Structural Calibration of the Rates of Amino Acid Evolution in a Search for Darwin in Drifting Biological Systems

Christina Toft[13] and Mario A. Fares[12*]

[1]Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, University of Dublin, Trinity College, Dublin, Ireland

[2]Integrative Systems Biology Group, Institute of Molecular and Cellular Biology (CSIC-Universidad Politécnica de Valencia (UPV)), Valencia, Spain.

[3]Current address: Department of Evolution, Genomics and Systematics, Uppsala University, Uppsala, Sweden.

*Correspondence Author: Mario A. Fares

Address: Department of Genetics, University of Dublin, Trinity College, Dublin 2, Dublin, Ireland.

Institute of Molecular and Cellular Biology (CSIC-Universidad Politécnica de Valencia (UPV)), Valencia, Spain.

Phone number: 34 1 3879934

Email address: mfares@ibmcp.upv.es

Keywords: Darwinian Evolution, Natural Selection, Endosymbiotic Bacteria, Calibrated Evolution

ABSTRACT

In the last two decades, many reports of proteins under positive selection have brought the neutral theory into question. However, the methods used to detect selection have ignored the evolvability of amino acids within proteins, which is fundamental to distinguishing positive selection from the relaxed constraints caused by genetic drift. Disentangling these two counterbalancing forces is essential to test the Neutral theory. Here, we calibrate rates of amino acid divergence by using structural information from the full set of crystallised proteins in bacteria. In agreement with previous reports, we show that rates of amino acid evolution correlate negatively with the number of per-amino acid atomic interactions. Calibration of the rates of evolution allows identifying signatures of selection in biological systems that evolve under strong genetic drift, such as endosymbiotic bacteria. Application of this method identifies different rates and evolutionary dynamics of evolution for highly-connected amino acids in the structure compared to sparsely-connected ones. We also unearth patterns of Darwinian selection in fundamental cellular proteins in endosymbiotic bacteria including the co-chaperonin GroES, ribosomal proteins, proteins involved in cell cycle control, DNA-binding proteins and proteins involved in DNA replication and repair. This is, to our knowledge, the first attempt to distinguish adaptive evolution from relaxed constraints in biological systems under genetic drift.

INTRODUCTION

"That natural selection will always act with extreme slowness I fully admit" - The Origin of Species (1857). With this statement, Darwin acknowledged the relative slowness of natural selection. This remark remains generally true, although exceptions highlight the occasionally saltational nature of evolution. For example, protein evolutionary rates usually shift between groups of organisms with different biological properties or distinct population dynamics. These shifts are generally correlated with functional changes in proteins that enable organisms to adapt to new environments. However, changes in environmental conditions are also accompanied by non-selective processes induced by genetic drift such as gene loss or accelerated rates of evolution in no-longer important (functionally redundant) genes (16). Although many studies have been conducted to identify adaptive evolution, little effort has been invested in disentangling the variation in protein's evolutionary rate caused by adaptive processes from that due to genetic drift. Resolving this question is fundamental to shedding light on the evolvability of proteins and hence in evaluating the potential of proteins to generate functional innovation.

Different methods based on the ratio between the rates of non-synonymous and synonymous nucleotide substitutions ($\omega = d_N/d_S$) have been devised to identify selection indiscriminately in all types of biological systems. Although such conservative measures of the intensity of selection can in theory distinguish between adaptive evolution ($\omega > 1$), neutrality ($\omega = 1$) and purifying selection ($\omega < 1$), they are subject to limitations. In particular, these methods systematically ignore the capacity of amino acids to evolve given their structural constraints. For example, the high conservation of amino acids that are confined to the core of a protein makes it difficult to identify punctual events of adaptive evolution using ω because such events are generally masked by strong purifying selection that operates at these sites during most of their evolution. Conversely, residues with little functional importance can show high rates of amino acid substitutions that, when the number of synonymous changes is small, can lead to inflated ω values erroneously supporting adaptive processes. Rather than being an exception, this problem underlies many biological systems that present complex population dynamics resulting from a change in their lifestyle. Accurate identification of selection in these biological systems may shed light on the molecular mechanisms enabling ecological innovation. Symbiosis is one of the most striking examples of biological innovation. Indeed, insects that have established symbiosis with proteobacteria have colonized a myriad of ecological niches, leading to the astonishing diversity of insect species (Price 1991). *Buchnera aphidicola*, the primary symbiotic bacterium of aphids, and *Candidatus Blochmannia sp*, the primary symbiotic bacterium of carpenter ants, are among the best-characterized endo-cellular symbiotic bacteria of insects. These bacteria experience strong population bottlenecks (Mira and Moran 2002) and they lack DNA-repair genes (Shigenobu et al. 2000; Moran, McCutcheon, and Nakabachi 2008). In the case of aphids, their endosymbiotic bacteria contain no prophages, a single rRNA operon, no long repeated sequences, and have lost genes involved in recombination or incorporation of foreign DNA (Suyama and Bork 2001; Tamas et al. 2002). This has resulted in the rapid fixation of slightly deleterious mutations by genetic drift, as has been extensively demonstrated (Lynch and Gabriel 1990; Moran 1996; Lynch 1997; Fares et al. 2002b). The average mutational load of endosymbiotic bacteria therefore increases between host generations in an irreversible fashion, a phenomenon previously recognized as an example of Muller's ratchet (Muller 1964). The strong effect of genetic drift in these bacteria makes it difficult to identify positively selected molecular changes that may have been beneficial for adaptation to the endo-cellular lifestyle.

The extent to which genetic drift dominates the selection-drift balance during the evolution of endosymbiotic bacteria remains the subject of many studies. Conflicting reports show either the action of selection (Fares et al. 2002a; Fares, Moya, and Barrio 2005) or genetic drift (Funk, Wernegreen, and Moran 2001; Herbeck et al. 2003) to be the main driving force in the fixation of mutations. Further, novel approaches have shown that selection for proteins robust to mistranslation errors as well as functional divergence have been operating throughout the evolution of endosymbiotic bacteria (Toft and Fares 2009; Toft, Williams, and Fares 2009). The fact that adaptive evolution and genetic drift both increase the rate of fixation of mutations makes identifying real patterns of selection in these endosymbiotic systems extremely challenging. The identification of the signature of selection is also hindered by the codon and nucleotide bias in endosymbiotic bacteria (Rispe et al. 2004) and by the saturation of synonymous sites or the action of selection at these sites, which may inflate the non-synonymous-to-synonymous rate ratio (Mayrose et al. 2007). In this manuscript we address the problem of identifying selection in systems that present high mutational loads by using a novel approach in which selection signatures are corrected by calibrating the rate of evolution at the molecular level. We show that: *i*) Amino acids within proteins are quantitatively constrained by their inter-residue interactions; *ii*) Amino acid mutational dynamics are determined by their location in the protein structure; *iii*)

Accounting for the evolvability of amino acids in the context of structures allows a more precise identification of Darwinian Selection; and *iv*) Endosymbiotic bacteria present evidence of strong selection despite the genetic drift underlying their population dynamics.

MATERIALS AND METHODS

Genomes and Alignments

We used four genomes of *Buchnera aphidicola* (hereafter *Buchnera*), the primary symbiotic bacterium of aphids, including strains *Acyrthosiphon pisum* (*BAp*: NC_002528), *Schizaphis graminum* (*BSg*: NC_004061), *Baizongia pistaciae* (*BBp*: NC_004545) and *Cinara cedri* (*BCc*: NC_008513). We did not use the two recently sequenced *BAp* genomes (Moran, McLaughlin, and Sorek 2009) to avoid biased results due to over-representation of one of the endosymbiotic genomes. In addition, these two new *BAp* genomes only represent intra-population variability at the very recent time scale while we were more interested in the variability at the species level. We also used the genomes of the endosymbiotic bacterium of carpenter ants *Blochmannia*, including *Candidatus Blochmannia pennsylvanicus* (*Bp*: NC_007292) and *Candidatus Blochmannia floridanus* (*Bf*: NC_005061). We used 85 complete genomes of gamma-3 proteobacteria to compare the evolutionary dynamics between free-living and endosymbiotic bacteria (see table 1 of Supplementary Information). These 85 genomes were those containing orthologs for the genes present in the endosymbiotic bacteria studied in this work. With each of the genes from the genomes of endosymbiotic bacteria we performed Blast searches to find their homologs in other genomes. We considered homologous genes only those showing reciprocal top best Blast hits with scores of less than or equal to 10^{-4}. For each one of the genes we built protein multiple sequence alignments using the ClustalW program with the default parameters (Thompson, Higgins, and Gibson 1994).

