

Slow protein evolutionary rates are dictated by surface–core association

Ágnes Tóth-Petróczy and Dan S. Tawfik¹

Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Gregory A. Petsko, Brandeis University, Waltham, MA, and approved May 25, 2011 (received for review October 27, 2010)

Why do certain proteins evolve much slower than others? We compared not only rates per protein, but also rates per position within individual proteins. For ~90% of proteins, the distribution of positional rates exhibits three peaks: a peak of slow evolving residues, with average $\log_2[\text{normalized rate}]$, $\log_2\mu$, of ca. -2 , corresponding primarily to core residues; a peak of fast evolving residues ($\log_2\mu \sim 0.5$) largely corresponding to surface residues; and a very fast peak ($\log_2\mu \sim 2$) associated with disordered segments. However, a unique fraction of proteins that evolve very slowly exhibit not only a negligible fast peak, but also a peak with a $\log_2\mu \sim -4$, rather than the standard core peak of -2 . Thus, a “freeze” of a protein’s surface seems to stop core evolution as well. We also observed a much higher fraction of substitutions in potentially interacting residues than expected by chance, including substitutions in pairs of contacting surface–core residues. Overall, the data suggest that accumulation of surface substitutions enables the acceptance of substitutions in core positions. The underlying reason for slow evolution might therefore be a highly constrained surface due to protein–protein interactions or the need to prevent misfolding or aggregation. If the surface is inaccessible to substitutions, so becomes the core, thus resulting in very slow overall rates.

core mutation | correlated mutation | protein mutations | surface mutation

Protein evolutionary rates—that is, the rate by which protein sequences change over time—differ by factors of 100 to 1,000, even within the same organism (1–9). Early theories surmised that rates are determined by the proportion of protein sites that are involved in function, and are therefore under strong purifying selection (10). However, although many plausible functional and biophysical constraints have been examined in relation to evolutionary rates, only weak correlations were found. Functional constraints manifested in gene essentiality show no correlation with evolutionary rates (7, 11). However, proteins with a larger fraction of interactive surface area tend to be more conserved (4, 12), supporting the notion that binding interfaces evolve slower than nonbinding surfaces (3). Unexpectedly, expression levels seem to be the strongest predictor of evolutionary rates (13, 14). Highly expressed proteins tend to evolve slowly, possibly because of their higher sensitivity to mistranslation and misfolding (2). Nonetheless, the divergence rates of protein sequences and of their expression levels are not correlated (6). Overall, the variety of functional and biophysical factors that constrain protein evolution complicates the understanding of evolutionary rates (1).

Evolutionary rates were thus far analyzed only as overall rates per protein (2–8). However, within a given protein, each position is under a different type and magnitude of selection pressure (10). A buried, core position would be under strong selection to maintain the protein’s configurational stability; a surface residue with no functional role would be under no, or weak selection; and a key active-site surface residue is likely to accept no mutations. We therefore examined the variation of rates at the residue level. We analyzed the distributions of positional rates in different proteins, and their relation to the overall, per protein rates. We also examined the association between sequence changes on the surface of a protein and in its core. At low divergence, the core barely changes, and the relative fraction of

core substitutions increases with divergence (15). We therefore examined the possibility that the acceptance of core mutations is dependent on earlier changes in surface residues. Examination of the positional distributions, and of other indications for the association between surface and core mutations, led us to hypothesize that if the surface of a protein is blocked to mutations, so is its core. The underlying cause of slow evolutionary rates may therefore be the inaccessibility of a protein’s surface to mutations.

Results

Canonical Positional Rate Distributions. We analyzed 3,778 proteins with orthologs in 10 yeast species of known phylogeny (16). Orthologs were aligned and placed within the known species tree. Positional rates were subsequently calculated using Rate4Site, a method that computes the rate for individual protein positions along the entire phylogeny, including the predicted nodes (17). The positional rates obtained by Rate4Site are normalized per phylogeny such that the mean rate for all sites in all analyzed proteins is 1. A positional rate, therefore, indicates the rate by which a given site evolves relative to the average rate of the entire protein phylogeny. Rates are presented at $\log_2[\text{normalized rate}]$, or $\log_2\mu$, as this scale gave graphically clear views. The normalized rates per protein were also calculated (μ^P), and found to be distributed as reported (Fig. 1) (18). Because we were interested in correlating rates of individual positions with their structural features, we initially examined a subset of proteins, and protein domains, with known structures ($n = 382$) (SI Appendix, Fig. S1 and Table S1). These domains evolve slower (mean $\mu^P = 1.3$) compared with all yeast proteins (mean $\mu^P = 2.0$), reflecting the tendency to select conserved, well-ordered proteins for crystallization.

