that elevated S:F in one branch (an excess of slow-site substitutions) can indicate a functionally relevant change, due to either positive selection or relaxed evolutionary constraint. In a genome-wide comparative study of γ -proteobacterial proteins, we find that cell-surface proteins involved with motility and secretion functions often have high S:F ratios, whereas information-processing genes do not. Change in evolutionary constraints in some species is evidenced by increased S:F ratios within functionally related sets of genes (for example, energy production in *Pseudomonas fluorescens*), whereas other species apparently evolve mostly by drift (for example, uniformly elevated S:F across most genes in *Buchnera* spp.). Overall, S:F reveals several species-specific, protein-level changes with potential functional/ecological importance. As microbial genome projects yield more species-rich gene trees, the S:F ratio will become an increasingly powerful tool for uncovering functional genetic differences among species.

The ISME Journal (2009) 3, 1180–1192; doi:10.1038/ismej.2009.51; published online 21 May 2009

Subject Category: evolutionary genetics

Keywords: comparative genomics; ecological adaptation; γ -proteobacteria; positive selection; reverse ecology

Introduction

Natural selection is an evolutionary force that promotes the spread of beneficial alleles in a population (positive/diversifying selection), and impedes the spread of deleterious alleles (negative/purifying selection). Selection is intimately tied to ecology: depending on the ecological niche of an organism (for example, its source of carbon and nutrients, interactions with predators, hosts, competitors, and so on), different mutations will be favored by selection. Genome-wide scans for natural selection have the potential to identify ecologically

relevant genetic adaptations, even when the adaptive traits themselves remain obscure (Li *et al.*, 2008). Such genome-wide approaches, sometimes referred to as ‘reverse ecology’, thus have great potential to elucidate the ‘hidden world’ of microbial ecology. Reverse ecology requires a sampling of related genomes to quantify genetic differences and similarities within or between species, and an appropriate genome-wide test for selection. Most tests for selection have been developed with sexual eukaryotes in mind, and may not always be amenable to microbes (Shapiro *et al.*, 2009). Evidence for selection can be detected over relatively recent time scales by studying allele frequencies within populations (Sabeti *et al.*, 2002; Zeng *et al.*, 2006), or over longer time scales by studying protein evolution between species (Yang, 1998; Jordan *et al.*, 2001; Shapiro and Alm, 2008).

At the protein sequence level, natural selection is often quantified using dN/dS (substitutions per

Correspondence: E Alm, Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building 48-317, Cambridge, MA, USA.

E-mail: ejalm@mit.edu

Received 9 December 2008; revised 9 March 2009; accepted 14 April 2009; published online 21 May 2009

nonsynonymous site/substitutions per synonymous site). The theoretical foundation of dN/dS can be traced back in the development of the neutral theory of molecular evolution, when Kimura made an important observation: while not all synonymous mutations are necessarily functionally neutral, ‘the possibility is very high that, on average, synonymous changes are subject to natural selection very much less than the mis-sense mutations’ (Kimura, 1977). When an excess of missense mutations is observed relative to nearly neutral silent mutations ($dN/dS > 1$), this provides strong evidence for positive selection on a protein or portion thereof. Meanwhile, dN/dS close to zero indicates strong selective constraint, and $dN/dS \approx 1$ indicates low or ‘relaxed’ selective constraint (for example, pseudogenes). Relaxed constraint amounts to reduced efficacy of purifying selection in purging deleterious mutations from a population. It has long been recognized that dN/dS loses power to detect positive selection over long divergence times because dS becomes ‘saturated’ with multiple substitutions. More recently, it has been recognized that dN/dS may also be unreliable over very short divergence times between species (Nozawa *et al.*, 2009) or when it is applied within a single population (Kryazhimskiy and Plotkin, 2008). These issues may be particularly acute in studies of microbial genomes, where populations may be ill-defined, or divergence times may be ancient (on the scale of millions to billions of years).

In this work, we introduce the slow:fast substitution ratio (S:F) as a metric for detecting variation in natural selection on biological sequences—either nucleotides or amino acids—and apply it to detect selection among bacterial species that are sufficiently diverged, that is, most synonymous sites have undergone multiple substitutions (saturated dS). The logic underlying this new method is analogous to the logic of dN/dS (see Supplementary Note 3): sequences with an excess of substitutions in sites (positions of a nucleotide or protein sequence alignment) of probable functional importance (slow-evolving), relative to the nearly neutral standard of substitutions in sites of less functional importance (fast-evolving), are candidate targets of positive or relaxed negative selection. Unlike dN/dS, which defines site categories based on the genetic code, the S:F ratio instead relies on each site’s observed substitution rate in a phylogeny of many species—and is thus applicable not just to codon sequences, but also to noncoding or protein sequences. Substitutions are first counted in each branch of the species phylogeny. For any extant or ancestral branch, we define S as the number of substitutions per ‘slow-evolving’ site, F as the number of substitutions per ‘fast-evolving’ site and S:F as their ratio (Figure 1). Along with S:F, we calculate an odds ratio (OR) and *P*-value to assess the significance of the branch’s deviation from the S:F

observed for that gene in the rest of the gene tree (Materials and methods).

A potential limitation of the S:F method is that it condenses all the complexity of a gene sequence into a single number. Gene sequences contain a multitude of sites, some of which may be under strong purifying selection, some selectively neutral and others under strong positive selection at any moment in time (Hughes and Nei, 1988; Hughes, 2007). Such an intricate pattern of selection across sites cannot be adequately captured by a simple summary statistic (prompting the development of site-specific models of dN/dS (Yang and Nielsen, 2002; Massingham and Goldman, 2005), although these model-based methods may suffer from false-positive and false-negative adaptive site predictions (Nozawa *et al.*, 2009)). However, S:F is not designed to summarize the complex pattern of selection across the gene, but instead to quantify the extent of change in that pattern. Thus, S:F is a simple statistic quantifying changes in the regime of selection, while still acknowledging that this regime may be complex (for example, sites and lineages with different selective constraints).

Even if a lineage is found to have a significantly high value of S:F, this may result from either adaptation (positive natural selection favoring novel mutations), or relaxed selective constraint (accumulation of neutral or mildly deleterious mutations). In cases of $S:F > 1$, positive selection is a likely explanation, but such cases are expected to be rare. In more subtle cases where S:F is excessively high (but still < 1) in one lineage, either positive or relaxed negative selection may be responsible, and S:F alone cannot be used to discriminate between these possibilities.

In this study, we calculated S:F for ~1000 protein families from 30 species of γ -proteobacteria, an ancient and ecologically diverse group, with evidence for species-specific positive selection on many of their core genes (Shapiro and Alm, 2008). We aimed to identify which genes—or functional modules of genes—contribute to species-specific adaptations. More specifically, we tested the hypotheses that (1) selective constraint, as measured with S:F, varies with protein function, and (2) that ecologically distinct species experience different regimes of selection on proteins of different functions. We describe several examples of elevated S:F in proteins with functions relevant to species ecology, suggesting ecological adaptation at the protein level.

