

Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

The Slow:Fast substitution ratio reveals changing patterns of natural selection in gamma-proteobacterial genomes

Permalink

<https://escholarship.org/uc/item/47h0k4tw>

Author

Shapiro, B. Jesse

Publication Date

2009-06-19

The Slow:Fast substitution ratio reveals changing patterns of natural selection in γ -proteobacterial genomes

B. Jesse Shapiro¹ & Eric Alm^{1,2,3,4,5§}

¹Program in Computational & Systems Biology, Massachusetts Institute of Technology, Cambridge, MA

²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA

³Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA

⁴The Virtual Institute of Microbial Stress and Survival, <http://vimss.lbl.gov>

⁵The Broad Institute of MIT and Harvard, Cambridge, MA

§Corresponding author

Email addresses:

BJS: jesse1@mit.edu

EA: ejalm@mit.edu

Running title: S:F ratio reveals changing patterns of selection

Subject Category: Evolutionary genetics

Abstract

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 use 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, while 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 (*e.g.* energy production in *Pseudomonas fluorescens*), while other species apparently evolve mostly by drift (*e.g.* 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.

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 (e.g. its source of carbon and nutrients, interactions with predators, hosts, competitors, etc.), different mutations will be favoured 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 (Zeng et al, 2006, Sabeti et al, 2002), or over longer time scales by studying protein evolution between species (Jordan et al, 2001, Yang, 1998, Shapiro & Alm, 2008).

At the protein sequence level, natural selection is often quantified using dN/dS (substitutions per 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 mis-sense 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 (e.g. 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 & Plotkin, 2008). These issues may be particularly acute in studies of microbial genomes, where populations may be ill-defined, or divergences 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 variation in natural selection among bacterial species that are sufficiently diverged that 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 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 (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 & 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 & Nielsen, 2002, Massingham & 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 (*e.g.* 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 & 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 MicrobesOnline database (Alm et al, 2005). Maximum-likelihood (ML) gene trees, and a consensus species tree topology were constructed using PhyML (Guindon & Gascuel, 2003; Supplementary Methods).

Calculation of Slow:Fast substitution ratio (S:F)

We performed joint reconstruction of ancestral sequences (Pupko et al, 2000), implemented in

PAML: Phylogenetic Analysis by Maximum Likelihood 4.0 (Yang, 1997) 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:

Equation 1.

S:F ratio = number of substitutions per slow-evolving site / number of substitutions per fast-evolving site

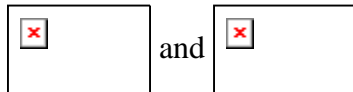
The number of substitutions per site was corrected for multiple hits using a Poisson correction for protein sequence (Equation 2) or a Jukes-Cantor correction for DNA (Equation 3).

Equation 2.

Equation 3.

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

The deviation of each branch from the expected S:F ratio was evaluated using a Fisher 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,

, respectively. We compute Fisher's Odds Ratio statistic, $O.R. =$

$(S_i/F_i) / (S_{tot}/F_{tot})$, and associated p -value to assess confidence in a branch i having S:F significantly greater ($O.R. > 1$) or less than ($O.R. < 1$) than 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 dataset of ~1000 genes. The *minSD* method aims to allocate sites into distinct 'slow' and 'fast' distributions, making the distributions as non-overlapping and as 'tight' as possible about their respective means.

In the *minFDR* method, we choose k to minimize the false discovery rate (FDR), a measure of the ratio of signal to noise in the data (Storey & 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.

5 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
10 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).

15

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 3 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 $dN/dS = 0.1$ to 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 & 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-grey bars). However, when many species (8/30 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 (4 species under selection; Figure 2, dark-grey bars), S:F behaves similarly to the 1-target-species case for $dN/dS \leq 2$, but plateaus around $dN/dS = 5$.

5

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, $p < 2.2e-16$; Figure S1A). While *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 $k = 0.75$.

10

15

Applied to protein sequence, *minFDR* converges on $k = 0.55$ (FDR = 0.18 for $p < 0.05$ and FDR=0.055 for $p < 0.005$; Figure S1B). However, the *minSD* method chooses values of k that are on average lower ($k = 0.30$; 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, $p < 2.2e-16$).