Over 90% of the structured proteins analyzed exhibited positional distributions that could be described by the overlay of three peaks: a slow peak, with an average $\log_2\mu$ of -2.0 ; a broader peak, with an average $\log_2\mu \sim 0.5$ (fast peak); and a third peak at $\log_2\mu \sim 2.0$ (very fast peak) (Fig. 2 and SI Appendix, Fig. S1). Adenylate kinase is a typical example of a protein with an average evolutionary rate (μ^P 0.5–2) (Fig. 1). This kinase exhibits a positional -2 peak and a much smaller 0.5 one, the latter corresponding primarily to surface residues, which are not functionally constrained (Fig. 24). Fast-evolving proteins exhibit not only a smaller slow peak ($\log_2\mu \sim -2$) and larger fast peak ($\log_2\mu \sim 0.5$), but also a peak of very fast-evolving residues ($\log_2\mu \sim 2$), comprising mostly loop and surface residues (Fig. 2B).

The above canonical distribution does not apply to slowly evolving proteins ($\mu^P \leq 0.5$) (SI Appendix, Tables S2 and S3). With few exceptions, slowly evolving proteins exhibited a different positional distribution (SI Appendix, Table S4). As expected, the fast peak corresponding to surface residues almost completely dis-

Author contributions: D.S.T. designed research; A.T.-P. performed research; A.T.-P. and D.S.T. analyzed data; and A.T.-P. and D.S.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence should be addressed. E-mail: dan.tawfik@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1015994108/-DCSupplemental.

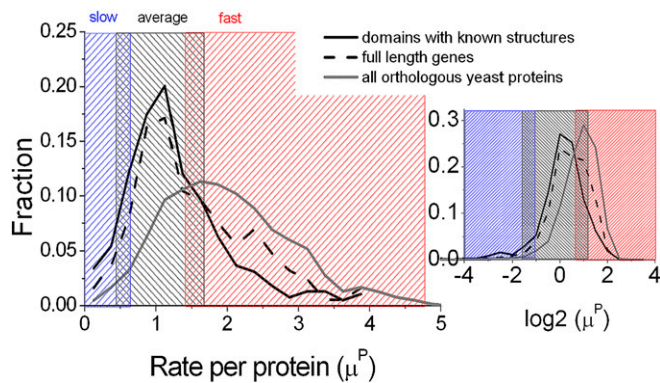


Fig. 1. Distribution of evolutionary rates per protein, or per protein domain (μ^P). Shown are the distributions for the complete set of yeast genes (gray line, $n = 3,778$, mean = 2.0 ± 0.9), for proteins and domains with known structures (black, $n = 382$, mean = 1.3 ± 0.7), and the full-length genes that encode these domains (dashed line, $n = 384$, mean = 1.5 ± 0.8). The differences between these distributions are statistically significant ($P < 0.0001$, ANOVA test). Slow, average, and fast evolving rates are marked with blue, black, and red backgrounds, respectively. (Inset) The same distributions on a \log_2 scale mark the range of slow evolutionary rates.

appeared. However, the most distinct and unexpected feature of slow-evolving proteins is the loss of the slow peak at -2 that corresponds to core residues. Instead, these proteins exhibit a peak at -4 (very slow peak) (Fig. 2C). It therefore appears that not only the surfaces, but also the cores of these proteins, evolve very slowly.

Analysis of the complete dataset ($n = 3,778$) indicated that most proteins (>90%) follow the standard canonical positional

rate distribution (SI Appendix, Fig. S2). Furthermore, $\sim 8\%$ of the structured domain set ($n = 32$), and $\sim 2\%$ of all yeast genes ($n = 70$), evolve very slowly ($\mu^P \leq 0.5$) and also exhibit a very slow peak ($\log_2 \mu^P \sim -4$) (SI Appendix, Tables S2 and S3). The described canonical distributions are therefore generic and apply to a very wide variety of proteins.

Grouping the analyzed proteins as slow, average, and fast, based on the relative area of the canonical peaks of their positional distributions, resulted in nearly perfect correlation with the per protein rates (Fig. 3A). For slow proteins, the -4 peak served as a marker. For average-rate proteins, the marker was larger -2 peak, and smaller 0.5 and 2 peaks, and for fast proteins, larger 0.5 and 2 peaks. Even a simpler measure, such as the relative area of the slow peak, correlated with the average rates for all proteins, except the very fast-evolving ones ($r = 0.87$, $P < 0.0001$).