Materials and methods

Data set

A set of 917 gene families (members of the same Cluster of Orthologous Groups; Tatusov *et al.*, 1997), each represented by a single copy in at least 16 of the 30 genomes in this study, was retrieved from the

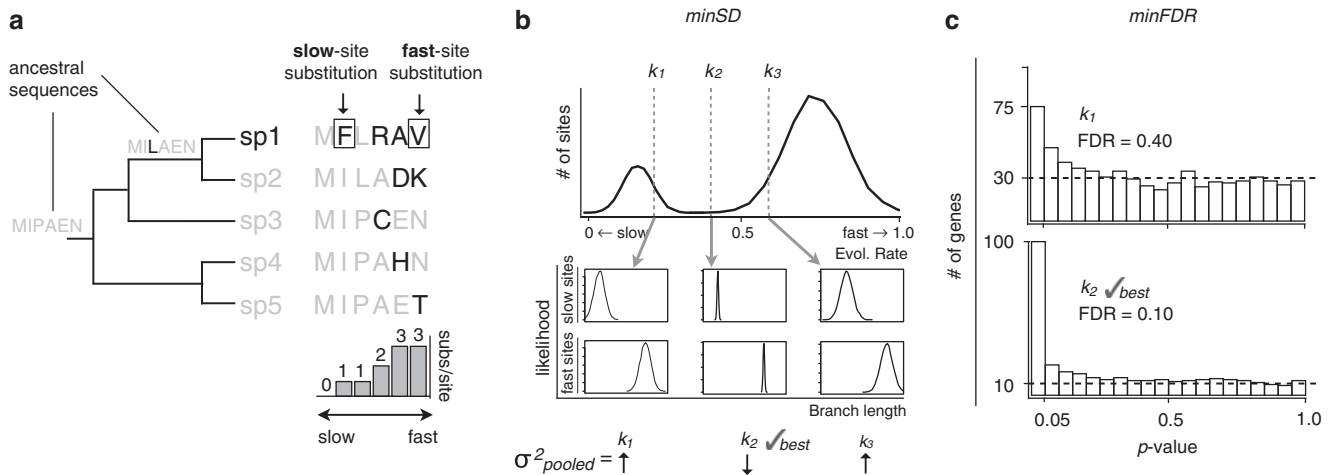


Figure 1 Overview of S:F methodology. **(a)** Hypothetical five-species phylogeny and multiple sequence alignment for a protein of six amino acids. Substituted amino acids are highlighted in black, and substitutions are ranked by the number of substitutions per site. Excluding the invariant site, there are two sites in the slowest category (one sub/site), two sites in the fastest category (three subs/site) and one site in an intermediate category (two subs/site). If the cutoff (k) were drawn such that the intermediate category is classed as ‘fast’, species 1 would have one substitution in two slow sites and three substitutions in three fast sites, yielding S:F = $(1/2)/(3/3) = 0.5$. **(b)** The minSD method to choose k . The sites in a protein are binned in a histogram (top panel) according to their evolutionary rate (relative number of subs/site, normalized to range from 0 to 1). Three possible choices of k are considered. For each k , slow and fast sites are considered separately to estimate branch lengths and likelihoods for the phylogeny. Branch-length distributions are shown for a representative branch (bottom panel). In practice, variances of all branch-length distributions in the phylogeny are computed and pooled. In this example, the intermediate choice (k_2) yields the lowest pooled variance and is thus the best choice. **(c)** The minFDR method to choose k . For each choice of k , S:F ratios and P -values are computed for all branches of all gene trees to produce a distribution of thousands of P -values, which are plotted in a histogram. The false-discovery rate (FDR) for the $P < 0.05$ bin is estimated as the average number of branches in bins with $P > 0.5$ (dashed lines) divided by the number of branches in the $P < 0.05$ bin. The value of k producing the lowest FDR (in this case, FDR = 0.10, meaning that of the 100 branches with significantly unusual S:F at the $P < 0.05$ confidence level, 10 are expected to be false positives) is chosen as the optimal cutoff.

MicrobesOnline database (Alm *et al.*, 2005). Maximum likelihood (ML) gene trees, and a consensus species tree topology, were constructed using PhyML (Guindon and Gascuel, 2003; LIRMM, Montpellier, France; Supplementary Methods).

Calculation of slow:fast substitution ratio

We performed joint reconstruction of ancestral sequences (Pupko *et al.*, 2000), implemented in PAML: Phylogenetic Analysis by Maximum Likelihood 4.0 (Yang, 1997; London, England, UK) using the ML gene-tree topologies. Sites in the protein or DNA sequence were rank-ordered (between 0 and 1, with 0 being the slowest and 1 the fastest evolving) by the number of substitutions inferred to have occurred in the site in all branches of the phylogeny (Figure 1a). A substitution-rate cutoff (k , also between 0 and 1) was then chosen to delineate slow (few substitutions in the phylogeny) and fast (many substitutions) sites. Invariant amino-acid sites (with no observed substitutions) were excluded, but invariant nucleotide sites were retained in the DNA analyses for consistency in comparison with dN/dS. S:F was computed as follows, after excluding branches with $F = 0$:

$$S:F \text{ ratio} = \frac{\text{number of substitutions per slow-evolving site}}{\text{number of substitutions per fast-evolving site}} \quad (1)$$

The number of substitutions per site was corrected for multiple hits using a Poisson correction for protein sequence (equation 2) or a Jukes–Cantor correction for DNA (equation 3).

$$d = a \left[\left(1 - \frac{c}{s} \right)^{-1/a} - 1 \right] \quad (2)$$

$$d = -\frac{3}{4} \ln \left(1 - \left(\frac{4}{3} \cdot \frac{c}{s} \right) \right) \quad (3)$$

where d is the corrected number of substitutions per site, c the number of observed substitutions, s the number of sites (fast or slow), and the parameter a is set to 2.4, as suggested for the JTT substitution model (Nei and Kumar, 2000).

The deviation of each branch from the expected S:F ratio was evaluated using Fisher’s exact test. For each branch i in a gene tree of N branches, we define S_i as the number of slow-site substitutions, and F_i as the number of fast-site substitutions in branch i . We define the total numbers of slow- and fast-site substitutions in all other branches ($x \neq i$) of the gene tree, $S_{\text{tot}} = \sum_{x \neq i}^N S_x$ and $F_{\text{tot}} = \sum_{x \neq i}^N F_x$, respectively. We compute Fisher’s OR = $(S_i/F_i)/(S_{\text{tot}}/F_{\text{tot}})$, and associated P -value to assess confidence in a branch i having S:F significantly greater (OR > 1) or less than (OR < 1) the rest of the gene tree.

Setting the cutoff (k) between slow and fast sites

We describe two methods to choose k . The minFDR method aims to maximize the sensitivity while controlling for the false-discovery rate (FDR) in the data set of ~ 1000 genes. The minSD method aims to allocate sites into distinct ‘slow’ and ‘fast’ distributions, making the distributions as nonoverlapping and as ‘tight’ as possible about their respective means.

In the minFDR method, we choose k to minimize the FDR, a measure of the signal-to-noise ratio in the data (Storey and Tibshirani, 2003). To do so, we range k from 0.05 to 0.95 in increments of 0.05 and re-compute S:F ratios and associated P -values for all branches and gene families for each value of k . The result is a distribution of P -values (over all genes and all branches) associated with each value of k . We choose the k yielding the distribution with the highest ratio of true/false positives, or the minimum FDR (Figure 1c; Supplementary Methods). We note that minFDR is only valid if the data actually contain a detectable signal of selection at the protein level, and the estimated P -values are unbiased.

In contrast, minSD chooses k to minimize instances of slow sites being miscategorized as fast (or *vice versa*) within a single gene (thus, a separate k can be chosen for each gene). Briefly, for each choice of k , slow and fast sites of the alignment are considered as two separate distributions, each used to infer the likeliest gene tree and branch lengths. If an excessively high k is chosen, some ‘true’ fast sites will be miscategorized as slow (and ‘true’ slow sites miscategorized as fast), thereby introducing greater variance into both distributions and both branch-length estimates. Our approach is thus to estimate standard deviations for both slow- and fast-site distributions, and choose the k that minimizes the standard deviation of both distributions (Figure 1b; Supplementary Methods).

Results

Performance of S:F under simulated evolution

To understand the response of S:F to changing regimes of selection, we generated simulated gene sequences along the γ -proteobacteria species tree (Supplementary Methods). The simulated sequences, 300 codons long, contained three site classes: (1) $dN/dS=0.2$ at 70% of sites, (2) $dN/dS=1.0$ at 20% of sites and (3) $dN/dS=0.1$ at 10% of sites. In the ‘baseline’ scenario, all branches in the tree evolved according to these site classes. Under ‘selection’ scenarios, dN/dS was increased to 0.5, 2.0 or 5.0 in site-class 3 (‘slow’ sites) for the designated ‘target’ branch(es). The resulting codon sequences were translated to amino acids, and S:F was calculated for each branch, with ‘slow’ and ‘fast’ defined such that sites ranking among the slowest 33% were considered ‘slow’ ($k=0.33$).