20

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 (Table S1) in 30 species of γ -proteobacteria (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).

- 5 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; Figure S2). Yet applied to amino acid sequences, S:F behaves differently than dN/dS (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
- 10 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)).

- 15 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 under-represented 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
- 20 lineage (*e.g.* 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 and Figure S3).

25

Most noticeably, genes involved in motility and secretion (function N) are significantly over-represented 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 -
5 proteobacteria, most notably in plant and enteric pathogens (Shapiro & Alm, 2008, Weber & Koebnik, 2006, Ma et al, 2006, Guttman et al, 2006). 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
10 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
15 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
20 negative selection, constituting a more 'adaptable', less constrained, component of these genomes.

In contrast, positive 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 under-represented 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

5 We next investigated to what extent function-specific selection may also be species-specific. In other words, does the set of cellular functions with 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
10 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 test: Odds Ratio = 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
15 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
20 (Moran, 1996, Ochman et al, 1999). 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 (Herbeck et al, 2003, Itoh et al, 2002, Fry & Wernegreen, 2005). This is recapitulated in the genome-wide S:F distributions for *Buchnera*, as well as the *Wigglesworthia* and *Candidatus*
25 *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.

5 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 (Table S4). Moreover, short branches have slightly higher dN/dS than longer branches (Figure S4), because purifying selection has had less time to purge deleterious mutations from the population (Rocha et al,
10 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 in order to be incorporated into an internal branch. Because S:F is not inflated in short leaf-branches (Figure S4), it appears that neither sequencing errors nor unpurged deleterious polymorphisms present a major source of bias in our results. Yet errors
15 in ancestral reconstruction may significantly bias S:F estimates in internal branches, and although they may be less biased than dN/dS (Table S4), S:F in internal branches should still be interpreted cautiously.

Selection example I: Redox metabolism in pseudomonads

20 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
25 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 *E. 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, Price-Whelan et al, 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 3 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→Tyr58, is in contact with a bound cardiolipin phospholipid (Yankovskaya et al, 2003), while the other two, Ala→Gly24 and Ile→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→Phe28 substitution, of 16 total *Pseudomonas* strains with SdhC 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 modification 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 (Price-Whelan et al, 2006, Mavrodi 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.

20

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

25

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 up-regulated 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 6 *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 & 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 pre-defined non-synonymous/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 (e.g. 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 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 (*e.g.* 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 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.

15

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.

20 Meanwhile, proteins involved in 'housekeeping' 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 gratefully acknowledge 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 U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contractDE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. 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**:1015-22
- 5 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
- 10 Clamp M, Cuff J, Searle SM and Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics* **20**:426-427
- Das M, Chopra AK, Cantu JM and 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
- 15 DeLano WL (2002) The PyMOL Molecular Graphics System. <http://www.pymol.org/>
- Falush D and Bowden R (2006) Genome-wide association mapping in bacteria? *Trends Microbiol.* **14**:353-355
- Fry AJ and Wernegreen JJ (2005) The roles of positive and negative selection in the molecular evolution of insect endosymbionts. *Gene* **355**:1-10
- 20 Guindon S and 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 and 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 and 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
- 25 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 and 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
- 30 Hughes AL, Friedman R, Rivallier P and 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
- 35 Hughes AL and 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 and Nei M (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proceedings of the National Academy of Sciences* **99**:12944-12948