Structural Assignment of Positional Rates. We assigned the different residue types, such as ordered/disordered, or core/surface, based on 3D structures ($n = 382$ set). The very fast peak corresponded mostly to disordered regions, as exemplified in Fig. 2D. For the complete set ($n = 3,778$), 84% of the predicted disordered residues appear in fast and very fast peaks (SI Appendix, Fig. S3). Comparison of the structured domains to the complete proteins from which they were derived shows that, on average, the latter exhibit higher evolutionary rates (Fig. 1, black line vs. dashed line, and Table 1). This difference is primarily because of disordered regions (e.g., long loops) or disordered domains that are missing in crystal structures. The average per protein rates (μ^P) for predicted disordered regions is approximately twofold higher than for predicted ordered regions (Table 1, and SI Appendix, Fig. S3 A and B). These observations are in

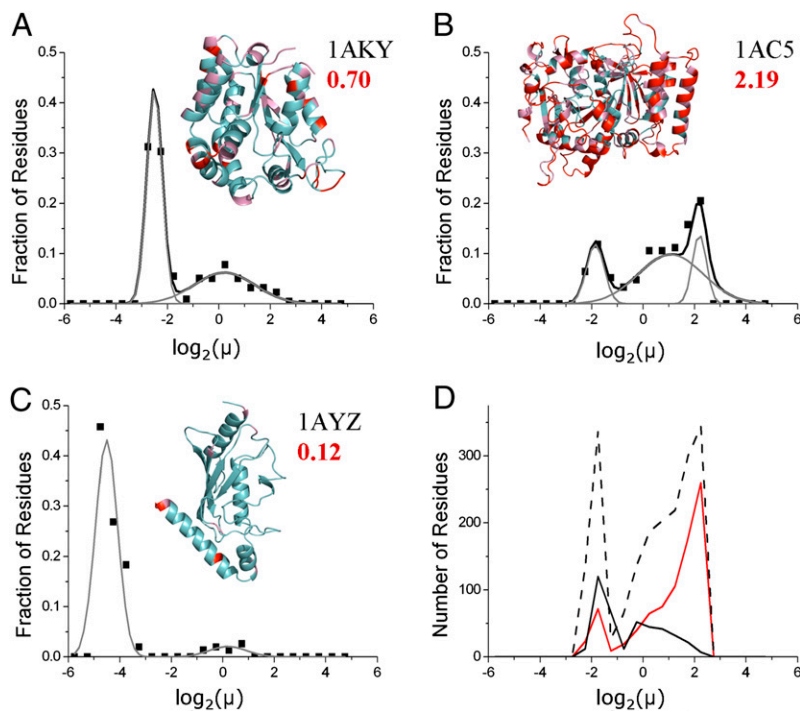


Fig. 2. Canonical types of positional rate distributions. To illustrate the canonical peaks, the computed distributions (black dots) were fitted by superposition of Gaussian peaks (in gray). The backbone structures mark residues by positional rates: in cyan, slow evolving, $\log_2 \mu < 0.5$; in pink, $\log_2 \mu = (0.5-2)$; in red, fast-evolving residues, $\log_2 \mu > 2$. PDB codes and average rates per protein (μ^P , in red) are denoted above each distribution. (A) Proteins evolving at average rates, such as adenylate kinase, exhibit a large slow peak at an average positional rate of about -2.0 , and smaller and broader fast peak at $\log_2 \mu \sim 0.5$. (B) Carboxypeptidase exemplifies fast-evolving proteins that, in addition to the slow and fast peaks, exhibit a very fast peak with an average $\log_2 \mu$ of ~ 2.0 . (C) Ubiquitin-conjugating enzyme RAD6 represents slowly evolving proteins in which the -2.0 peak is absent, and a very slow peak at -4.0 appears. (D) Histograms of positional rates for BNI1: its structured domain (PDB code: 1UX5; in black); the complete gene (dashed line), and its predicted disordered domains ($\sim 43\%$ of the protein; in red). The majority of disordered residues (89%) appear in the very fast peak.

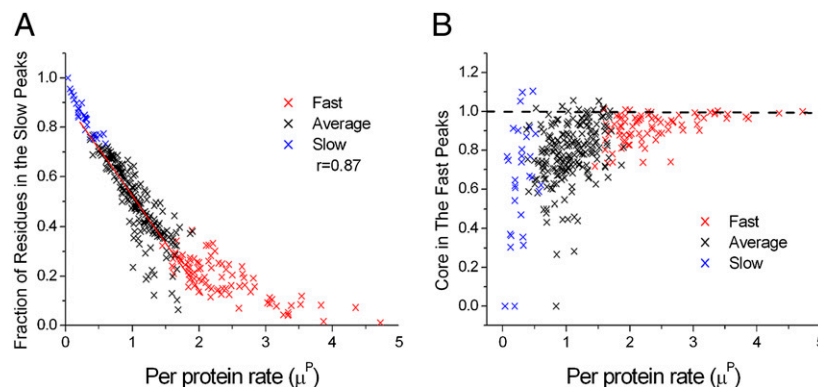


Fig. 3. Classification of proteins by the topology of their distributions of positional rates. (A) The fraction of residues within the slow or very slow positional peaks of a given protein ($\log_2\mu \sim -2.0$, or -4.0 , respectively) versus its overall evolutionary rate (μ^P). The line ($r = 0.87$, $P < 0.0001$) represents a fit of all proteins except the fast evolving ones ($\mu^P > 1.5$). (B) The fraction of core residues in the fast and very fast peaks of a protein ($\log_2\mu = 0.5$, and 2.0 , respectively) divided by the overall fraction of core residues in that protein is plotted versus the protein's overall evolutionary rate (μ^P). The majority of proteins exhibit a relative fraction much smaller than one, indicating that core residues are less likely to appear in the fast peak.

accordance with hypothesis that disordered segments evolve faster than ordered ones (19).