In simulations mimicking species-specific positive selection, a single ‘target’ species was assigned an accelerated nonsynonymous substitution rate in the ‘slow’ 10% of codons (site-class 3), which remained slow evolving in the other branches ($dN/dS=0.1$). Even with a moderately elevated dN/dS in ‘slow’ sites (from $dN/dS=0.1$ –0.5), the median S:F ratio in the target branch increased significantly from the ‘baseline’ scenario in 100 replicate simulations (Kolmogorov–Smirnov test: $D=0.22$; $P=0.016$). We then modeled more dramatic branch-specific positive selection, keeping dN/dS within the range previously observed (Yang and Nielsen, 2002). Dramatic increases in dN/dS in the target branch’s slow-sites led to a monotonic increase in S:F (Figure 2, light-gray bars). However, when many species (8 of 30 species, interspersed over the tree) were targeted by selection, S:F became less sensitive to detect selection (Figure 2, black bars). With so many species experiencing substitutions in ‘slow’ sites, they could no longer be classed as ‘slow’, and thus did not result in high S:F. This illustrates how S:F is sensitive to species-specific changes in selective pressure, yet relatively blind to positive selection on many/all branches. In an intermediate scenario (four species under selection; Figure 2, dark-gray bars), S:F behaves similarly to the one-target-species case for $dN/dS \leq 2$, but plateaus around $dN/dS=5$.

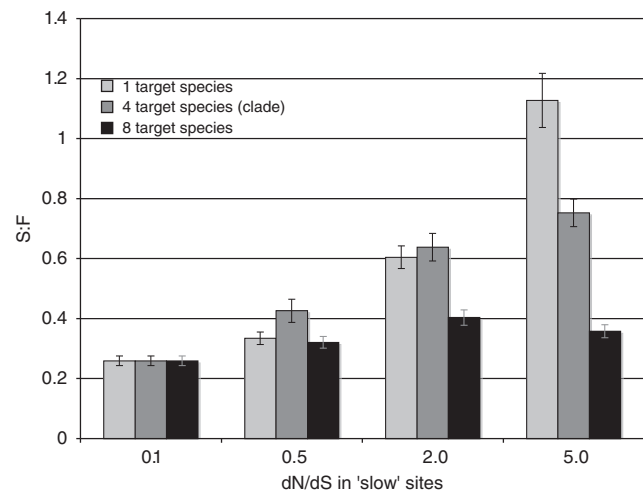


Figure 2 Response of S:F to different selection scenarios. S:F ratios for a single branch (*V. cholerae*) under selection at slow sites (light-gray bars). Eight branches (*V. cholerae*, *V. vulnificus*, *X. oryzae*, *P. syringae*, *S. oneidensis*, *P. multocida*, *E. coli*, *B. aphidicola* APS) under selection (black bars). Clade of four species (*P. syringae*, *P. aeruginosa*, *P. fluorescens* and *P. putida*) under selection (dark-gray bars). The x-axis shows dN/dS in ‘slow-evolving’ sites (class 3), comprising 10% of each sequence, and set to 0.1 in all nontarget branches in the tree. Each bar shows the mean S:F in the target branch(es) for 100 replicate simulations, with error bars showing \pm the standard error of the mean. When multiple branches are targeted, a single representative branch is displayed, chosen at random.

Delineating 'fast' and 'slow' sites

The S:F approach relies on empirical definitions of 'slow' and 'fast' sites, necessitating an optimal cutoff (k) between 'slow' and 'fast' sites. We evaluated both methods for choosing k based on mutual consistency, and consistency with dN/dS. Applied to codon data, dN/dS correlates best with S:F when $k=0.75$ (Pearson's correlation = 0.92, $P < 2.2E-16$; Supplementary Figure S1A). Although minSD allows each gene to have a different k , its average estimates of S:F also correlate best with $k=0.75$. This agrees with minFDR, which finds the minimum false-positive rate at $k=0.75$.

Applied to protein sequence, minFDR converges on $k=0.55$ (FDR = 0.18 for $P < 0.05$ and FDR = 0.055 for $P < 0.005$; Supplementary Figure S1B). However, the minSD method chooses values of k that are on average lower ($k=0.30$; Supplementary Figure S1B). Thus minSD is 'stricter', allocating fewer sites into the 'slow' category. Nevertheless, the two methods agree fairly well with one another (Pearson's correlation = 0.47, $P < 2.2E-16$).

Henceforth, we use minSD estimates of S:F (Dataset S1) as these generally provide a stricter definition of 'slow' sites, and the method makes fewer assumptions about the signature of selection in the data, but we report minFDR estimates for comparison (Dataset S2).

Regimes of natural selection on different protein functions

We set out to quantify variation in selective pressures on 917 gene families (Supplementary Table S1) in 30 species of γ -proteobacteria (Supplementary Table S2). For each branch of each gene tree, we computed S:F (using amino-acid and codon sequences), and estimated dN/dS (using codon sequences). When applied to codon sequences with an appropriate cutoff ($k=0.67$, approximating the expected proportion of nonsynonymous sites), S:F closely tracks dN/dS (Pearson's correlation = 0.91, $P < 2.2E-16$; Supplementary Note 1; Supplementary Figure S2). Yet applied to amino-acid sequences, S:F behaves differently than dN/dS (Supplementary Table S3; correlations in range 0.2–0.3, $P < 2.2E-16$). The poor correlation between the amino-acid-based S:F measure and dN/dS may be due to saturation of dS over the relatively long time scales investigated.

To test the hypothesis that different cellular functions are under different regimes of selection, we compared S:F ratios among proteins annotated with different biological functions (from the Clusters of Orthologous Groups (COG) database; Tatusov *et al.*, 1997). We picked proteins with values of S:F in the top 10% of their respective branch, and pooled together all branches into a 'high-S:F' subset (Figure 3a). We then used a hypergeometric test to determine if any functional categories were over- or underrepresented in the high-S:F subset, relative to the entire set of proteins. We used a percentile cutoff

for S:F values within a genome to control for any genome-wide inflations or deflations of S:F in a particular lineage (for example, inflation in *Buchnera* likely due to relaxed negative selection). The high-S:F protein set should therefore reflect protein-specific variation in S:F, rather than genome-wide variations in mutation rate, generation time or effective population size. We also applied an additional P -value cutoff, reducing the size of the data set while preserving its main features (Supplementary Note 2; Supplementary Figure S3).

Most noticeably, genes involved in motility and secretion (function *N*) are significantly overrepresented in the high-S:F subset (Figure 3a). This is consistent with the notion that these genes, which often code for surface proteins targeted by immune systems, predators or phage, are frequent targets of positive selection, as has been documented previously in the γ -proteobacteria, most notably in plant and enteric pathogens (Guttman *et al.*, 2006; Ma *et al.*, 2006; Weber and Koebnik, 2006; Shapiro and Alm, 2008). Nonetheless, this result is not necessarily anticipated because S:F cannot detect genes that are under positive selection in all lineages (Figure 2). Thus, not only are motility/secretion genes subject to strong positive selection, but selection must frequently apply to different genes, or different sets of amino acids, in each lineage. Elevated S:F ratios in function *N* are observed using both amino-acid (AA) and codon (DNA) sequences, both estimators of k , and dN/dS (Figure 3a), providing evidence for recurrent diversifying selection spanning ancient to more recent time scales. Motility/secretion genes are also significantly enriched among the set of genes with dN/dS > 1 (hypergeometric test, $P = 0.005$), supporting the hypothesis of frequent positive selection on these genes, rather than relaxed negative selection.

Genes involved in cell envelope biosynthesis (M), ion transport and metabolism (P), and signal transduction (T) also tend to have high S:F, although with less statistical significance. Nevertheless, these functions may be common targets of lineage-specific positive or relaxed negative selection, constituting a more 'adaptable', less constrained, component of these genomes.