- Jordan IK, Kondrashov FA, Rogozin IB, Tatusov RL, Wolf YI and Koonin EV (2001) Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol.* **2**:RESEARCH0053
- 5 Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**:275-276
- Kryazhimskiy S and Plotkin JB (2008) The Population Genetics of dN/dS. *PLoS Genet.* **4**:e1000304
- Li YF, Costello JC, Holloway AK and Hahn MW (2008) "Reverse ecology" and the power of population genomics. *Evolution* **62**: 2984-2994
- 10 Ma W, Dong FF, Stavrinos J and 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 and Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**:1753-1762
- Mavrodi DV, Blankenfeldt W and Thomashow LS (2006) Phenazine compounds in fluorescent *Pseudomonas* spp. biosynthesis and regulation. *Annu.Rev.Phytopathol.* **44**:417-445
- 15 Moran NA (1996) Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences* **93**:2873-2878
- Nei M and Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol.Biol.Evol.* **3**:418-426
- Nei M and Kumar S (2000) *Molecular Evolution and Phylogenetics*, Oxford University Press, Oxford, UK
- 20 Nozawa M, Suzuki Y and Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences* (doi: 10.1073/pnas.0901855106, Early edition)
- Ochman H, Elwyn S and Moran NA (1999) Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* **96**:12638-43
- 25 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 and Alm EJ (2006) OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC Bioinformatics* **7**:19
- 30 Price-Whelan A, Dietrich LE and 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 and Newman DK (2006) Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nat.Chem.Biol.* **2**:71-78
- Pupko T, Pe'er I, Shamir R and Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol.Biol.Evol.* **17**:890-896
- 35 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**:226-35
- 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
- 40 Shapiro BJ and Alm EJ (2008) Comparing Patterns of Natural Selection across Species Using Selective Signatures. *PLoS Genet.* **4**:e23

- Shapiro BJ, David LA, Friedman J and Alm EJ (2009) Looking for Darwin's footprints in the microbial world. *Trends Microbiol* **17:5** (in press).
- 5 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 and Tibshirani R (2003) Statistical significance for genomewide studies. *Proc.Natl.Acad.Sci.U.S.A.* **100**:9440-9445
- Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science* **278**:631-637
- 10 Weber E and Koebnik R (2006) Positive selection of the Hrp pilin HrpE of the plant pathogen *Xanthomonas*. *J.Bacteriol.* **188**:1405-1410
- Woolfit M and Bromham L (2003) Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol.Biol.Evol.* **20**:1545-1555
- 15 Xu C, Wang S, Ren H, Lin X, Wu L and 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 and 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**:568-73
- 20 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
- 25 Zeng K, Fu YX, Shi S and Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**:1431-1439

Titles and legends to figures

Figure 1 - Overview of S:F methodology.

(A) Hypothetical 5-species phylogeny and multiple sequence alignment for a protein of 6 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 2 sites in the slowest category (1 sub/site), 2 sites in the fastest category (3 subs/site), and 1 site in an intermediate category (2 subs/site). If the cutoff (k) were drawn such that the intermediate category is classed as 'fast', species 1 would have 1 substitution in 2 slow-sites and 3 substitutions in 3 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, $FDR = 0.05$, meaning that of the 100 branches with significantly unusual S:F at the $p < 0.05$ confidence level, 5 are expected to be false positives) is chosen as the optimal cutoff.

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). 8
branches (*V. cholerae*, *V. vulnificus*, *X. oryzae*, *P. syringae*, *S. oneidensis*, *P. multocida*, *E.*
coli, *B. aphidicola* APS) under selection (black bars). Clade of 4 species (*P. syringae*, *P.*
5 *aeruginosa*, *P. fluorescens*, *P. putida*) under selection (dark-grey 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 non-target branches in the tree. Each bar shows the mean S:F in the target branch(es) for
100 replicate simulations, with error bars showing +/- the standard error of the mean. When
multiple branches are targeted, a single representative branch is displayed, chosen at random.

10

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 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 over-represented in the high-S:F set of
15 genes are colored in maroon, and those under-represented in blue, with color saturation
proportional to the significance of the Hypergeometric test for enrichment/depletion. The
results are repeated using 5 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)
20 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.

(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 (Supp. Methods). The number of genes in each branch for which S:F
25 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).

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 grey on the tree). Data is 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) 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%. Grey: 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 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.

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 (grey). 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); MSA image using Jalview (Clamp et al, 2004).
- (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 grey text.