Protein structures

All available protein structures for gamma-3 proteobacteria (mainly in *Escherichia coli*: *Ec*) were downloaded from the protein data bank (PDB: http://www.rcsb.org/pdb/). In total, we downloaded 1,000 (but 1075 different chains) structures (Accession numbers in Table 4 of Supplementary Information). In those protein structures with several chains or subunits we blasted each of the subunit amino acid sequences against the representative sequence in gamma-3 proteobacteria and stored that chain of the protein structure sequence with the highest score. These structures represented approximately 20% to 25% of the proteome of *Ec*. This set of structures also contained 221 proteins in *Buchnera* (about 50% to 80% of the remaining genes in *BAp* and *BCc*, respectively) and 335 proteins from *Blochmannia* (constituting about 50% to 60% of the proteins set conserved in *Blochmannia* genomes). None of the main functional categories contained in the cluster of orthologous groups (COG), were over-represented (we found no significant enrichment of any of the categories for any of the structures, and the proportion of represented proteins in the three main categories varied only between 28% and 37%). Despite this, we found significant differences in the rates of evolution of crystallized proteins versus non-crystallized ones in *Buchnera* (T-test of independent samples: t = 3.319, *d.f.* = 156, P = 0.001) and in *Blochmannia* (t = 2.470, *d.f.* = 454, P = 0.013), with crystallized proteins being the most variable. Highly conserved proteins therefore do not bias our analyses, and we have a sufficient amount of evolutionary noise in our proteins so as to put our approach to the test.

Calibrating Natural Selection Using Structural Constraints

Amino acids within a protein do not contribute equally to protein's function or structure and are therefore under different constraints. It has been shown that, in general, amino acid residues exposed to the surface of the protein evolve faster than those within the core of the protein (Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1998; Bustamante, Townsend, and Hartl 2000; Mintseris and Weng 2005; Bloom et al. 2006b; Conant and Stadler 2009). Here, we explore the relationship between the number of amino acid atomic contacts and their evolutionary divergence. We then use this information to calibrate the rates of amino acid evolution and identify those residues that present selection signals beyond the stochastic error caused by genetic drift and beyond the intrinsic evolvability of amino acids. We hypothesise that amino acids with greater number of contacts are structurally more constrained because their mutation is likely to affect many residues in the protein, and that they therefore evolve slowly. This means that, under genetic drift, amino acids with low number of contacts may present artifactual signals of selection and accelerated rates due to relaxed constraints rather than to adaptive processes. To demonstrate the precise relationship between structural constraints and divergence, we subdivided protein alignments from gamma-3 proteobacteria into different divergence categories. We did that by estimating the average amino acid variability for each of the proteins in *Ec*. These categories were named 10%, 20%, 30% and 40% and comprised proteins with 0 to 10% average pairwise protein sequence divergence, 11 to 20% divergence, 21 to 30% divergence and 31 to 40% divergence, respectively. We then measured the number of amino acid interactions with other residues (hereafter called amino acid density) within each of the proteins by measuring the structural distance among all pairs of amino acid sites using the Euclidean distance:

$$d=\frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\sqrt{\left(X_i-X_j\right)^2+\left(Y_i-Y_j\right)^2+\left(Z_i-Z_j\right)^2}$$

Here *N* is the number of atoms in amino acid *i* while *K* is that number in amino acid *j*. X, Y and Z represent the three-dimensional coordinates of the atoms corresponding to each of the amino acids. We considered two amino acids to contact each other when the distance between their closest atoms was 4Å.

We categorized amino acid densities into classes that included interactions, 3 to 5, 6 to 8, 9 to 11, 12 to 14, 15 to 17 and 18 to 20 interactions. We determined the classes and their limits using Stutgarts equation. For each of the sets of proteins with divergence belonging to one of the levels (10% to 40%), we had 7 categories of amino acid densities. The number of data per category is summarized in supplementary table 3 of supplementary information. Finally, we subdivided protein sequence alignments into sub-alignments comprising amino acid sites that shared similar amino acid density – that is, we picked from the alignment those amino acids with similar densities and built a sub-alignment containing them - and stored such sub-alignments in the different density categories. This was further performed in the case of alignments including endosymbiotic sequences that were classified into the categories of divergence levels according to the divergence of their orthologous free-living bacterial alignments. That is to say, within each of the 28 categories endosymbiont sub-alignments are a subset of free-living sub-alignments that are within the same category, so that the proteins within each of the categories in free-living bacteria are being compared with the same proteins in endosymbiotic bacteria.

For each one of the 28 categories of protein sequence sub-alignments we measured the average pairwise amino acid distance corrected for multiple hits using Poisson. Then we plotted Poisson-corrected distances against the classes of amino acid densities. This allowed us to construct curves that provided a background protein evolutionary divergence against which we could compare individual sub-alignments. Using this evolutionary background signal we tested and identified signatures of selective constraint deviating from expectation at individual amino acid site categories within proteins.

Identifying Calibrated Selection Constraints in Endosymbiotic Proteins

We built multiple sequence alignments for all those endosymbiotic proteins for which a crystal structure is available. The final set of proteins for which we built alignments consisted of 221 proteins for *Buchnera* and 335 proteins for *Blochmannia*. After classifying proteins into the different divergence levels according to the divergence of their orthologs in free-living bacteria, we classified amino acid sites within these alignments into the different density categories and measured their evolutionary divergence as the pairwise Poisson-corrected amino acid distance in the multiple sequence alignment at that site.

RESULTS

The number of residue interactions constrains amino acid evolution

Does the numbers of residue interactions constrain amino acid evolution? Although many other authors have partially addressed this question, the quantitative relationship between these two parameters remains unexplored. We first examined the dynamics of evolutionary rates in free-living bacteria (the 85 genomes from the gamma-3 proteobacterium group). To carry out these analyses, we first calculated the structural density for each amino acid in *Ec* protein structures. Then we classified amino acids within bins of densities (see Material and Methods for details) and measured their evolutionary divergence using the average Poisson-corrected pairwise amino acid distances. We therefore calculated two numbers for each amino acid: structural density and mean evolutionary divergence. As predicted, our analysis of the full set of crystallized proteins demonstrated a negative correlation between amino acid density and evolutionary rate: the greater the amino acid density of a residue, the lower is its divergence (Figure 1A). This relationship is reproducible regardless of the divergence levels of the proteins considered (Figure 1A), and it remains true in *Buchnera* and *Blochmannia* (Figure 1B and C) despite the effects of genetic drift. We repeated this analysis controlling for the overall divergence level of the proteins (that is, we analyzed the relationship between divergence measured as mean Poisson distance and amino acid density in the structure for the set of proteins with average divergence levels of 10%, 20%, 30%). Correlation coefficients were low due to the high number of data points, although they were highly significant in all the correlation curves of Figure 1 (Table 1). Unlike free-living bacteria, the variance of the relationship between the structural density of amino acids and their divergence in endosymbiotic bacteria is better explained by a curvilinear (quadratic) rather than linear model (Figure 1A and B). In this model, the slopes of the curves are only slightly negative at amino acids presenting low numbers of interactions (for example, low amino acid density) and increase sharply at higher amino acid densities (Figure 1B and C). A plausible explanation for this pattern is that endosymbiotic bacterial genomes have fixed mutations neutrally at amino acids with low densities (possibly due to lower constraints) that became quickly saturated over time. This may have led to a slope that becomes insensitive to the variance of amino acid density when density values are low. This result was not biased by the codon or nucleotide composition because there was no correlation between the structural location of the amino acid and its nucleotide composition (Spearman Correlation: $r = 0.087$; $P >> 0.2$).