In contrast, positive selection and relaxed negative selection are much less frequent among genes involved in information processing and central metabolism (functions C, E, F, G and J). These COG functions are all underrepresented among the 'high-S:F' component of genomes (Figure 3a), and are likely under similar regimes of mostly purifying selection.

Species-specific, function-specific variation in selection

We next investigated to what extent function-specific selection may also be species specific. In other words, does the set of cellular functions with

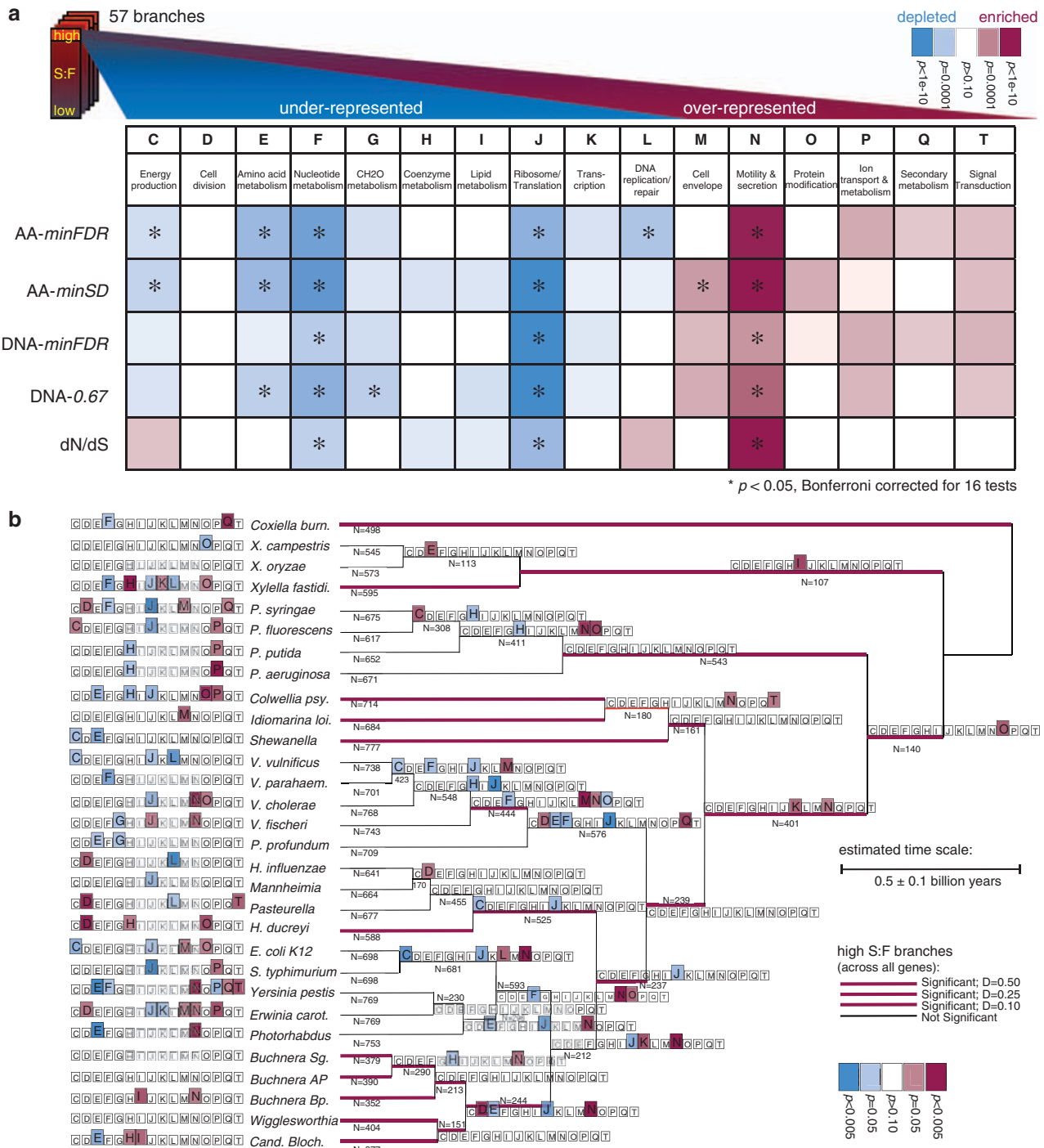


Figure 3 Enrichment/depletion of cellular functions in the high-S:F subset of genes. **(a)** Schematic of hypergeometric test results for enrichment or depletion of Clusters of Orthologous Groups (COG) functional categories of genes in the top 10% highest values of S:F within each genome, pooled over all 57 branches in the species tree. Functional categories overrepresented in the high-S:F set of genes are colored in maroon, and those underrepresented in blue, with color saturation proportional to the significance of the hypergeometric test for enrichment/depletion. The results are repeated using five different metrics: (1) S:F applied to amino-acid sequences (AA), estimating k with minFDR ($k=0.55$) or (2) estimating k with minSD, (3) S:F applied to nucleotide (DNA) sequences of the same set of genes, estimating k with minFDR, or (4) setting $k=0.67$, such that 2/3 of sites are considered slow and 1/3 fast, and (5) dN/dS estimated with the NG86 method (Nei and Gojobori, 1986). **(b)** Functional enrichment/depletion is mapped onto each branch of the γ -proteobacterial species tree, with branch lengths (time \pm standard deviation) estimated using a relaxed molecular clock model (Supplementary Methods). The number of genes in each branch for which S:F was calculated (N) is shown below each branch. Genes were only included in an internal branch if that internal branch was present in the gene tree, otherwise it was excluded. Enrichment/depletion of each functional category among the high-S:F gene set (top 10% S:F values in the branch; minSD method) is shown in maroon/blue colored boxes to the left of each species (terminal branch), or above each internal branch. Branches with positively shifted S:F distributions are highlighted in maroon (genome-wide distribution of S:F is shifted to higher values than all-genome pooled distribution; assessed by K-S test D statistic, $P < 0.05$ after Bonferroni correction for 57 branches).

unusually high S:F differ among branches of the γ -proteobacteria species tree? To address this, we again looked for enrichment/depletion of COG functions in the highest 10% of S:F values in each branch, this time on a branch-by-branch basis. By visual inspection, branches clearly differ in the set of COG functions with unusually high or low S:F ratios (Figure 3b). This difference is statistically significant: when choosing pairs of genes from the pooled high-S:F set, pairs from the same branch are more likely to have the same function than pairs from different branches (Fisher's exact test: OR = 1.33, $P < 2.2E-16$). For example, the tendency toward high S:F in motility/secretion genes is attributable mostly to enterobacteria and members of the *Vibrio* clade (Figure 3b), perhaps due to unusually strong diversifying selection on cell-surface proteins in these species.

Certain lineages have globally skewed rates of evolution across all their genes, due to species-specific differences in effective population size, mutation rates or generation times (Moran, 1996; Ochman *et al.*, 1999; Woolfit and Bromham, 2003). The *Buchnera* clade of aphid endosymbionts is a classic example: *Buchnera* experience population bottlenecks in each transmission cycle, reducing the efficacy of purifying selection, and allowing frequent fixation of deleterious mutations (Itoh *et al.*, 2002; Herbeck *et al.*, 2003; Fry and Wernegreen, 2005). This is recapitulated in the genome-wide S:F distributions for *Buchnera*, as well as the *Wigglesworthia* and *Candidatus Blochmannia* species of insect endosymbionts, which are all biased toward high S:F ratios (Figure 3b). The bias applies across all gene functions: *Buchnera* show little functional enrichment or depletion among their high-S:F genes, consistent with reduced efficacy of selection relative to genetic drift.