In line with previous analyses of per protein rates, we observed no correlation of positional rates with secondary structure (20), but could correlate rates of individual positions with the degree of their surface exposure (15, 20). The slow peak corresponds mostly to core residues, and the fast and very fast peaks to surface residues. For slow-evolving proteins, almost no residues appear in the fast peaks, whereas in fast-evolving proteins almost all residues appear within these peaks (Fig. 3B and *SI Appendix*, Fig. S3D). Nonetheless, in agreement with the observation that surface residues evolve faster than core ones (15), $\sim 95\%$ of the proteins have less core residues in the fast and very fast peaks than expected by chance. Surface mutations are more tolerated because of their weaker destabilizing effects (21). However, many surface residues are highly constrained by their engagement in binding or other functions. These conflicting trends explain the presence of surface residues in both the slow and fast peaks.

Association Between Surface and Core Mutations. Similar to previous reports (5, 15), our dataset also indicates that, on average, core residues exchange approximately fourfold slower than surface residues. However, the relative rates of core and surface

residues are not constant. If the accumulation of surface substitutions was simply four-times faster than of core residues, we would expect both types of substitutions to show a linear rate with respect to the overall level of divergence. However, it appears that core positions exhibit a lag, and only start to change rapidly when a relatively large fraction of surface positions had exchanged (Fig. 4A). This finding implies that core substitutions tend to accumulate once the surface has sufficiently drifted.

Is there a direct association between surface and core mutations, namely, of surface changes enabling, or facilitating changes in the core? Unfortunately, the order by which individual positions exchanged cannot be revealed: 10 sequences and 8 predicted ancestral nodes per protein represent >100 million years of drift. There are, however, several indirect indications that support our hypothesis. First, substitutions tend to cluster in space much more than expected by chance. Namely, mutations tend to occur in potentially interacting residues. The clustering of mutations increases as the drift extends, suggesting an evolutionary link between physically interacting residues (Fig. 4B). Second, within a given protein, core residues that evolve faster tend to be in contact with other fast-evolving residues (Fig. 4C). Specifically, for slowly evolving residues ($\mu < 0$), which typically are core residues, the dependency is steeper, indicating that a core

Table 1. Average evolutionary rates for different protein subsets

Proteins	Mean rate (median)*	SD; variance*	Residue type	Mean rate (median)	SD; variance
All orthologous yeast proteins $n = 3,722^\dagger$	2.0 (1.9)	0.9; 0.8	Ordered [‡] $n' = 1,782,923$	1.9 (1.3)	1.7; 3.0
			Disordered $n' = 283,302$	3.2 (3.3)	1.9; 3.4
Proteins with known structure [§]	1.5 (1.3)	0.8; 0.7	Ordered $n' = 165,385$	1.5 (0.9)	1.6; 2.4
			Disordered $n' = 21,690$	3.0 (3.1)	1.9; 3.7
Domains with known structure $n = 382$	1.3 (1.1)	0.7; 0.5	Surface $n' = 45,420$	1.6 (1.1)	1.5; 2.3
			Core $n' = 62,585$	1.1 (0.5)	1.2; 1.5
Proteins with ≥ 1 disordered segment $n = 98$	2.0 (2.0)	0.8; 0.7			

*Because the distributions of positional rates are not normal, both the mean values and the median values (in parenthesis) are shown, and the SDs as well as the variances of the averages are listed. The differences between the mean values between the protein subsets are statistically significant ($P < 0.001$, ANOVA test). Differences between the distribution of ordered and disordered residues are statistically significant ($P < 0.05$, Mann-Whitney test).

[†]Sample sizes relate to the number of proteins or domains (n), or residues (n').

[‡]Disordered segments were defined as ≥ 30 residue-long segments with >0.4 IUPred score; all residues within these segments are included in the "disordered" category. Conversely, residues with ≤ 0.4 score were defined as "ordered."

[§]The three rows below relate to evolutionary rates of proteins in which a domain, and sometimes the entire protein, has a known structure. The first row lists the evolutionary rates for the entire protein, including parts with unknown structure. The second row specifies the rates for domains, or whole proteins, with known structures. The third row lists the evolutionary rates of a subset of the first row comprising proteins that are predicted to have at least one disordered segment.