We investigated the relationship of the evolutionary dynamics and structural constraints in proteins from the different functional categories of the Clusters of Orthologous Genes (COG). We ran the same analyses on the set of genes that were classified within the categories of metabolism (met), information storage and processing (isp) and cellular processes and signalling (cps) (Table 4 of Supplementary Information). Free-living bacteria presented the same dynamics in each of the functional categories, with amino acid densities correlating negatively with their divergence (Figure 2A). The variance between these functional categories was very low. Indeed, the boundaries between the divergence levels were clear and the difference in divergence among protein divergence curves remained significant, such that there was no significant overlap between error bars at different divergence levels for the three COG categories (Figure 2B). In the case of the endosymbiotic bacteria *Buchnera* and *Blochmannia*, the results were slightly different because, despite the reproducibility of negative correlation between rates and amino acid densities in all COG categories, the high variance between the curves resulted in the overlap of divergence rates among the different protein divergence levels (Figure 2C to F). This supports the fact that amino acid sites in these systems have become saturated by replacements. Interestingly, the *met* comprised the most highly evolving proteins, while *isp* compiled the slowest evolving proteins in *Buchnera* and *Blochmannia* (Figure 2C to F). The same pattern was not observed in free-living bacteria, where the amount of contributed variation within each divergence category varied randomly among the COG classes (Figure 2A and B and Table 4 of Supplementary Information). It was also noteworthy that, in *Buchnera* and *Blochmannia*, residues with the highest amino acid densities presented substantially lower divergence levels in the 30% divergence category compared to the 10% and 20% divergence categories (Figure 2C to F). In other words, amino acids in the protein core seemed more constrained in proteins with 30% divergence than in those with 10% or 20% divergence. This may indicate functional divergence at these amino acids followed by strong constraints to preserve their new functional roles, as previously suggested (Toft, Williams, and Fares 2009).

Heterogeneous evolutionary dynamics of amino acids across protein structures

In the previous section we showed that amino acid sites are constrained by the number of interactions they establish. Here we sought to determine whether the mode of evolution of amino acid sites with high structural densities is equivalent to that of surface-exposed amino acids despite their different divergence levels. We addressed this in free-living bacteria and endosymbiotic bacteria separately because they experience different population structures and selection patterns. To test the correlation between the mode of evolution and the structural density we studied possible changes throughout the phylogeny in the evolutionary dynamics of amino acids falling within particular density categories. We examined whether the average evolutionary rate of amino acids belonging to one density category changes between different phylogenetic levels (that is, we compared the rates of evolution estimated from closely-related organisms to those obtained from distantly related organisms in the same phylogeny) and whether these changes (which we call Phylogenetic Selection Shifts, PhSS) are of the same magnitude in the different density categories. To compare PhSS in evolutionary rates between density categories, we calculated the increments in evolutionary divergence across the phylogeny for each density category. Due to the fact that for some genes, distance between the two sister taxa can be greater than the mean tree divergence, we assumed both these divergence levels follow a binomial distribution. Therefore, the increment of divergence for each of the amino acids within a sub-alignment and density category was:

$$\Delta D = 1 - \left(\frac{\left| \overline{P}_{tree} - P_{Ec-St} \right|}{\overline{P}_{tree} + P_{Ec-St}} \right)$$

Here *ΔD* is the increment of the Poisson-corrected distance for each amino acid within a particular amino acid density category, $\overline{P}_{tree}$ is the average pairwise Poisson-corrected amino acid distance for the entire free-living bacterial tree for that amino acid, while $P_{Ec\text{-}St}$ is that distance between *Escherichia coli* (*Ec*) and

Salmonella typhimurium (*St*). It is noteworthy that in the tree *Ec* and *St* are the closest possible species and hence P_{tree} should be greater than $P_{Ec\text{-}St}$. Accelerated evolution between *Ec* and *St* in a particular density category will imply that $P_{Ec\text{-}St}$ approaches $\overline{P}_{tree}$, the ratio will decrease, and hence ΔD will approach 1. The inverse situation will lead to increments closer to 0. We then conducted the same approximation in *Buchnera* and calculated the ratio for $P_{BAp\text{-}BSg}$ (the closest endosymbiotic bacterial species in the tree) and the average distance for the four *Buchnera* genomes along the phylogeny, $P_{Buchnera}$. The results in figure 3 show that free-living bacteria present regression models with changing slopes, going from a positive slope in the 10% and 20% divergence curves to a negative slope in the 30% divergence curve (Figure 3A and Table 2). At low protein divergence levels, therefore, proteins behave as expected: amino acids with high structural densities evolve slowly regardless of the phylogenetic level (for example, they present similar number of fixed mutations when we compare closely related species or distant species, respectively) and hence the ratio is close to 0 and ΔD close to 1. The inverse situation occurs with amino acids at low densities. At high protein divergence levels, such as the curve of 30% divergence, however, it is the amino acids with high densities that display greater shifts in their divergence rates, yielding ΔD values closer to 0 (Figure 3A). In contrast, the *Buchnera* curves all had positive slopes, indicating that when the divergence level between species increases, the residues with high amino acid densities evolve proportionally slower than those with low densities (Figure 3B). The results obtained for *Buchnera* were also reproduced in *Blochmannia* (results not shown). These results lead to two conclusions: *i*) that amino acids experience different evolutionary dynamics depending on the structural constraints; and *ii*) that despite their population dynamics, endosymbiotic bacteria display evolutionary patterns strongly suggesting the action of natural selection at highly constrained amino acids.

Identifying calibrated Darwinian selection in endosymbionts

We used the evolutionary divergence curves as a starting point for identifying selection in endosymbiotic bacteria. To do so, we subdivided the multiple sequence alignments for each of the proteins for which a crystal structure was available into different sub-alignments that each contained residues with similar amino acid densities. We then estimated the average Poisson distance for these sub-alignments and compared this distance with that estimated for the full dataset (full dataset as available in Figure 1). The comparison was done using the divergence level curve appropriate to each protein (for instance, we compared the sub-alignments derived from proteins at the 10% divergence level to the curve of the proteins with 10% divergence levels in Figure 1). Those sub-alignments falling within the curve and its confidence interval were considered to evolve neutrally. Outliers were considered to be proteins that were evolving either under adaptive/accelerated evolution (if falling above the curve) or under strong selective constraints (falling below the curve). This is a very convenient way of identifying selection because it is not subject to the assumptions of parameters used to measure selection, such as neutrality at synonymous sites or lack of codon bias, which are mostly violated in the case of endosymbiotic bacteria. This approach identified several outlier proteins, some of which presented one category of amino acid density under either strong constraints or accelerated evolution, while other proteins included residues from different categories under distinct constraints (Figure 4). We subdivided the proteins identified as outliers into those showing unexpected evolutionary patterns at amino acids with high structural densities and those with low structural densities (Table 3). We identified accelerated evolution in highly dense residues in proteins related to the translational and transcriptional machinery (Table 3). We also found accelerated rates of evolution in signal recognition particle receptors and excretion systems, as well as in proteins related to the metabolism of amino acids (Figure 4 and Table 3). Importantly, these proteins mediate the endosymbiotic lifestyle of these bacteria, which experience low replication rates and synthesize essential amino acids that are later exported to the host (Gray, Burger, and Lang 1999; Shigenobu et al. 2000). We also detected accelerated evolution at amino acids with low densities (therefore exposed on the surface of the protein) in proteins from the ribosomal system and in many enzymes and nucleotide binding proteins (Figure 4 and Table 3). This hints at the possibility that this acceleration may be affecting the ability of these proteins to bind other proteins or nucleotides. Among the highly constrained proteins we identified ribosomal proteins and the essential chaperonin cofactor GroES that, in conjunction with the chaperonin GroEL, is involved in the protein folding cycle. Other essential proteins were also identified such as the cell division protein FTSZ, ssDNA-binding protein SSB and the mechanosensitive ion channel protein MSCS (Figure 4 and Table 3). Interestingly, while FTSZ showed constraints at buried amino acids, proteins involved in interactions with other proteins or with DNA displayed constraints at surface-exposed residues as well as at buried sites. Taken together, these results suggest new functional roles in proteins that either buffer the effects of intra-cellular life in endosymbionts, such as ameliorating the effects of destabilizing mutations through GroEL/S (Moran 1996; Fares et al. 2002a; Fares, Moya, and Barrio 2004; Fares, Moya, and Barrio 2005), or that slow down cell division in order to adapt to intra-cellular life (Toft, Williams, and Fares 2009).

We examined the distribution of the genes that showed accelerated or constrained evolution in the different categories of the cluster of orthologous groups (COG) and we identified Fast-evolving and slow-evolving genes in all three COG categories (met, isp and cps). However, unlike in other COG categories, , we did not find evidence of enrichment for constrained (or slowly-evolving) genes in met (Figure 5 and Table 5 of supplementary information). In *Blochmannia* we also identified a significant impoverishment for slow-evolving genes in met (Figure 5). Conversely, fast-evolving and slow-evolving genes were present in the other two categories in *Buchnera* and *Blochmannia* (Figure 5).