In addition to the insect endosymbionts, several other branches are shifted toward high values of S:F. For example, internal branches often have high S:F (and even higher values of dN/dS), perhaps due to ancestral sequence reconstruction errors (Supplementary Table S4). Moreover, short branches have slightly higher dN/dS than longer branches (Supplementary Figure S4), because purifying selection has had less time to purge deleterious mutations from the population (Rocha *et al.*, 2006). The short-branch effect, like sequencing error, is only expected to influence dN/dS or S:F in leaf branches, because the same deleterious mutation (or sequencing error) would have to occur twice independently to be incorporated into an internal branch. Because S:F is not inflated in short leaf branches (Supplementary Figure S4), it appears that neither sequencing errors nor unpurged deleterious polymorphisms present a major source of bias in our results. Yet errors in ancestral reconstruction may significantly bias S:F estimates in internal branches, and although they may be less biased than dN/dS (Supplementary Table S4), S:F in internal branches should still be interpreted cautiously.

Selection example I: redox metabolism in *Pseudomonas*

Proteins involved in energy production (function C) tend to have low S:F in most species (Figure 3), consistent with uniform purifying selection. The only exception is the *Pseudomonas* clade, notably *P. fluorescens*, which has an excess of high-S:F energy production genes (Figure 3b). Many of these genes are co-expressed on the same operon (Figure 4) and tend to have elevated S:F in *Pseudomonas* but rarely in other clades (Figure 4). This pattern of selection is discernible at the protein level (both minSD and minFDR methods), but is weaker at the codon level, partially due to saturation of dS (Figure 4, bottom panel). For example, the pyruvate dehydrogenase E1 component (AceE; COG 2609) has high S:F at the protein level in *P. putida*, but codon-level selection is not detectable with dN/dS. Consistent with species-specific, protein-level adaptation, significant structural differences are known to have occurred in AceE between *P. putida* and *Escherichia coli* (Arjunan *et al.*, 2002). Moreover, pseudomonads often inhabit oxygen-limited biofilms, where they produce alternative electron acceptors such as phenazines to maintain redox homeostasis (Price-Whelan *et al.*, 2006, 2007). Phenazines may interact with AceE: inhibiting it by generating superoxides, or promoting its activity by re-oxidizing one of its products, NADH (Price-Whelan *et al.*, 2007). These potentially *Pseudomonas*-specific biochemical interactions may impose lineage-specific selective pressures on AceE and other redox metabolism genes.

In another example, we found both transmembrane subunits of the succinate/ubiquinone dehydrogenase complex, SdhC and SdhD, among the high-S:F subset of *P. fluorescens* genes (Figure 4). This complex shuttles electrons from succinate to ubiquinone as part of the electron transport chain. SdhC has high S:F in two *Pseudomonas* species, but no other lineages (Figure 4), suggesting a lineage-specific evolutionary change. SdhC is in the 1% highest values of S:F in the *P. fluorescens* genome, due to three slow-site substitutions (S:F = 1.5, $P < 0.05$, minSD; Table 1). We mapped these substitutions onto the *E. coli* Sdh protein structure (Yankovskaya *et al.*, 2003) and discovered that one substitution, Phe \rightarrow Tyr58, is in contact with a bound cardiolipin phospholipid (Yankovskaya *et al.*, 2003), whereas the other two, Ala \rightarrow Gly24 and Ile \rightarrow Phe28, fall in the path of electron transport between the 3Fe-4S cluster and ubiquinone (Figure 5a). The latter site, Ile28, makes up part of the ubiquinone-binding site, and is perfectly conserved across species in this study except *P. fluorescens*. Further confirming the species specificity of this substitution, *P. fluorescens* Pf-5 (the strain used in this study) and *P. fluorescens* PfO-1 (the only other *P. fluorescens* genome in MicrobesOnline) are the only 2 strains harboring the Ile \rightarrow Phe28 substitution, of 16 total *Pseudomonas* strains with SdhC

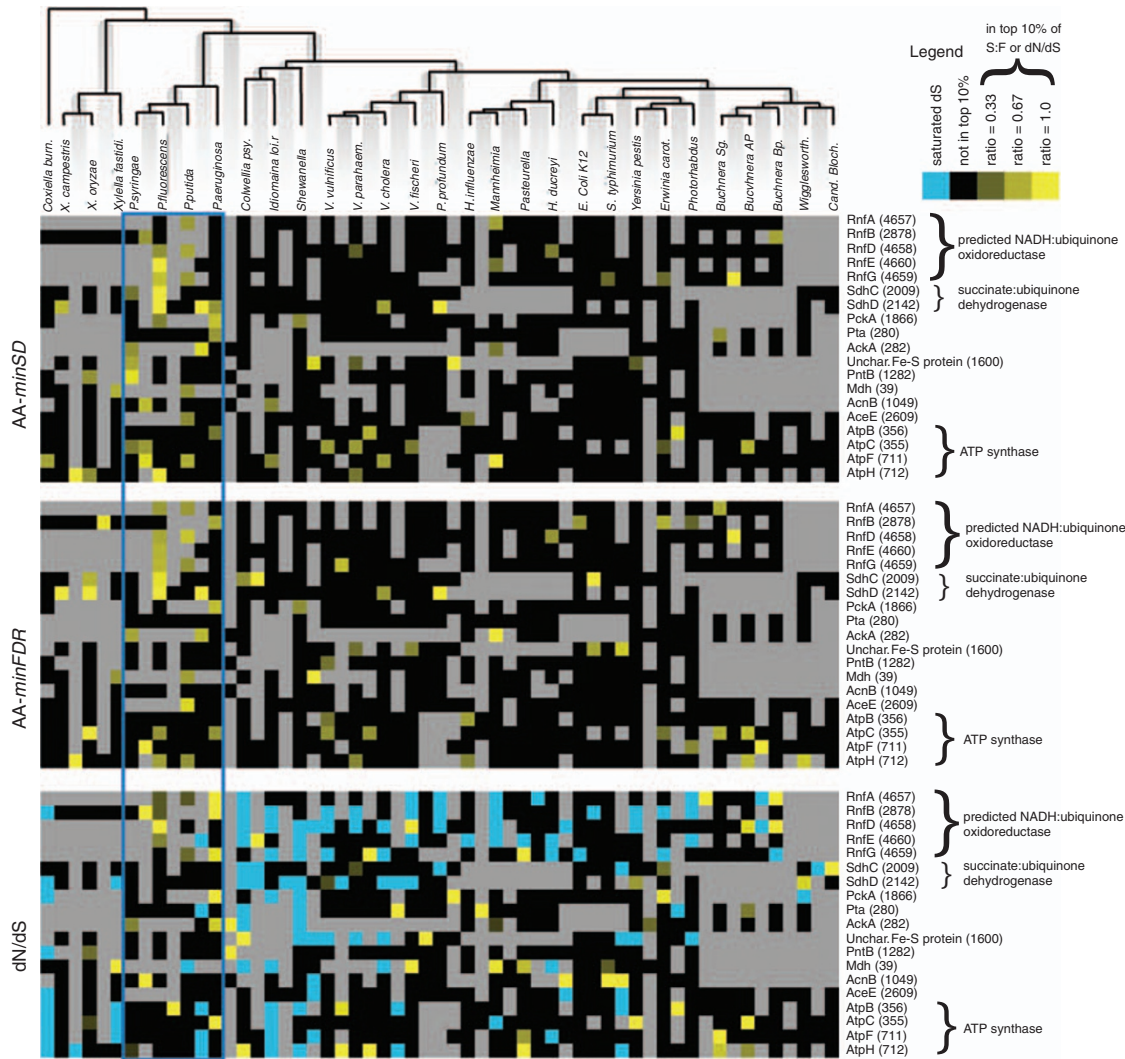


Figure 4 Genes involved in energy production have elevated S:F in pseudomonads. Gene-by-branch heatmap for genes in category C (energy production) in top 10% of S:F in one or more branches of the *Pseudomonas* clade (blue box). Columns represent either terminal branches, or internal nodes (highlighted in gray on the tree). Data are presented for three different methods: S:F applied to amino-acid data using minSD to choose k (top), using minFDR to choose k (middle) or dN/dS (NG86 method; Nei and Gojobori, 1986) applied to codons (bottom). Red, gene in top 10% of S:F (or dN/dS) values in the branch, with saturation proportional to the magnitude of the S:F (or dN/dS) ratio; black, gene is present in the branch but not among top 10%; gray: gene is not present in the branch, or the branch in the gene tree does not correspond to a monophyletic clade in the species tree; blue, ratio not estimated because the denominator (F or dS) is saturated with substitutions. Genes are listed by the short name of their *E. coli* ortholog, with Clusters of Orthologous Groups (COG) number in parentheses. Branch lengths in the species tree are not to scale. Genes on the same operon in *E. coli* (Price *et al.*, 2006) are grouped together with curly brackets.