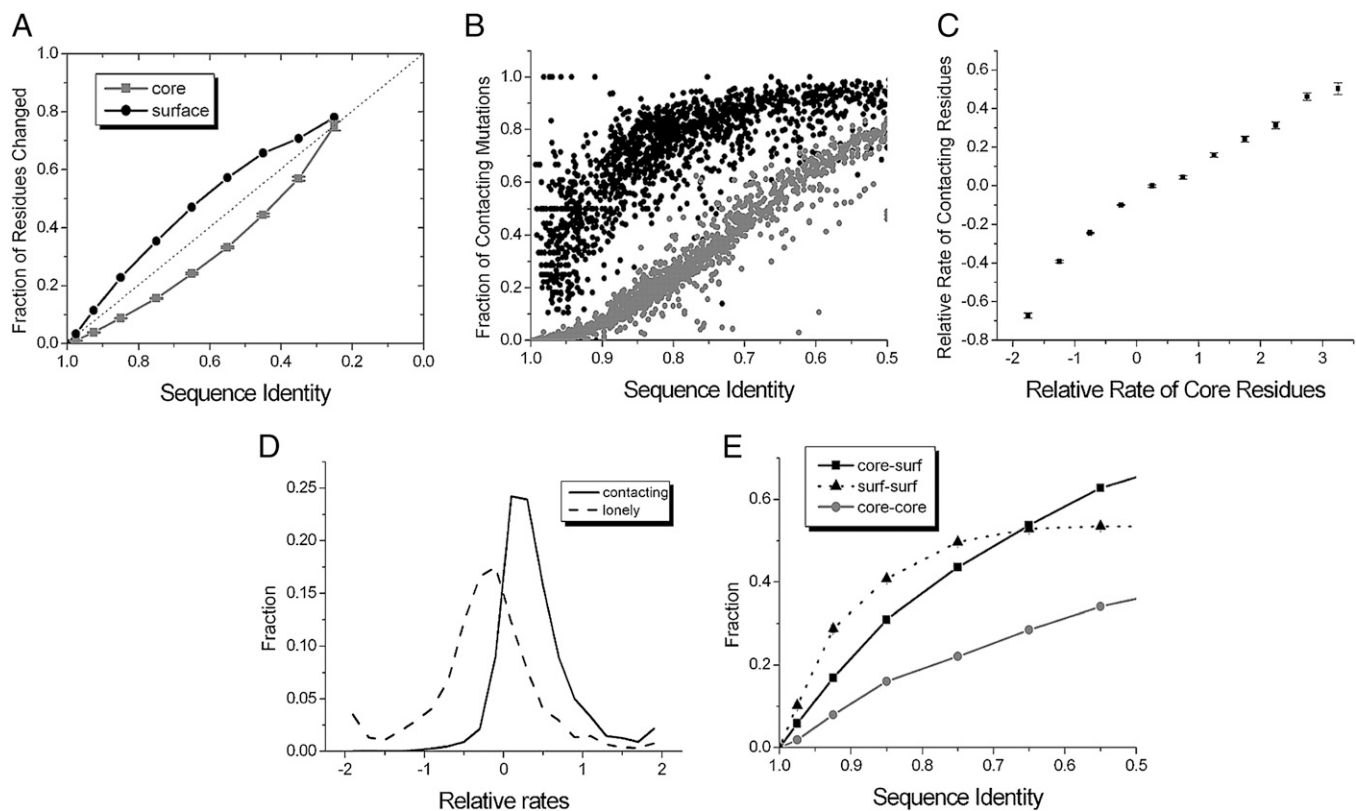


Fig. 4. Surface-core association. Substitutions were defined as sequence changes compared with the *S. cerevisiae* orthologs ($n = 3,456$ sequence pairs, resulting from the comparison of 10 orthologs of the 382 domains with known structure). Standard deviations are shown as bars (often too small to be observed). (A) The fraction of substitutions in the core and in the surface; the values were binned and plotted as a function of sequence identity. (B) The fraction of potentially interacting substitutions (identified as described in *Methods*) as a function of sequence identity. Black dots indicate the fraction of contacting substitutions in the dataset ($n = 3,456$ sequence pairs). The control set of random substitutions was generated with the only constraint of having the same number of substitutions in the core and the surface regions as in the original dataset. A hundred randomly mutated sequences were generated for each protein, and the average fraction of contacting substitutions in the random set is presented as gray dots. (C) For each evolving core position, the relative evolutionary rate ($\mu - \mu^c$), and the relative rates of all its contacting residues, was obtained. Data were binned by the relative rates of the evolving core residues. (D) Distributions of the relative evolutionary rates ($\mu - \mu^c$) for “lonely” substitutions (mean = 0.40 ± 0.52) versus “contacting” substitutions (mean = -0.26 ± 0.77 ; $P < 0.05$, Kolmogorov-Smirnov test). (E) The fraction of contacting substitutions within core-surface, surface-surface, and core-core residues averaged within bins of 0.1 ($n = 3,456$).

residue is much more unlikely to change if its neighboring residues remain unchanged.

That clustering correlates with faster evolutionary rates is also seen in the analysis of evolutionary rates of “lonely” positions (no mutations occurred in neighboring residues) versus “contacting” positions (at least one potentially interacting residue diverged). Contacting mutations were overrepresented within residues that evolve faster than the average rate of the corresponding protein (Fig. 4D). Furthermore, at the early divergence stages, contacting surface-surface substitutions accumulate the fastest, followed by core-surface and core-core substitutions (Fig. 4E). The fractions of surface-surface and core-surface contacting substitutions becomes equal only at an average of $\sim 35\%$ sequence divergence.