Identification of calibrated constraints in two case studies

To determine the biological significance of the information obtained from selection analyses, we focused on two essential protein-coding genes. These two genes were *infB,* which encodes the prokaryotic translation initiation factor (IF2) (Laursen et al. 2002) and *ssb*, which is the single-stranded DNA (ssDNA) binding protein. ssDNA is a transient state in DNA metabolic processes such as replication, recombination and repair (Lohman and Ferrari 1994; Wold 1997). Both proteins are essential for cell viability under normal conditions. The binding sites for SsB protein are known to include residues at positions 55 to 61 (highlighted in green in figure 6a). Analysis of selection patterns in SsB after calibrating amino acid evolvability by structural constraints identifies several amino acids as being under strong purifying selection (spheres in the structure of figure 6a). From the perspective of structural constraints, we would expect these residues to evolve more quickly than core-protein residues because they are involved in a low number of residue interactions. However, the function of these residues is essential and they are therefore expected to evolve slowly despite their low structural densities. Identification of amino acid sites with low densities and high constraints reveals that residues from the binding region as well as those surrounding it are evolving slowly (figure 6a).

In the case of the InfB protein, accelerated evolution at highly structurally constrained amino acids was detected (for example, at residues with amino acid densities of 15 or greater, figure 6b). In endosymbiotic bacteria, replication and translation are constrained by the cell space because these bacteria live within specialized host cells called bacteriocytes. These functions, therefore, may have shifted to adapt to the new space limitations. Accelerated evolution at important regions in InfB may be related with these shifts and may have been accompanied by structural modification given the location of the affected amino acids in the protein core.

Discussion

The factors governing the rate of protein evolution remain to be elucidated. This task is difficult because the evolutionary rate of a protein is the result of the interactions between many parameters, including gene expression (Jovelin and Phillips 2009), number of per-protein interactions, modularity (Fraser et al. 2002; Fraser, Wall, and Hirsh 2003; Fraser 2005) and translational robustness (Drummond et al. 2005; Drummond, Raval, and Wilke 2006). In contrast to inter-protein rate variation, the effects of expression level, translational robustness, and other interaction factors are equally felt by all amino acids within a protein. Therefore, differences in evolvability among amino acids within a protein may be due to their unequal contribution to the structural stability and/or function of that protein. The rates at which amino acids evolve vary greatly across protein structures. Epistatic interactions between amino acids belonging to interacting proteins may contribute substantially to this variation. Evidence in support of the importance of epistatic interactions between proteins in the evolution of amino acids has been previously reported (Fraser et al. 2002; Fraser, Wall, and Hirsh 2003). Nonetheless, the effects of these amino acids on general patterns of evolution should be negligible because only a very few amino acids generally participate in the interaction between proteins. In this report, we have analyzed the constraints that structure imposes on the evolvability of amino acid sites in proteins. We show that, indeed, structural constraints can be defined as the number of atomic interactions that amino acids establish within the protein. Here we assumed that amino acids within 4Å contact one another. The number of atomic contacts is negatively correlated with divergence level as we show in group 3 of the gamma-proteobacteria. The negative correlation between the solvent accessibility of amino acids and their evolutionary rate has been previously addressed (Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1998; Bustamante, Townsend, and Hartl 2000; Mintseris and Weng 2005; Bloom et al. 2006a; Conant 2009; Conant and Stadler 2009). In spite of the clear inter-dependence between these two factors, no formal correlation analyses have been carried out to quantify this relationship across protein structures. In our study we assumed that number of atomic interactions is a greater determinant of amino acid evolutionary rate than solvent accessibility – these properties are not necessarily correlated because, in small proteins, amino acids may have low solvent accessibility and yet establish a low number of atomic interactions. We also used our quantitative measure of the relationship between divergence levels and atomic interactions to build curves of evolution in order to calibrate the evolutionary constraints operating within proteins. The importance of these curves stems from the potential they confer to tease apart adaptive evolution from genetic drift in biological systems under strong drift effects. Using ω in these systems would be useless because both drift and positive selection provide similar signatures, especially at the population level. In our study we assume protein structures are conserved from *Ec* to endosymbionts, and although this is likely to be generally true for the proteins analysed here, exceptions may exist. However, the effect of any such exceptions is likely to be minimal since we expect most protein structures to be conserved. An additional limitation is that crystal structures are only static images of the real protein that fail to capture important conformational changes during the binding to cofactors or other proteins. This means that amino acids far apart in the crystal structure may contact transiently after a protein conformational change. Again, we expect this to have a negligible effect on our conclusions, because the number of amino acids involved is likely to be very limited. Indeed, if this were a general phenomenon we would expect the resulting stochastic noise to hide the patterns observed in our study and therefore our results are, overall, conservative.

The endosymbiotic bacteria of insects are often under strong bottlenecks during their intergenerational transmission, and their genomes are characterized by high AT-content and saturation of synonymous sites (Rispe et al. 2004). Proteins from these bacteria are therefore very likely to violate the assumption of neutrality at synonymous sites. Synonymous sites may also be under selection (Chamary, Parmley, and Hurst 2006; Mayrose et al. 2007). This poses difficulties in using ω as a measure of the intensity of selection in these systems. Using calibrating curves we identified several proteins, involved in fundamental processes in the cell, which present evidence of selection beyond the background noise of slightly deleterious fixed mutations. We also showed that not only can amino acids evolve under different rates but that they can also present different mutational dynamics depending on their structural location. For instance, different structural regions of the GroEL protein are evolving under different constraints, so that the entire protein experiences evolutionary dynamics more complex than those previously reported (Fares et al. 2002a; Herbeck et al. 2003; Fares, Moya, and Barrio 2004; Fares, Moya, and Barrio 2005). This is a good example of the fact that even in drifting systems such as the endosymbiotic bacteria of insects, a delicate balance between selection and drift is at work in essential proteins. Only accounting for the forces shaping the evolvability of amino acids allows these signatures to be distinguished.

In summary, unlike many authors we use the link between structure and evolvability to identify signatures of selection. This allows the identification of selection signatures against a neutral background enhanced by drift. Our work demonstrates that we need to account for the structural constraints on amino acid evolvability to accurately infer selection. We propose using calibrated curves of evolution as a new approximation for identifying adaptive Darwinian selection regardless the system under study.

Acknowledgements

We thank the Evolutionary Genetics and Bioinformatics lab members for a careful review of the manuscript and useful suggestions. We especially thank Tom A. Williams for reading the manuscript and making useful suggestions on the writing. This work was supported by Science Foundation Ireland under the President of Ireland Young Researcher Award [04/YI1/M518] to MAF. Part of the research carried out in this study was supported by grant number (BFU2009-12022) from the Spanish Ministerio de Ciencia e Innovación.

References

Bloom, J. D., D. A. Drummond, F. H. Arnold, and C. O. Wilke. 2006a. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol **23**:1751-1761.

Bloom, J. D., S. T. Labthavikul, C. R. Otey, and F. H. Arnold. 2006b. Protein stability promotes evolvability. Proc Natl Acad Sci U S A **103**:5869-5874.

Bustamante, C. D., J. P. Townsend, and D. L. Hartl. 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. Mol Biol Evol **17**:301-308.

Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7**:98-108.

Conant, G. C. 2009. Neutral evolution on mammalian protein surfaces. Trends Genet **25**:377-381.

Conant, G. C., and P. F. Stadler. 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. Mol Biol Evol **26**:1155-1161.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102**:14338-14343.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23**:327-337.

Fares, M. A., E. Barrio, B. Sabater-Munoz, and A. Moya. 2002a. The evolution of the heat-shock protein GroEL from Buchnera, the primary endosymbiont of aphids, is governed by positive selection. Mol Biol Evol **19**:1162-1170.

Fares, M. A., A. Moya, and E. Barrio. 2005. Adaptive evolution in GroEL from distantly related endosymbiotic bacteria of insects. J Evol Biol **18**:651-660.

Fares, M. A., A. Moya, and E. Barrio. 2004. GroEL and the maintenance of bacterial endosymbiosis. Trends Genet **20**:413-416.

Fares, M. A., M. X. Ruiz-Gonzalez, A. Moya, S. F. Elena, and E. Barrio. 2002b. Endosymbiotic bacteria: groEL buffers against deleterious mutations. Nature **417**:398.

Fraser, H. B. 2005. Modularity and evolutionary constraint on proteins. Nat Genet **37**:351-352.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science **296**:750-752.

Fraser, H. B., D. P. Wall, and A. E. Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol **3**:11.

Funk, D. J., J. J. Wernegreen, and N. A. Moran. 2001. Intraspecific variation in symbiont genomes: bottlenecks and the aphid-buchnera association. Genetics **157**:477-489.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445-458.