orthologs in the database. This also serves as tentative confirmation that the substitution is fixed in *P. fluorescens*, and is not simply a slightly deleterious polymorphism segregating in the population (Hughes *et al.*, 2008). The substituted side-chain (Phe) is substantially larger than Ile, and would clash directly with ubiquinone unless there were some local modifications of the protein structure (Figure 5a). Moreover, mutations at the equivalent site in human Sdh cause disease (Astuti *et al.*, 2001), and result in oxidative stress in nematodes (Ishii *et al.*, 1998) due to electron leakage (Yankovskaya *et al.*, 2003). Often associated with superoxide-producing plants, *P. fluorescens* has a

number of mechanisms for coping with oxidative stress (Paulsen *et al.*, 2005). The Ile→Phe28 substitution might therefore be tolerated by *P. fluorescens* due to relaxed negative selection against free radical production. However, the occurrence of another nearby substitution in the path of electron transport (Ala→Gly24) suggests an adaptive change. Given the diversity and ecological importance of secondary ‘respiratory pigments’ produced by pseudomonads (Mavrodi *et al.*, 2006; Price-Whelan *et al.*, 2006), it is not unreasonable to speculate that central metabolic respiratory pathways involving redox balance may be under positive selection to better interface with these secondary pathways.

Table 1 Substitutions in slow and fast sites in *P. fluorescens* SdhC (COG 2009)

Method	k	Slow (S)		Fast (F)		S:F		Fisher's test	
		No. of subs.	No. of sites	No. of subs.	No. of sites	Ratio ^a	Rank	OR	P
AA minFDR	0.55	5	66	3	37	0.93	2% ^b	2.42	0.28
AA minSD	0.35	3	30	5	73	1.50	1% ^b	5.26	0.043
DNA minFDR	0.75	17	275	23	88	0.20	61%	0.82	0.63
DNA $k=0.67$	0.67	8	228	32	135	0.13	55%	0.59	0.22
dN/dS NG86	NA	dN = 0.042		dS = 0.39		0.11	37%		NA
dN/dS PAML	NA	dN = 0.042		dS = 0.29		0.14	27%		NA

Data are shown for S:F applied to amino-acid data (AA) and nucleotide data (DNA), and for dN/dS, each estimated by two different methods (PAML and NG86). For each method, sites were percentile-ranked based on the number of substitutions/site, and divided into slow and fast at the rank cutoff (k) indicated (Materials and methods). The 'Rank' column indicates the percent rank of the ratio (S:F or dN/dS) among genes in the *P. fluorescens* genome, with 1% indicating very high S:F. In the Fisher's test column, OR > 1 indicates that the S:F ratio is greater in *P. fluorescens* than other branches of the gene tree.

^aCorrected for multiple substitutions.

^bS:F ratio is among the top 10% highest in the genome.

Selection example II: outer membrane in *V. cholerae*
Another potential ecological adaptation is presented by the outer membrane protein OmpW (COG 3047) of the human pathogen *Vibrio cholerae*. Low DNA-S:F and dN/dS show that OmpW is highly conserved in *V. cholerae*, with few amino-acid-altering substitutions relative to silent substitutions (Table 2). Yet of these few amino-acid changes, an unexpectedly high number occur in slow-evolving sites, suggesting lineage-specific positive or relaxed negative selection (S:F = 1.38, $P < 0.05$, minSD; Table 2). We focus on this protein because it is present in all known *V. cholerae* strains, is highly immunogenic, suggesting it may be subject to immune selection (Das *et al.*, 1998), and is upregulated in related vibrios under high-NaCl stress (Xu *et al.*, 2005), suggesting a role in osmoregulation. Of the 12 substitutions inferred in *V. cholerae* using the minFDR method, the 6 in slow-sites cluster more closely with one another in the 3D structure (Hong *et al.*, 2006) than do the 6 in fast-sites (mean pairwise Euclidean distance between C_{α} atoms = 21.5 Å for slow sites; 27.2 Å for fast sites; two-sample one-sided Wilcoxon test: $W = 75$, $P = 0.06$), suggesting that the slow-site substitutions may represent structurally coordinated adaptive changes. Indeed, the substitutions in the two most highly conserved sites, Leu → Val55 and Leu → Phe83, are adjacent in the protein structure, despite being distant in the linear protein sequence (Figure 5b). They are localized just below the putative exit channel, where a small molecule may exit the hydrophobic barrel and enter the outer membrane (Hong *et al.*, 2006). The substitutions might thus alter substrate specificity or transport kinetics of the channel. All six slow-site substitutions are present in the additional six *V. cholerae* strains (V51, V52, RC385, O395, MO10 and 2740-80) with sequences in MicrobesOnline, consistent with functional significance of these substitutions, and confirming that they are not slightly deleterious polymorphisms or sequencing errors. However, these substitutions are

not all unique to *V. cholerae*: Leu → Val55 and Leu → Phe83 both occur in *V. splendidus* 12B01 and *Photobacterium profundum* 3TCK (not in *P. profundum* SS9, which appears to have lost COG3047). Horizontal gene transfer could be responsible for this phylogenetically incongruence, but would require two separate transfer events because *V. cholerae* contains an insertion of the sequence SGGELG between residues 67 and 68, which is not present in either potential donor, *P. profundum* or *V. splendidus* (Figure 5b). Therefore, convergent evolution is the more parsimonious explanation for this covarying pair of substitutions, and this likely implies positive selection (Sokurenko *et al.*, 2004; Falush and Bowden, 2006; Holt *et al.*, 2008).

Discussion

S:F as a method to detect changes in the regime of selection

We have described a method for detecting selection at the protein or DNA level that is conceptually similar to dN/dS, but is more general, relying on empirical definitions of 'slow' and 'fast' sites rather than predefined nonsynonymous/synonymous sites. In general, S:F identifies deviations from a sequence's 'usual' regime of selection, whether that regime is neutral, involves strong purifying or diversifying selection, or some complex combination of these regimes. An advantage of S:F over dN/dS is its suitability to anciently diverged clades, such as the γ -proteobacteria, in which synonymous sites are often saturated with multiple substitutions. Applied to more closely related strains, it may lack power due to paucity of substitutions, but should still be more conservative (for example, fewer false positives) than branch- and site-based models of dN/dS (Nozawa *et al.*, 2009).