20

Tables

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

Method	<i>k</i>	slow (S)		fast (F)		S:F		Fisher Test	
		# subs.	# sites	# subs.	# sites	Ratio [^]	Rank	O.R.	<i>p</i>
AA <i>minFDR</i>	0.55	5	66	3	37	0.93	2%*	2.42	0.28
AA <i>minSD</i>	0.35	3	30	5	73	1.50	1%*	5.26	0.043
DNA <i>minFDR</i>	0.75	17	275	23	88	0.20	61%	0.82	0.63
DNA <i>k=0.67</i>	0.67	8	228	32	135	0.13	55%	0.59	0.22
dN/dS NG86	n/a	dN = 0.042		dS = 0.39		0.11	37%	n/a	
dN/dS PAML	n/a	dN = 0.042		dS = 0.29		0.14	27%	n/a	

[^] Corrected for multiple substitutions

* S:F ratio is among the top 10% highest in the genome

5

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. 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 (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 Test column, O.R. >1 indicate that the S:F ratio is greater in *P. fluorescens* than other branches of the gene tree.

10

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

Method	<i>k</i>	slow (S)		fast (F)		S:F		Fisher Test	
		# subs.	# sites	# subs.	# sites	Ratio [^]	Rank	O.R.	<i>p</i>
AA <i>minFDR</i>	0.55	6	75	6	49	0.63	1.5%*	1.71	0.38
AA <i>minSD</i>	0.35	2	16	10	108	1.38	0.1%*	8.25	0.037
DNA <i>minFDR</i>	0.75	34	360	86	135	0.07	25%	0.46	0.00015
DNA <i>k=0.67</i>	0.67	25	325	95	170	0.08	49%	0.46	0.00056
dN/dS NG86	n/a	dN = 0.07		dS > 0.75**		< 0.093	n/a**	n/a	
dN/dS PAML	n/a	dN = 0.08		dS = 3.70		< 0.022	42%	n/a	

[^] Corrected for multiple substitutions

* S:F ratio is among the top 10% highest in the genome

** saturated in NG86, rank & ratio are only approximate

15

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.

20

Supplementary information available at the ISME website.

Supplementary Data Set 1 – S:F ratios in the γ -proteobacteria dataset (*minSD* method to estimate *k*)

S:F ratios (corrected for multiple substitutions), Odds Ratios, and *p*-values are shown for
5 each of 917 COGs in each branch of the species-tree. The numbers of substitutions in both
fast and slow-sites (uncorrected) are also shown in the table. Internal branches are named as
the concatenation of all the terminal branches they encompass.

Supplementary Data Set 2 – S:F ratios in the γ -proteobacteria dataset (*minFDR* method to estimate *k*)

10 Same as Supplementary Data Set 1, except with *k* estimated with the *minFDR* method.

Supplementary Table 1 – List and description of COGs used in the study

COG ID numbers, functional category (one-letter code), brief description, and short names
are given for each of the 917 COGs used in the study.

Supplementary Table 2 – List of species used in the study

15 NCBI/MicrobesOnline taxonomy IDs, species names, and abbreviated names are given for
each of the 30 species used in the study.

Supplementary Table 3 – Pearson's correlations between dN/dS and S:F

Included in Supplementary Information.

**Supplementary Table 4 – Comparison of Internal and Terminal branch lengths, S:F,
20 and dN/dS ratios**

See Supplementary Notes for detailed figure legend.

Supplementary Figure 1 – Finding an optimal cutoff (*k*) between fast and slow-sites

See Supplementary Notes for detailed figure legend.

Supplementary Figure 2 – Agreement between dN/dS and S:F applied to codons

See Supplementary Notes for detailed figure legend.

5 **Supplementary Figure 3 – Additional Fisher *p*-value filter on Enrichment/Depletion of Cellular Functions among high-S:F subset of genes**

See Supplementary Notes for detailed figure legend.

Supplementary Figure 4 – Relationships between internal and terminal branch lengths, S:F and dN/dS

10 See Supplementary Notes for detailed figure legend.

Supplementary Information

Supplementary Notes 1-3, Supplementary Figure legends 1-4, Table legend S4, and Table S3.

Supplementary Methods

15 Methods to construct gene- and species-trees, estimate dN/dS and divergence times, simulate sequences, and set the cutoff (*k*) between Slow and Fast-sites