Gray, M. W., G. Burger, and B. F. Lang. 1999. Mitochondrial evolution. Science **283**:1476-1481.

Herbeck, J. T., D. J. Funk, P. H. Degnan, and J. J. Wernegreen. 2003. A conservative test of genetic drift in the endosymbiotic bacterium Buchnera: slightly deleterious mutations in the chaperonin groEL. Genetics **165**:1651-1660.

Jovelin, R., and P. C. Phillips. 2009. Evolutionary rates and centrality in the yeast gene regulatory network. Genome Biol **10**:R35.

Laursen, B. S., A. S. S. A. de, J. Hedegaard, J. M. Moreno, K. K. Mortensen, and H. U. Sperling-Petersen. 2002. Structural requirements of the mRNA for intracistronic translation initiation of the enterobacterial infB gene. Genes Cells **7**:901-910.

Lohman, T. M., and M. E. Ferrari. 1994. Escherichia coli single-stranded DNA-binding protein: multiple DNA-binding modes and cooperativities. Annu Rev Biochem **63**:527-570.

Lynch, M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. Mol Biol Evol **14**:914-925.

Lynch, M., and W. Gabriel. 1990. Mutation load and the survival of small populations. Evolution **44**:1725-1737.

Mayrose, I., A. Doron-Faigenboim, E. Bacharach, and T. Pupko. 2007. Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. Bioinformatics **23**:i319-327.

Mintseris, J., and Z. Weng. 2005. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A **102**:10930-10935.

Mira, A., and N. A. Moran. 2002. Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. Microb Ecol **44**:137-143.

Moran, N. A. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A **93**:2873-2878.

Moran, N. A., J. P. McCutcheon, and A. Nakabachi. 2008. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet **42**:165-190.

Moran, N. A., H. J. McLaughlin, and R. Sorek. 2009. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science **323**:379-382.

Muller, H. J. 1964. The relation of recombination to mutational advance. Mutation Research **1**:2-9.

Price, P. W. 1991. The web of life: development of over 3.8 billion years of trophic relationships. In symbiosis as a Source of Evolutionary Innovation: Speciation and Morphogenesis. MA: MIT Press, Cambridge.

Rispe, C., F. Delmotte, R. C. van Ham, and A. Moya. 2004. Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids. Genome Res **14**:44-53.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature **407**:81-86.

Suyama, M., and P. Bork. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet **17**:10-13.

Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A. S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, and S. G. Andersson. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296**:2376-2379.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**:4673-4680.

Thorne, J. L., N. Goldman, and D. T. Jones. 1996. Combining protein evolution and secondary structure. Mol Biol Evol **13**:666-673.

Toft, C., and M. A. Fares. 2009. Selection for Translational Robustness in Buchnera aphidicola, Endosymbiotic Bacteria of Aphids. Mol Biol Evol.

Toft, C., T. A. Williams, and M. A. Fares. 2009. Genome-wide functional divergence after the symbiosis of proteobacteria with insects unraveled through a novel computational approach. PLoS Comput Biol **5**:e1000344.

Wold, M. S. 1997. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. Annu Rev Biochem **66**:61-92.

Figure Legends

Figure 1. Amino acid evolution correlates with atomic interactions. In this figure we analyze the correlation between the mean Poisson distance calculated for the amino acids and their number of atomic interactions. We calculated the number of atomic interactions for a particular residue as the number of amino acids within 4 Angstroms radius of the residue using the Euclidean distance. Correlation analyses between these two parameters were carried out for free-living bacteria (A), endosymbiotic bacteria of aphids *Buchnera aphidicola* (B) and endosymbiotic bacteria of carpenter ants *Blochmannia sp*. (C). We performed three sets of correlation analyses per organism, one for proteins with up to 10% average divergence level between orthologs (blue colour), one for proteins with divergence levels ranging from more than 10% and up to 20% (red colour) and one for proteins with levels of divergence between 20% and 30% (green colour).

Figure 2. Preserved correlation of density and divergence among functional categories. We calculated the correlation between average evolutionary rate of amino acids and the number of their contacts with other residues within the protein for each of the functional categories as represented in Cluster of Orthologous Genes (COG). We calculated the number of atomic interactions for a particular residue as the number of amino acids within 4 Angstroms radius of the residue using the Euclidean distance. We did this for the three main COG categories, including metabolism, cellular processes and information storage and processing. The analyses were carried out for free-living bacteria (A and B), endosimbiotic bacteria of aphids *Buchnera aphidicola* (C and D) and endosymbiotic bacteria of carpenter ants *Blochmannia sp.* (E and F). This analyses was done with three sets of proteins that showed divergence levels of up to 10%, between 10% and 20% divergence and between 20% and 30% divergence levels. All the different correlation analyses are colour coded.

Figure 3. Amino acid structural contacts determine their evolutionary dynamics. To identify the evolutionary dynamics at different categories of amino acid contacts, we estimated the proportional increment of evolutionary rate in free-living bacteria (A) and endosymbiotic bacteria of aphids (B) and compared it between different density categories. We estimated the density of amino acids as the number of residue contacts. We considered two residues to contact in the structure if they were within 4 Angstroms distance from one another. We calculated the proportional increments focusing in the pairs *Buchnera Acyrthosiphon pisum* (*BAp*) and Schizaphis graminum (*BSg*) for endosymbionts and *Escherichia coli* (*Ec*) and *Salmonella typhimurium* (*St*) for free-living bacteria because both these two pairs present similar divergence times (50 to 75 million years). The increment of distance in these sub-trees was calculated in comparison with the total trees for Buchnera and for free-living bacteria, respectively. Distance were calculated using Poisson correction. Numerators were transformed into the absolute values of the differences (abs). This analyses was done with three sets of proteins that showed divergence levels of up to 10% (blue), between 10% and 20% (red) divergence and between 20% and 30% (green) divergence levels. All the different correlation analyses are colour coded.

Figure 4. Identification of constraints in endosymbiotic bacteria. We calibrated selection signatures by the structural constraints and identified three types of evolutionary dynamics in endosymbiotic proteins. These included protein regions evolving under neutrality (grey cells), under accelerated evolution (red cells) and under strong purifying selection (blue cells). The distribution of the amino acids under constraints in the different atomic density categories is shown. Density categories are represented as the number of amino acids that are in contact with a residue within the structure. We estimated the distance between residues using Euclidena distance and considered two residues to contact if they were within 4 Angstroms from one another. The analyses were carried out for each of the categories of the Cluster of Orthologous Genes (COGs).

Figure 5. Enrichment analysis of functional categories for constrained amino acids. We counted the number of outliers within each of the functional categories of metabolism, information storage and processing and cellular processes and signalling. The significance of the enrichment for constrained or relaxed genes was calculated using chi-square test. We indicate enriched categories by a * ($P < 0.05$) or ** ($P < 0.01$). The different functional categories are colour coded.

Figure 6. Case study of proteins with calibrated selection constraints. We present two examples where calibration helps to elucidate the main undergoing constraints once calibrated by amino acid evolvability. (A) Analysis of the ssDNA binding protein (SSB) permits identifying strong constraints around (spheres) and in binding regions (green spheres). (B) The *infB* that encodes the prokaryotic translation initiation factor (IF2) presents strong purifying selection at amino acids with high atomic interactions (spheres). Only amino acids with spheres structure representation show significant constraints patterns once corrected by their evolvability.

Table 1. Evolutionary rates correlate with amino acids contacts density.

Divergence	Organisms	ρ_{Pearson}[a]	*t*[b]	*d.f.*[c]	Probability
10%	Free-Living	-0.191	-57.422	87,534	$< 2.2\text{x}10^{-16}$
	Buchnera	-0.183	-28.088	22,828	$< 2.2\text{x}10^{-16}$
	Blochmannia	-0.056	-10.037	31.336	$< 2.2\text{x}10^{-16}$
20%	Free-Living	-0.227	-98.777	179,775	$< 2.2\text{x}10^{-16}$
	Buchnera	-0.165	-30.118	32,268	$< 2.2\text{x}10^{-16}$
	Blochmannia	-0.060	-14.239	57,812	$< 2.2\text{x}10^{-16}$
30%	Free-Living	-0.245	-40.356	25,491	$< 2.2\text{x}10^{-16}$
	Buchnera	-0.185	-10.620	3,173	$< 2.2\text{x}10^{-16}$
	Blochmannia	-0.040	-2.563	4,705	0.010
40%	Free-Living	-0.095	-2.877	911	0.004
	Buchnera	-	-	-	-
	Blochmannia	-	-	-	-

[a] Pearson correlation coefficient was utilised, although Spearman coefficient rendered approximately the same significance levels.