As an empirical method, S:F exploits the availability of species-rich protein families, made possible by whole-genome sequencing of related

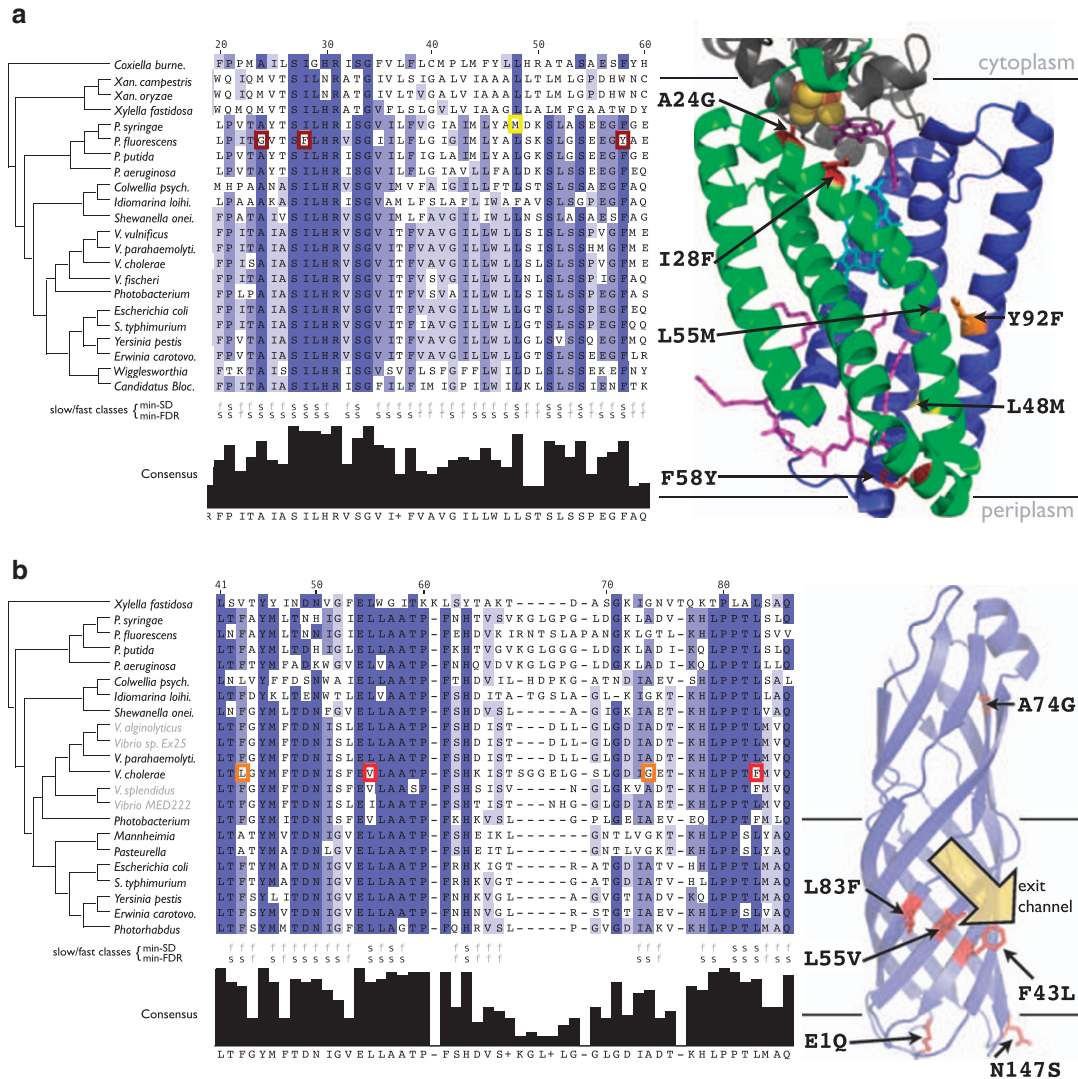


Figure 5 Alignment and structure of proteins with high S:F. **(a)** Multiple sequence alignment (MSA) of SdhC transmembrane helix I (left), with positions numbered according to the *E. coli* structure (Yankovskaya *et al.*, 2003) (right). Columns of the MSA are colored by conservation, with site categories (slow or fast), as determined by both minSD and minFDR methods, as well as the majority-rule consensus, shown below each column. Perfectly conserved columns were not assigned a slow:fast category. Slow-site substitutions in *P. fluorescens*, assigned by the minSD method, are boxed in red, *P. syringae* in yellow. The structure shows subunits SdhC (green), SdhD (blue) and part of SdhB (gray). Slow sites with substitutions in *P. fluorescens* are colored in red, *P. syringae* in yellow and *P. aeruginosa* in orange. The *E. coli* side chains, not the substituted *Pseudomonas* residues, are depicted. Other molecules in the structure are: ubiquinone (purple), heme b (cyan), cardiolipin (magenta) and the 3Fe–4S iron–sulfur cluster (yellow/orange spheres). Branch lengths in the species tree are not to scale. Structure image generated using PyMOL (DeLano, 2002; Palo Alto, CA, USA); MSA image using Jalview (Clamp *et al.*, 2004; Dundee, Scotland, UK). **(b)** MSA of OmpW (left), with positions numbered according to the *E. coli* structure (Hong *et al.*, 2006) (right). *V. cholerae* substitutions in slow-sites are boxed in orange if identified by the minFDR method, or in red if by the minSD method. The same color scheme is used on the structure, with *E. coli* side chains depicted. Species added to the MSA but not among the 30 species used in other analyses are shown in gray text.

species. Depending on the diversity and breadth of species included, S:F will identify different sets of slow- and fast-evolving sites. The method is therefore flexible, and potentially sensitive to selection at different time scales. In this study, we investigated the relatively broad hypothesis that patterns of function-specific natural selection vary among ecologically distinct species. The method also lends itself well to more specific hypotheses, aimed at particular groups of interest.

Distinguishing adaptive evolution

Elevated S:F may be attributed to either positive selection or species-specific relaxation of negative selection. Both scenarios have the potential to be biologically informative, and may suggest ecological adaptation. For example, the Ile→Phe28 substitution in *P. fluorescens* SdhC may have been ‘passively tolerated’ by relaxed selective constraint on this residue, or ‘actively’ pushed to fixation by positive selection for a novel or improved function. Without within-population sampling (for example,

Table 2 Substitutions in slow and fast sites in *V. cholerae* OmpW (COG 3047)

Method	k	Slow (S)		Fast (F)		S:F		Fisher's test	
		No. of subs.	No. of sites	No. of subs.	No. of sites	Ratio ^a	Rank	OR	P
AA minFDR	0.55	6	75	6	49	0.63	1.5% ^b	1.71	0.38
AA minSD	0.35	2	16	10	108	1.38	0.1% ^b	8.25	0.037
DNA minFDR	0.75	34	360	86	135	0.07	25%	0.46	0.00015
DNA $k=0.67$	0.67	25	325	95	170	0.08	49%	0.46	0.00056
dN/dS NG86	NA	dN = 0.07		dS > 0.75 ^c		< 0.093	NA ^c		NA
dN/dS PAML	NA	dN = 0.08		dS = 3.70		< 0.022	42%		NA

Data as described in Table 1 legend. The 'Rank' column indicates the percent rank of the ratio (S:F or dN/dS) among genes in the *V. cholerae* genome, with 1% indicating very high S:F.

^aCorrected for multiple substitutions.

^bS:F ratio is among the top 10% highest in the genome.

^cSaturated in NG86, rank and ratio are only approximate.

McDonald–Kreitman tests; Li *et al.*, 2008; Shapiro *et al.*, 2009), it is difficult to distinguish between these scenarios. Yet the substitution is lineage specific (Figures 4 and 5a), strongly suggesting some sort of functional re-wiring of redox metabolism and electron transport in *P. fluorescens*. The substitution is also gene-specific: SdhC has an S:F ratio in the top 1% of the *P. fluorescens* genome (Table 1) and therefore cannot be attributed it to a genome-wide shift in substitution rates, or possible biases in S:F due to branch length. By further accounting for *P. fluorescens*' ecology—a phenazine-producing, plant-associated organism with a high metabolic capacity—we gain confidence in the adaptive value of substitutions in an electron transport protein. Similar lines of evidence lend support to the hypothesis that OmpW has acquired ecologically adaptive substitutions in *V. cholerae*. In both examples, further experimental work is needed to fully understand and validate the predictions of our 'reverse ecology' approach.

Conclusions

In our analysis of adaptive protein evolution across 30 γ -proteobacteria, we were able to glean several insights, both global and specific. Globally, we found that proteins localized to the cell surface (functioning in motility/secretion or cell envelope biosynthesis) are frequent targets of positive or relaxed negative selection, showing elevated S:F ratios across many species, especially those involved in host–pathogen or host–symbiont interactions. Meanwhile, proteins involved in 'house-keeping' roles tend to be under purifying selection, which we observe as low S:F ratios. Yet there are exceptions to this rule: we observe instances of species- or clade-specific reversals of purifying selection, for example the unusually high S:F ratios observed in a suite of energy metabolism proteins in pseudomonads.