The clustering of mutations could also be a result of functional constraints that restrict mutations to certain regions of the protein. To exclude this possibility, we analyzed a mutational drift performed in the laboratory. Analysis of drifted sequences of TEM-1 β -lactamase (22) showed a far higher fraction of mutations in contacting residues than expected by chance (9.8% and 0.5% respectively, empirical P value $< 10^{-6}$) (SI Appendix, Table S5). To exclude the effect of functional constraints, we repeated the simulated random drift but excluded all functionally relevant residues (36 residues in and around TEM-1’s active-site). We found, however, that this constraint does not account for the high frequency of mutations in contacting residues. Thus, irrespective of functional constraints, mutations tend to occur in

contacting, potentially interacting residues both in natural and laboratory drifts.

Discussion

The above analyses support the notion of a direct association between the mutations, and specifically between surface and core substitutions: a change in one residue, most frequently a surface residue, facilitates sequence changes in contacting residues, and most notably in core residues. The surface-core association seeks to explain the observed shift in the positional rates of core residues of very slowly evolving proteins, from -2 to -4 . In the absence of such association, the core positions of all proteins would diverge at similar rates, regardless of how variable their surfaces are. Indeed, on average, core residues evolve approximately four-times slower than surface ones (15). However, the shift of the slow peak ($\log_2 \mu \sim -2$) to the very slow peak ($\log_2 \mu \sim -4$) indicates that in proteins where the surface evolves slowly, the core evolves ~ 16 -times slower.

The complexity of the determinants of protein sequence implies a variety of independent properties that correlate with slow evolution (1, 7), including multiple interaction partners (3), essentiality (7, 8), evolutionary age (18), and high expression levels (2). Nonetheless, we suggest that the underlining reason for slow evolutionary rates is a highly constrained surface. Surface residues of slow evolving proteins show extraordinarily slow evolutionary rates. As discussed below, these surfaces seem to be

constrained, a result of both function and stability factors. However, why do slowly evolving proteins also exhibit exceptionally slow evolutionary rates of core residues? The data summarized in Fig. 4 support the notion that sequence changes in certain positions are associated with substitutions in other positions, thus slowing down the rate of evolution (23, 24). We specifically observed an association between surface and core exchanges. We thus hypothesize that surface substitutions enable subsequent changes in the core. If core substitutions were enabling surface substitutions, we would observe proteins with diverged cores and conserved surfaces. However, 89% of the slowly evolving proteins ($\mu^P \leq 0.5$) exhibit an unusually slow evolving core ($\log_2 \mu \sim -4$) (SI Appendix, Tables S2 and S3).

The stability effects of mutations also supports an order of surface substitutions followed by core. Apart from their possible effects on function, mutations primarily affect the configurational stability of proteins (25). Core mutations tend to be highly destabilizing. However, among surface residues there are few highly destabilizing mutations and the average $\Delta\Delta G$ per mutation is nearly neutral (~ 0.5 kcal/mol) (21). The predicted stability effects of core and surface substitutions in the yeast protein phylogeny show similar trends (SI Appendix, Fig. S4B). Strongly destabilizing mutations ($\Delta\Delta G \geq 3$ kcal/mol) are generally purged as seen in laboratory drifts (26). Among the substitutions that were tolerated, 4% of the core substitutions have predicted $\Delta\Delta G$ values of ≥ 3 kcal/mol in comparison with 0.7% of the surface substitutions. Thus, surface substitutions that may have initially appeared as neutral may enable deleterious core substitutions to accumulate at later stages. This mechanism (and other unknown ones) may account for the fact that sequence changes that are deleterious in one protein are tolerated in closely related proteins, and for the epistatic nature of protein sequence divergence (23, 24).

Why are the surfaces of slowly evolving proteins highly constrained? Two major causes may apply: (i) selection to maintain functionally important residues and, in particular, residues involved in protein–protein interactions; and (ii) selection to prevent misfolding or aggregation at high-expression levels. Surface residues engaged in function (e.g., transient binding of ligands or proteins, regulation, or oligomerization) are highly conserved. The evolution of protein–protein interface residues is nearly as slow as core residues (3, 5, 27). Among the slowest evolving proteins, many—if not most—proteins are part of large complexes (such as RNA polymerase, proteasome, ribosome) and are involved in multiple interactions (e.g., ubiquitin, nuclear transport factor, and so forth) (SI Appendix, Tables S2 and S3). Indeed, the average number of interacting partners for slow-evolving proteins (10 ± 8) is nearly threefold larger than for the remaining dataset (3.6 ± 4.2 ; $P < 0.0001$) (SI Appendix, Table S3).