[b] Student *t* value for independent samples.

[c] Degrees of freedom.

Table 2. Regression of the increments of rates of evolution over categories of amino acid contact densities.

Divergence	Amino acid Density[a]							
Free-living	5	8	11	14	17	20	R[b]	Probability
10%	0.419	0.441	0.502	0.599	0.700	0.746	0.984	0.001
20%	0.259	0.277	0.279	0.318	0.429	0.379	0.871	0.008
30%	0.282	0.244	0.202	0.189	0.224	0.148	-0.856	4.66x10^{-4}
Buchnera								
10%	0.641	0.701	0.735	0.817	0.839	0.843	0.968	1.98x10^{-5}
20%	0.616	0.634	0.669	0.711	0.763	0.816	0.988	3.45x10^{-6}
30%	0.593	0.613	0.614	0.635	0.923	0.923	0.867	0.013

[a] Amino acid density is calculated as the number of amino acids within 4 Angstroms radius of the residue. We calculated distances between amino acids using the Euclidean distance.

[b] Regression Coefficient.

Table 3. Proteins evolutionarily accelerated (A) or constrained (C) identified after structural calibration of evolutionary rates.

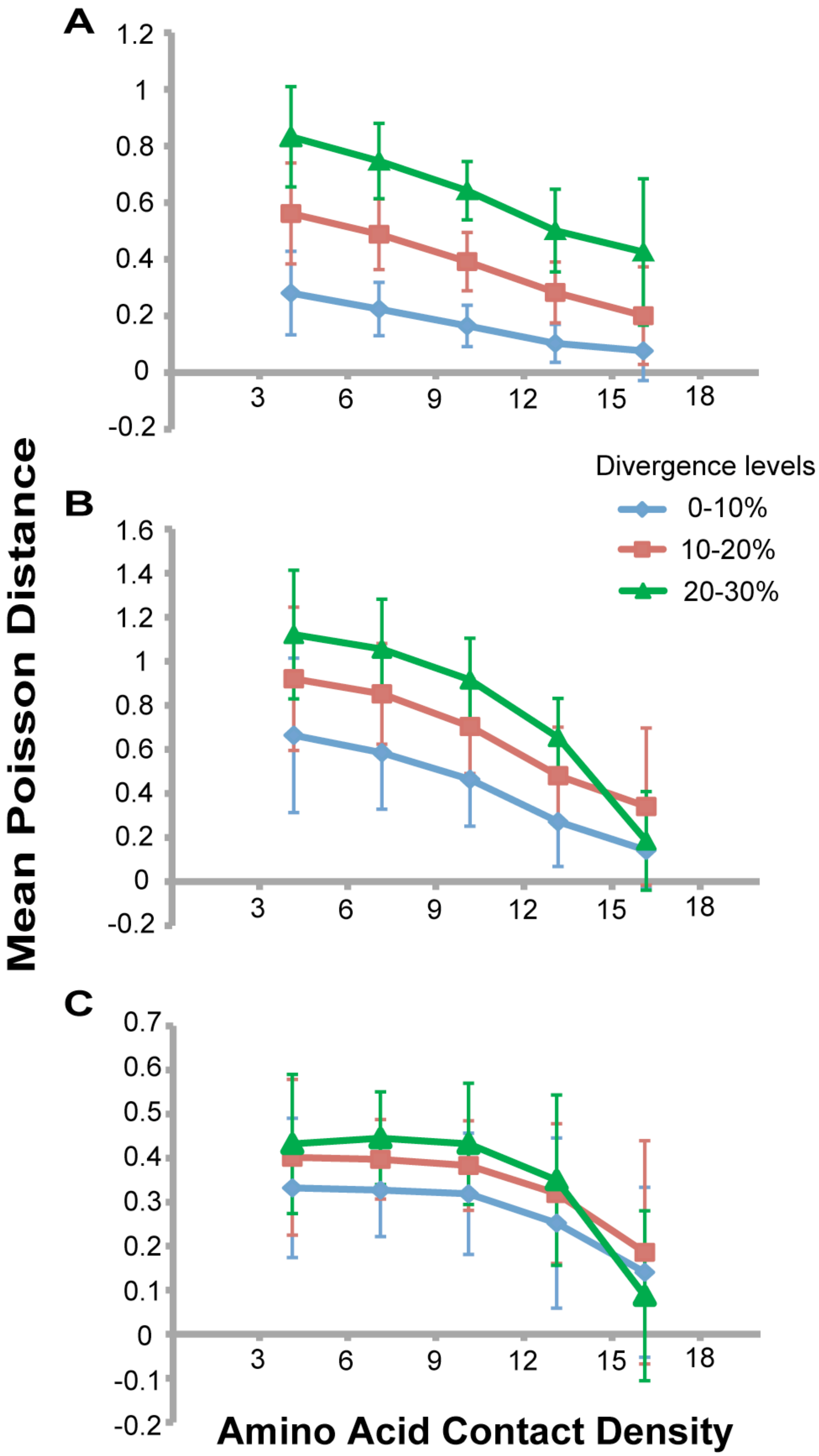
Gene[a]	Biological Process (Function)[b]	Density[c]
InfB (A)	Translation initiation factor & interacts with ribosome	H
rplM (A)	50S ribosomal subunit protein L13	H
rplL (A)	50S ribosomal subunit protein L7/L12	H
rpmB (A)	50S ribosomal subunit protein L28	H
rpmE (A)	50S ribosomal subunit protein L31	H
rpsO (A)	15S ribosomal subunit protein	L
greA (A)	Transcription elongation factor	H
dnaN (A)	Component of DNA polymerase III, β subunit	H
nfo (A)	Endonuclease IV, involved in DNA repair	L
mutY (A)	Adenine glycosylase mismatch repair enzyme	L
recC (A)	DNA helicase, recombination and repair	H
yeaZ (A)	Peptidase	L
ftsY (A)	Signal Recognition Particle Receptor	H & L
secA (A)	Component of SEC protein secretion system	H
yidC (A)	Inner membrane protein insertion factor	H
fdx (A)	Oxidized ferrodoxine	L
lpd (A)	Dihydrolipoyl dehydrogenase	H
pta (A)	Subunit of phosphatase acetyltransferase	H
talA (A)	Transaldolase, Carbohydrate metabolic process	L
iscS (A)	Cystein desulfurase, Fe-S cluster assembly	H
ilvH (A)	Regulatory subunit of acetolactate synthase III	H
thrA (A)	aspartate kinase I, K biosynthetic process	H
cyaY (A)	Protein complex assembly	H & L
rplY (C)	50S ribosomal subunit protein L25	H & L
rpmF (C)	50S ribosomal subunit protein L32	H & L
rpsB (C)	30S ribosomal subunit protein S2	L
rpsN (C)	30S ribosomal subunit protein S14	H & L

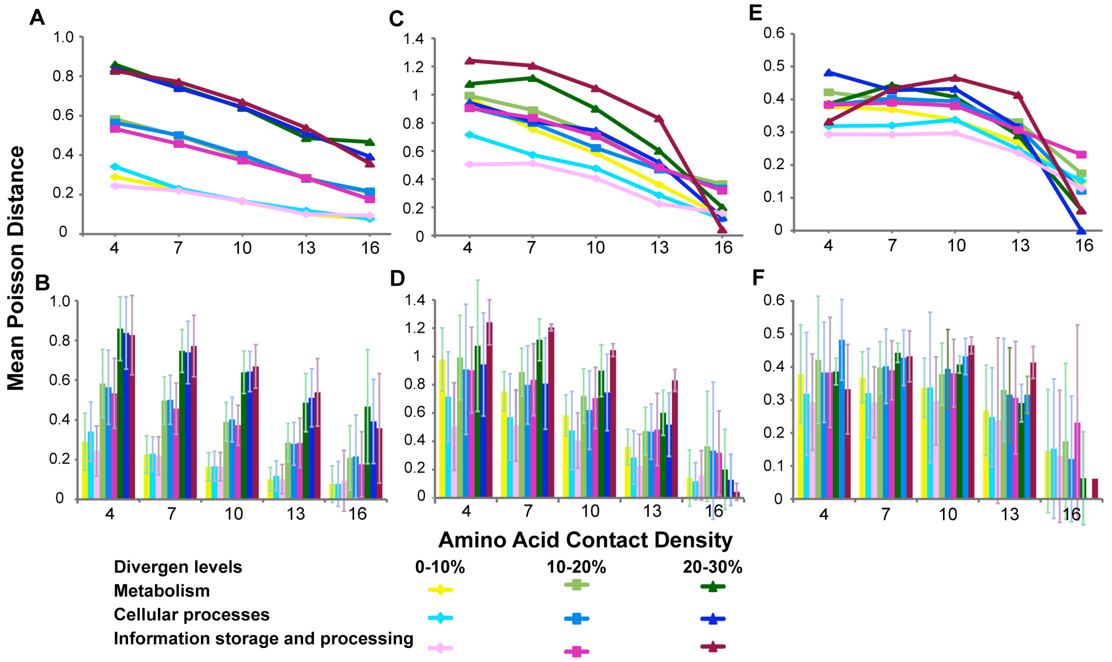
rpsT (C)	30S ribosomal subunit protein S20	L
ssb (C)	ssDNA binding protein, DNA recombination	L
ftsZ (C)	Essential cell division protein, Cytokinesis	H
groS (C)	GroEL co-chaperone, protein folding	H & L
mscS (C)	Mechanosensitive channel, cellular water homeostasis	L
era (C)	GTP binding protein, cell cycle	L