The method we describe is a flexible, empirical approach for detecting varying regimes of natural

selection. It can be applied to study selection on protein-coding sequences, or non-coding genomic sequences, such as promoters and non-coding RNAs. In this work, we showed how S:F can be applied over evolutionary time scales beyond the reach of dN/dS. Discriminating between positive and relaxed negative selection remains a challenge, but we reason that both scenarios are ecologically informative. As we accumulate whole-genome sequences for more and more ecologically diverse species, the S:F method will be useful in detecting the protein-level adaptations that functionally distinguish between them.

Acknowledgements

We thank Dianne Newman and Alexa Price-Whelan for useful discussion and insights into phenazine metabolism in pseudomonads. This work was part of the Virtual Institute for Microbial Stress and Survival (<http://VIMSS.lbl.gov>) supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and US Department of Energy. BJS was funded by a National Institutes of Health (NIH) training grant and a Natural Sciences and Engineering Research Council of Canada (NSERC), Canada Graduate Scholarship.

References

- Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL *et al.* (2005). The MicrobesOnline web site for comparative genomics. *Genome Res* **15**: 1015–1022.
- Arjunan P, Nemeria N, Brunskill A, Chandrasekhar K, Sax M, Yan Y *et al.* (2002). Structure of the pyruvate dehydrogenase multienzyme complex E1 component from *Escherichia coli* at 1.85 Å resolution. *Biochemistry* **41**: 5213–5221.
- Astuti D, Latif F, Dallol A, Dahia PL, Douglas F, George E *et al.* (2001). Gene mutations in the succinate

- dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. *Am J Hum Genet* **69**: 49–54.
- Clamp M, Cuff J, Searle SM, Barton GJ. (2004). The Jalview Java alignment editor. *Bioinformatics* **20**: 426–427.
- Das M, Chopra AK, Cantu JM, Peterson JW. (1998). Antisera to selected outer membrane proteins of *Vibrio cholerae* protect against challenge with homologous and heterologous strains of *V. cholerae*. *FEMS Immunol Med Microbiol* **22**: 303–308.
- DeLano WL. (2002). The PyMOL Molecular Graphics System <http://www.pymol.org/>.
- Falush D, Bowden R. (2006). Genome-wide association mapping in bacteria? *Trends Microbiol* **14**: 353–355.
- Fry AJ, Wernegreen JJ. (2005). The roles of positive and negative selection in the molecular evolution of insect endosymbionts. *Gene* **355**: 1–10.
- Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.
- Guttman DS, Gropp SJ, Morgan RL, Wang PW. (2006). Diversifying selection drives the evolution of the type III secretion system pilus of *Pseudomonas syringae*. *Mol Biol Evol* **23**: 2342–2354.
- Herbeck JT, Funk DJ, Degnan PH, Wernegreen JJ. (2003). A conservative test of genetic drift in the endosymbiotic bacterium *Buchnera*: slightly deleterious mutations in the chaperonin groEL. *Genetics* **165**: 1651–1660.
- Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I *et al.* (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. *Nat Genet* **40**: 987–993.
- Hong H, Patel DR, Tamm LK, van den Berg B. (2006). The outer membrane protein OmpW forms an eight-stranded beta-barrel with a hydrophobic channel. *J Biol Chem* **281**: 7568–7577.
- Hughes AL, Friedman R, Rivaille P, French JO. (2008). Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol Biol Evol* **25**: 2199–2209.
- Hughes AL. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**: 364–373.
- Hughes AL, Nei M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167–170.
- Ishii N, Fujii M, Hartman PS, Tsuda M, Yasuda K, Senoo-Matsuda N *et al.* (1998). A mutation in succinate dehydrogenase cytochrome *b* causes oxidative stress and ageing in nematodes. *Nature* **394**: 694–697.
- Itoh T, Martin W, Nei M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc Natl Acad Sci* **99**: 12944–12948.
- Jordan IK, Kondrashov FA, Rogozin IB, Tatusov RL, Wolf YI, Koonin EV. (2001). Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol* **2**: RESEARCH0053.
- Kimura M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275–276.
- Kryazhimskiy S, Plotkin JB. (2008). The population genetics of dN/dS. *PLoS Genet* **4**: e1000304.
- Li YF, Costello JC, Holloway AK, Hahn MW. (2008). ‘Reverse ecology’ and the power of population genomics. *Evolution* **62**: 2984–2994.
- Ma W, Dong FF, Stavrinides J, Guttman DS. (2006). Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS Genet* **2**: e209.
- Massingham T, Goldman N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**: 1753–1762.
- Mavrodi DV, Blankenfeldt W, Thomashow LS. (2006). Phenazine compounds in fluorescent *Pseudomonas* spp. biosynthesis and regulation. *Annu Rev Phytopathol* **44**: 417–445.
- Moran NA. (1996). Accelerated evolution and Muller’s ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci* **93**: 2873–2878.
- Nei M, Gojobori T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**: 418–426.
- Nei M, Kumar S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press: Oxford, UK.
- Nozawa M, Suzuki Y, Nei M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci* (doi: 10.1073/pnas.0901855106, early edition).
- Ochman H, Elwyn S, Moran NA. (1999). Calibrating bacterial evolution. *Proc Natl Acad Sci USA* **96**: 12638–12643.
- Paulsen IT, Press CM, Ravel J, Kobayashi DY, Myers GS, Mavrodi DV *et al.* (2005). Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat Biotechnol* **23**: 873–878.
- Price MN, Arkin AP, Alm EJ. (2006). OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC Bioinformatics* **7**: 19.
- Price-Whelan A, Dietrich LE, Newman DK. (2007). Pyocyanin alters redox homeostasis and carbon flux through central metabolic pathways in *Pseudomonas aeruginosa* PA14. *J Bacteriol* **189**: 6372–6381.
- Price-Whelan A, Dietrich LE, Newman DK. (2006). Rethinking ‘secondary’ metabolism: physiological roles for phenazine antibiotics. *Nat Chem Biol* **2**: 71–78.
- Pupko T, Pe’er I, Shamir R, Graur D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol* **17**: 890–896.
- Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH *et al.* (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* **239**: 226–235.
- Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832–837.
- Shapiro BJ, Alm EJ. (2008). Comparing patterns of natural selection across species using selective signatures. *PLoS Genet* **4**: e23.
- Shapiro BJ, David LA, Friedman J, Alm EJ. (2009). Looking for Darwin’s footprints in the microbial world. *Trends Microbiol* **17**: 5 (in press).
- Sokurenko EV, Feldgarden M, Trintchina E, Weissman SJ, Avagyan S, Chattopadhyay S *et al.* (2004). Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. *Mol Biol Evol* **21**: 1373–1383.

- Storey JD, Tibshirani R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* **100**: 9440–9445.
- Tatusov RL, Koonin EV, Lipman DJ. (1997). A genomic perspective on protein families. *Science* **278**: 631–637.
- Weber E, Koebnik R. (2006). Positive selection of the Hrp pilin HrpE of the plant pathogen *Xanthomonas*. *J Bacteriol* **188**: 1405–1410.
- Woolfit M, Bromham L. (2003). Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol Biol Evol* **20**: 1545–1555.
- Xu C, Wang S, Ren H, Lin X, Wu L, Peng X. (2005). Proteomic analysis on the expression of outer membrane proteins of *Vibrio alginolyticus* at different sodium concentrations. *Proteomics* **5**: 3142–3152.
- Yang Z, Nielsen R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**: 908–917.
- Yang Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**: 568–573.
- Yang Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.
- Yankovskaya V, Horsefield R, Tornroth S, Luna-Chavez C, Miyoshi H, Leger C *et al.* (2003). Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science* **299**: 700–704.
- Zeng K, Fu YX, Shi S, Wu CI. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**: 1431–1439.

Supplementary Information accompanies the paper on The ISME Journal website (<http://www.nature.com/ismej>)