The divergence of surface residues might also be constrained by the propensity to misfold because of translational errors under high expression rates (28). As shown before (1, 13), the yeast protein phylogeny shows a significant correlation of evolutionary rates with expression levels (SI Appendix, Fig. S5). Indeed, $\sim 70\%$ of slow-evolving proteins exhibit expression levels above average (SI Appendix, Table S3). The unusually slow evolutionary rates of surface as well as core residues may also relate to a selection to maintain higher configurational stability and lower aggregation propensity, regardless of translational errors (29). In agreement with previous analysis (30), slow-evolving proteins share compositional properties with thermophilic proteins (SI Appendix, Fig. S4A). Misfolding-minimizing amino acids were also shown to be overrepresented and to be conserved in highly expressed proteins (29). Indeed, we found that slowly evolving proteins are enriched in the IVYWREL amino acids whose fraction is correlated with higher thermostability (31). In particular, we observed a higher fraction of Arg and Glu on the surface (SI Appendix, Fig. S4A). Thus, the need to maintain high configurational stability and reduce aggregation imposes sequence

constraints on surface positions that are not required for function. These constraints might be particularly strong in proteins in which a significant fraction of the surface is engaged in protein–protein interactions. Aggregation-resistant surface regions may also be present at binding interfaces, and on highly expressed proteins, thus preventing aggregation as well as undesirable reversible interactions (32). Highly conserved patches that resemble the protein's core comprise the key part of protein–protein interfaces (27, 33). In general, interface regions were shown to be more prone to aggregation than other surface regions (32) (in our dataset, however, surface residues of slowly evolving proteins are not significantly more hydrophobic than those of all other proteins). A single residue exchange may elicit a binding potential and, hence, an undesirable cross-reactivity (27). Similar to other known stability–activity tradeoffs in proteins (34), reducing aggregation propensity (i.e., decreasing surface stickiness, or increasing “solubility”) demands certain changes in surface residues, such as exchanging hydrophobic to polar or charged amino acids, but these changes may abolish essential protein–protein interactions.

As shown here, high fraction of surface residues that are highly restricted, and thereby evolve very slowly, results in other residues, and core residues in particular, evolving very slowly. This “auto-catalytic freeze” relates to the need to maintain both specific protein functions (such as catalysis or binding) and general ones, such as stability, charge, and reduced aggregation propensity (10). The need to maintain all these functions results in tradeoffs, and hence in slower evolution rates of surface residues. The surface–core association further expands this freeze into the protein's core resulting in very slow overall rates.

Methods

Yeast Protein Database. Orthologous sequences were collected from the sequenced genomes of *Saccharomyces cerevisiae* and nine closely related species: *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Candida glabrata*, *Saccharomyces castellii*, *Saccharomyces kluyveri*, *Kluyveromyces lactis*, *Kluyveromyces waltii*, and *Ashbya gossypii*. Orthologous gene annotations and sequences were taken from the Fungal Orthogroups Repository (www.broadinstitute.org/reggev/orthogroups) (35). When more than one ortholog was assigned, the sequence with the highest sequence identity to *S. cerevisiae* was chosen. Only genes that have orthologs in all 10 species were included ($n = 3,778$). For analyzing proteins of known structure, we used *S. cerevisiae* proteins, or protein domains, with crystal structures at ≤ 3 Å resolution, available from the PDB (release 09.04.2010).

Sequence Analysis. Alignments were created with MUSCLE (36). For the structure-based analysis, the protein parts for which structure is available were considered. Proteins, or domains, for which $>10\%$ of the sequence lacked structural information were excluded. In total, 382 proteins and domains fitted the above-mentioned criteria (SI Appendix, Table S1). The solvent accessible surface area (ASA) was calculated with CSU (37). ASA was defined as the ratio of the solvent accessible surface for a given residue within the structured protein vs. in the free residue. Residues were classified as core for $ASA \leq 0.15$, and surface for $ASA > 0.15$. The threshold used previously was 0.25, but ASA values were calculated with a different program (21). In both cases, however, ca. half the residues (for an average protein size of 200 amino acids) were classified as core. Disorder prediction was carried out by IUPred (38). Contacts between residues were defined based on CSU analysis of the structures, which considers atom types, distance, and interacting surface area for the definition of potentially interacting residues (37).