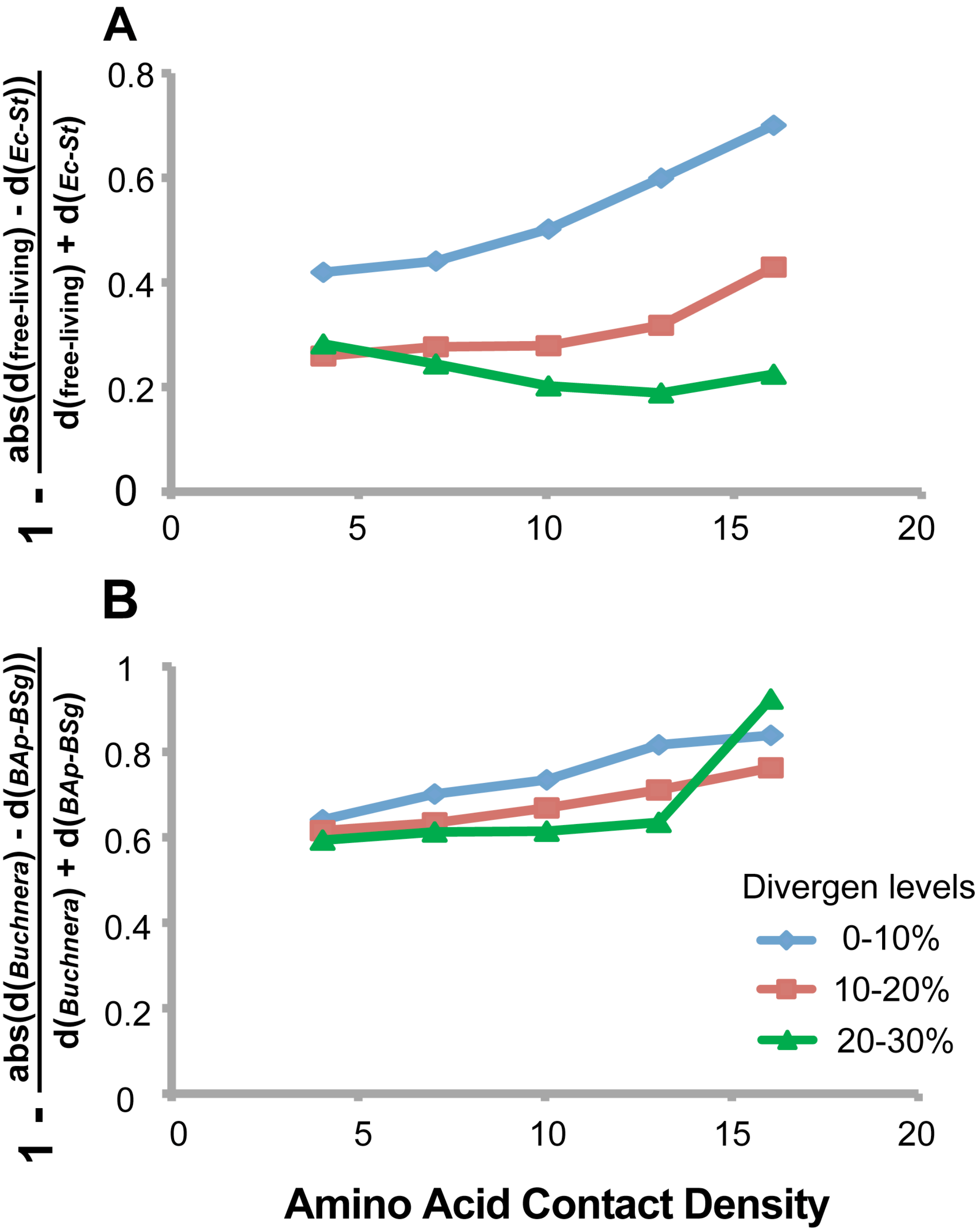
[a] Gene name

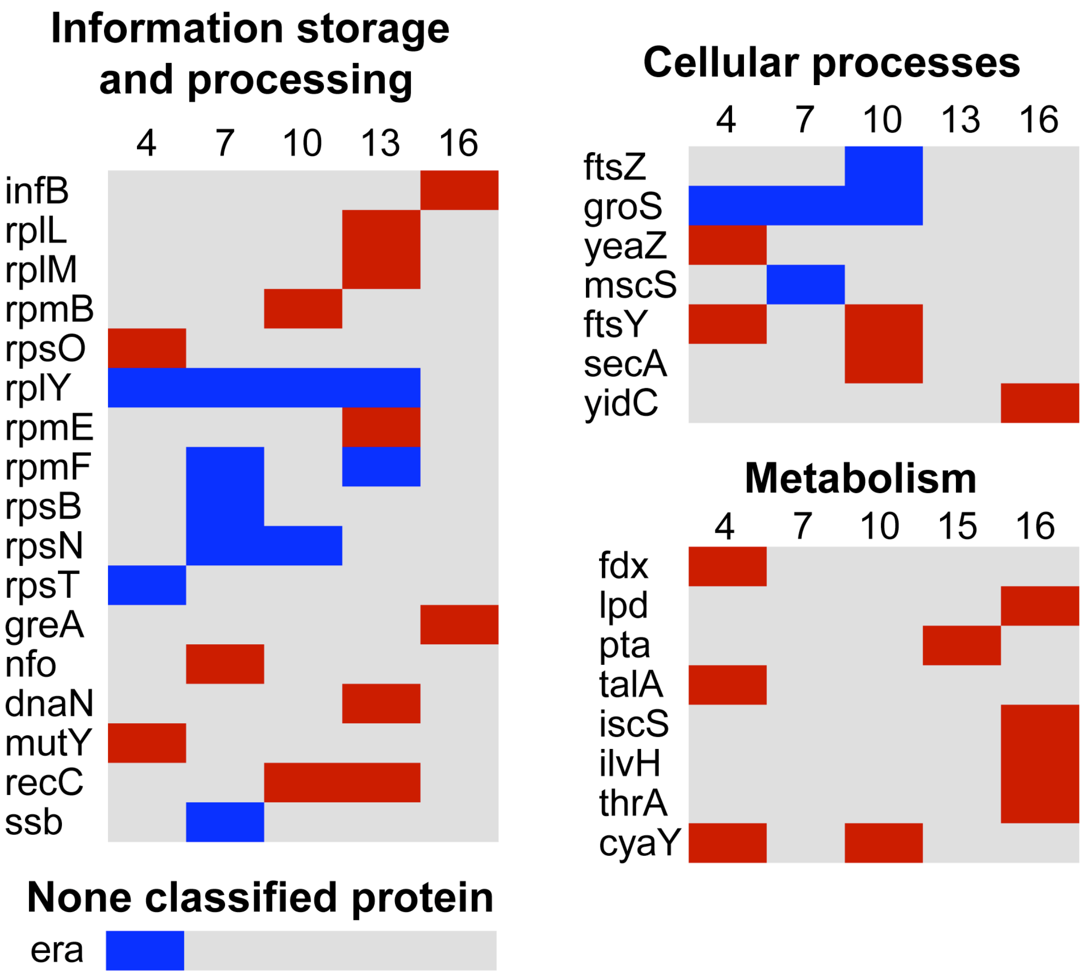
[b] Protein product and its biological function.

[c] Amino acid density calculated as the number of amino acids within 4 Angstroms of the constraints outlier residue in the structure. High densities (H) are those residues with equal or greater than 11 amino acids within 4 Angstroms in the structure. Low densities (L) refer to residues with less than 11 amino acids closeby in the structure.


A
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
3
6
9
12
15
18
B
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
3
6
9
12
15
18
Divergence levels
0-10%
10-20%
20-30%
C
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
3
6
9
12
15
18
Mean Poisson Distance
Amino Acid Contact Density


A
B
C
D
E
F
Mean Poisson Distance
1.0
0.8
0.6
0.4
0.2
0
1.4
1.2
1
0.6
0.5
0.3
0.1
4
7
10
13
16
Amino Acid Contact Density
Divergen levels
Metabolism
Cellular processes
Information storage and processing
0-10%
10-20%
20-30%


A
1 - abs(d(free-living) - d(Ec-St)) / d(free-living) + d(Ec-St)
0.8
0.6
0.4
0.2
0
0
5
10
15
20
B
1 - abs(d(Buchnera) - d(BAp-BSg)) / d(Buchnera) + d(BAp-BSg)
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
Divergen levels
0-10%
10-20%
20-30%
Amino Acid Contact Density



Information storage and processing

	4	7	10	13	16
infB	⬜	⬜	⬜	⬜	🟥
rplL	⬜	⬜	⬜	🟥	⬜
rplM	⬜	⬜	⬜	🟥	⬜
rpmB	⬜	⬜	🟥	⬜	⬜
rpsO	🟥	⬜	⬜	⬜	⬜
rplY	🟦	🟦	🟦	🟦	⬜
rpmE	⬜	⬜	⬜	🟥	⬜
rpmF	⬜	🟦	⬜	🟦	⬜
rpsB	⬜	🟦	⬜	⬜	⬜
rpsN	⬜	🟦	🟦	⬜	⬜
rpsT	🟦	⬜	⬜	⬜	⬜
greA	⬜	⬜	⬜	⬜	🟥
nfo	⬜	🟥	⬜	⬜	⬜
dnaN	⬜	⬜	⬜	🟥	⬜
mutY	🟥	⬜	⬜	⬜	⬜
recC	⬜	⬜	🟥	🟥	⬜
ssb	⬜	🟦	⬜	⬜	⬜

Cellular processes

	4	7	10	13	16
ftsZ	⬜	⬜	🟦	⬜	⬜
groS	🟦	🟦	🟦	⬜	⬜
yeaZ	🟥	⬜	⬜	⬜	⬜
mscS	⬜	🟦	⬜	⬜	⬜
ftsY	🟥	⬜	🟥	⬜	⬜
secA	⬜	⬜	🟥	⬜	⬜
yidC	⬜	⬜	⬜	⬜	🟥

Metabolism

	4	7	10	15	16
fdx	🟥	⬜	⬜	⬜	⬜
lpd	⬜	⬜	⬜	⬜	🟥
pta	⬜	⬜	⬜	🟥	⬜
talA	🟥	⬜	⬜	⬜	⬜
iscS	⬜	⬜	⬜	⬜	🟥
ilvH	⬜	⬜	⬜	⬜	🟥
thrA	⬜	⬜	⬜	⬜	🟥
cyaY	🟥	⬜	🟥	⬜	⬜

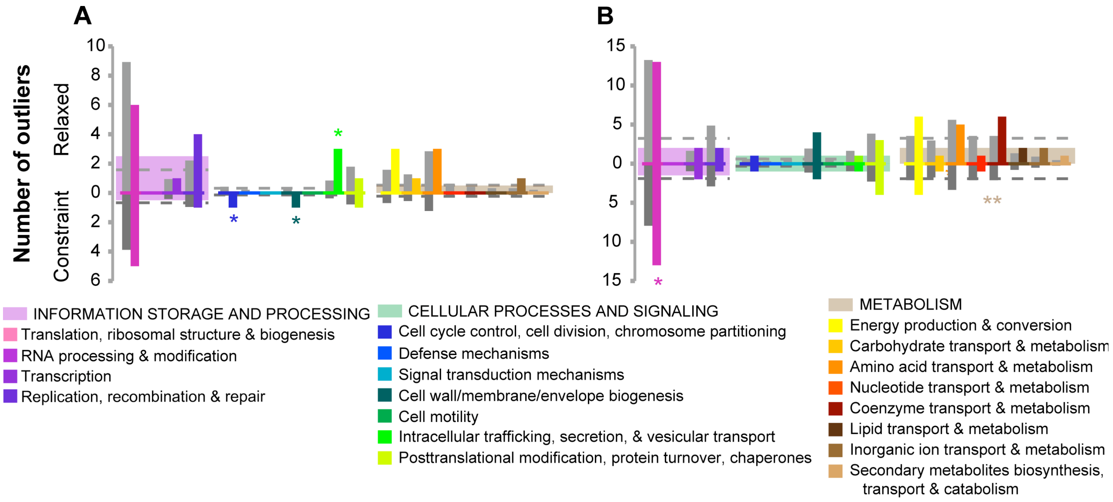
None classified protein

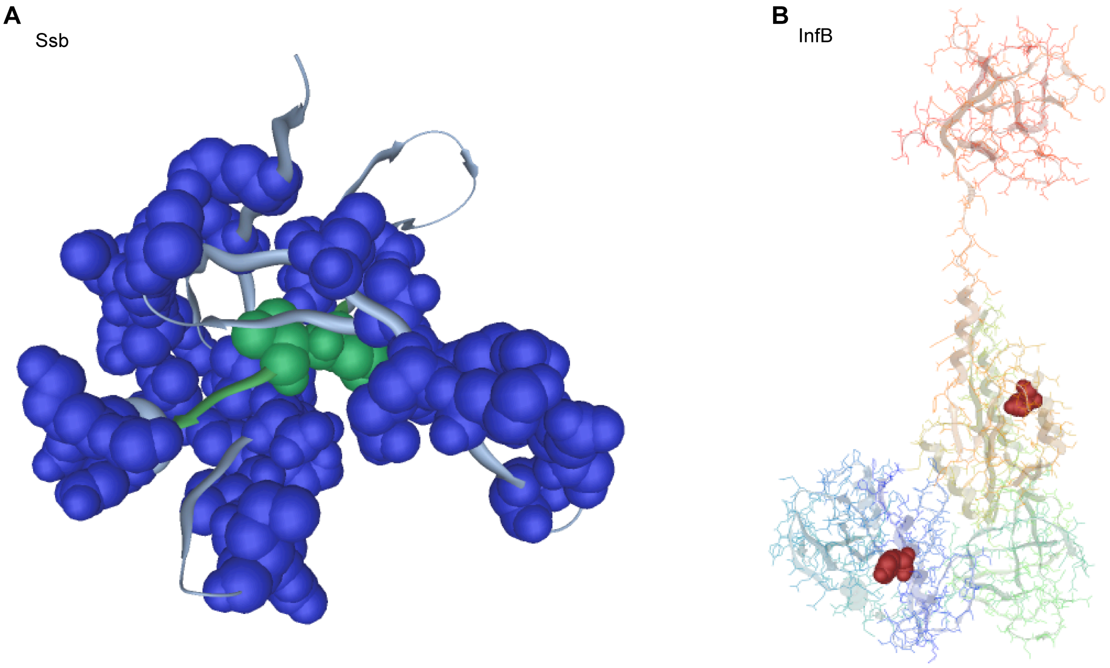
era	🟦	⬜	⬜	⬜	⬜

⬜ Evolving under expected selection contrains

🟥 Representing accelereted evolution

🟦 Evolving under stronger selection contrains


A
B
Number of outliers
Relaxed
Constraint
10
8
6
4
2
0
2
4
6
15
10
5
0
5
10
15
*
**
INFORMATION STORAGE AND PROCESSING
Translation, ribosomal structure & biogenesis
RNA processing & modification
Transcription
Replication, recombination & repair
CELLULAR PROCESSES AND SIGNALING
Cell cycle control, cell division, chromosome partitioning
Defense mechanisms
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Cell motility
Intracellular trafficking, secretion, & vesicular transport
Posttranslational modification, protein turnover, chaperones
METABOLISM
Energy production & conversion
Carbohydrate transport & metabolism
Amino acid transport & metabolism
Nucleotide transport & metabolism
Coenzyme transport & metabolism
Lipid transport & metabolism
Inorganic ion transport & metabolism
Secondary metabolites biosynthesis, transport & catabolism


A
Ssb
B
InfB