Phylogenetic Analysis and Evolutionary Rates. A maximum-likelihood tree was built by PhyML (39), using a concatenated alignment of 180 proteins representing the 10 yeast genomes, the JTT matrix, and 16 different substitution-rate categories (SI Appendix, Fig. S6A). The tree exhibited the known topology for these species. Individual phylogenetic trees were also calculated for each protein and compared with the species tree. Two proteins that showed high deviations, possibly because of horizontal gene transfer, were thereby excluded (SI Appendix, Fig. S6). Evolutionary rates are commonly calculated per protein, either by the number of nonsynonymous changes

between two sequences divided by the total number of nonsynonymous sites (dN) (2, 3, 6, 20), or by dN/dS ratios (4), or both (7, 14). The former ignores the common situation of amino acid exchanges that proceed through intermediates, and none of these methods provides rates per individual positions. We used Rate4Site (40), which performs more accurate Bayesian rate estimations (17), and applied it using the JTT matrix (other matrices gave identical results) (*SI Appendix, Fig. S7B*). This method is more accurate because it considers the predicted ancestral sequences, and thereby includes intermediate changes. Rate4Site also assumes a different rate for each residue and describes the rate distribution with a relatively unrestricted γ -shaped function. A site-specific rate (μ) is calculated, which indicates how fast this site evolves relative to the average rate across all sites in all included proteins. The average per protein rates (μ^p) were obtained by calculating the mean value of all positional rates for a given protein. For comparison, dN/dS ratios were computed using Codeml from the PAML package and assuming one substitution rate for all branches (model = 0) (41). The correlation with the Rate4Site rates is low ($R = 0.5$) (*SI Appendix, Fig. S7A*), but the distributions of per protein rates we obtained (Fig. 1) resemble those obtained with dN/dS (18), and so do the correlations of evolutionary rates and expression levels (*SI Appendix, Fig. S5*) and of ordered versus disordered

domains (Table 1). To ensure that the computed Rate4Site rate distributions are unbiased, calculations were repeated with combinations of two or three γ -functions using the GammaMixture program (42). This process did not significantly affect the analysis (*SI Appendix, Fig. S7*), as anticipated for single-domain proteins (42). We also applied maximum likelihood as an alternative to the default Bayesian method. As expected, this process gave different rate distributions, yet supported the same conclusions (*SI Appendix, Fig. S8*).

Computational and Statistical Analyses. Computational and statistical analyses were performed using perl scripts, and the graphs and statistical analyses were performed with Matlab and Origin. Statistical significance of the differences of the means was tested with a Student *t* test for normally distributed data, and with the Kolmogorov-Smirnov test and Mann-Whitney test for nonnormally distributed data.

ACKNOWLEDGMENTS. We thank Tal Pupko and Itay Mayrose for their invaluable support and Itay Tirosh and Ben Lehner for insightful comments. This work was supported in part by the Meil de Botton Aynsley and the Israel Science Foundation. D.S.T. is the Nella and Leo Benozziyo Professor of Biochemistry.

- Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7:337–348.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102:14338–14343.
- Eames M, Kortemme T (2007) Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Structure* 15:1442–1451.
- Kim PM, Lu LJ, Xia Y, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314:1938–1941.
- Lin YS, Hsu WL, Hwang JK, Li WH (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol* 24:1005–1011.
- Tirosh I, Barkai N (2008) Evolution of gene sequence and gene expression are not correlated in yeast. *Trends Genet* 24(3):109–113.
- Wall DP, et al. (2005) Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci USA* 102:5483–5488.
- Zhang J, He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* 22:1147–1155.
- Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. *Curr Opin Struct Biol* 20:385–389.
- Zuckermandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J Mol Evol* 7(3):167–183.
- Pál C, Papp B, Hurst LD (2003) Genomic function: Rate of evolution and gene dispensability. *Nature* 421:496–497, discussion 497–498.
- Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296:750–752.
- Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
- Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23:327–337.
- Sasidharan R, Chothia C (2007) The selection of acceptable protein mutations. *Proc Natl Acad Sci USA* 104:10080–10085.
- Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449(7158):54–61.
- Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. *Mol Biol Evol* 21:1781–1791.
- Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA* 106:7273–7280.
- Brown CJ, Johnson AK, Daughdrill GW (2010) Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol* 27:609–621.
- Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol* 23:1751–1761.
- Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369:1318–1332.
- Bershtein S, Goldin K, Tawfik DS (2008) Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol* 379:1029–1044.
- Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. *Nature* 465:922–926.
- Lunzer M, Golding GB, Dean AM (2010) Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* 6:e1001162.
- Yue P, Li Z, Moutl J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 353:459–473.
- Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19:596–604.
- Levy ED (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol* 403:660–670.
- Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
- Yang JR, Zhuang SM, Zhang J (2010) Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol Syst Biol* 6:421.
- Cherry JL (2010) Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol Biol Evol* 27:735–741.
- Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 3(1):e5.
- Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M (2009) Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proc Natl Acad Sci USA* 106:10159–10164.
- Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots—A review of the protein-protein interface determinant amino-acid residues. *Proteins* 68:803–812.
- Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11:572–582.
- Wapinski I, et al. (2010) Gene duplication and the evolution of ribosomal protein gene regulation in yeast. *Proc Natl Acad Sci USA* 107:5505–5510.
- Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
- Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics* 15:327–332.
- Dosztányi Z, Csizmók V, Tompa P, Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347:827–839.
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
- Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18 (Suppl 1): S71–S77.
- Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
- Mayrose I, Friedman N, Pupko T (2005) A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics* 21 (Suppl 2):ii151–ii158.