Quantifying Evolutionary Importance of Protein Sites: A Tale of Two Measures

Short title: *Quantifying Evolutionary Importance of Protein Sites*

Avital Sharir-Ivry¹, Yu Xia^{1*}

¹Department of Bioengineering, McGill University, Montreal, QC, Canada

* Corresponding author:

E-mail: brandon.xia@mcgill.ca

Abstract

A key challenge in evolutionary biology is the accurate quantification of selective pressure on proteins and other biological macromolecules at single-site resolution. The evolutionary importance of a protein site under purifying selection is typically measured by the degree of conservation of the protein site itself. A possible alternative measure is the strength of the site-induced conservation gradient in the rest of the protein structure. However, the quantitative relationship between these two measures remains unknown. Here, we show that despite major differences, there is a strong linear relationship between the two measures such that more conserved protein sites also induce stronger conservation gradient in the rest of the protein. This linear relationship is universal as it holds for different types of proteins and functional sites in proteins. Our results show that the strong selective pressure acting on the functional site in general percolates through the rest of the protein via residue-residue contacts. Surprisingly however, catalytic sites in enzymes are the principal exception to this rule. Catalytic sites induce significantly stronger conservation gradients in the rest of the protein than expected from the degree of conservation of the site alone. The unique requirement for the active site to selectively stabilize the transition state of the catalyzed chemical reaction imposes additional selective constraints on the rest of the enzyme.

Introduction

The evolutionary importance of protein sites under purifying selection can be quantified in two very different ways. The classical, “intrinsic” measure for the evolutionary importance of a protein site is the degree of conservation or evolutionary rate of the protein site itself. Protein residues experience different degrees of selective pressure as a result of the different roles they play in protein stability and function(1,2). For example, residues in a protein core are generally under stronger selective pressure than surface residues due to their importance in stabilizing the protein. Indeed, structural determinants such as solvent exposure(3–7) and degree of packing (8–10) were shown to explain a large portion of the variability in the observed site-specific evolutionary rates. In addition, residues in functional sites such as catalytic sites(11) and ligand-binding sites are also under stronger selective pressure than non-functional residues.

An alternative, “extrinsic” measure for the evolutionary importance of a protein site is the conservation gradient the site exerts on the rest of the protein. Rather than quantifying evolutionary conservation of the protein site itself, this measure captures how the evolutionary conservation of other residues surrounding the site gradually decreases with distance from the site in the tertiary structure. Several studies have indicated the possibility for selective pressure to propagate from the functional site to the rest of the protein via physical interactions between neighboring residues in the three-dimensional structure (12–15). While it is clear that the two measures of the evolutionary importance of protein sites are substantially different, it remains unknown how the two measures relate to each other.

In this paper, we addressed the fundamental question whether there is a direct relationship between the intrinsic and extrinsic measures of the evolutionary importance of protein sites. We have based our study on a dataset of homology-based structural models of the yeast proteome(5,7). Despite their major differences, we show here that there is a strong linear relationship between the degree of conservation of a protein site and the conservation gradient induced from it. In other words, more conserved protein sites tend to induce stronger conservation gradient in the rest of the protein, as selective pressure acting on the protein site percolates via residue-residue interactions. This linear evolutionary conservation-percolation relationship is universal in that it holds for different types of proteins as well as for different types of functional sites in proteins. Remarkably however, catalytic sites in enzymes are the only exception to this universal rule, as catalytic sites induce significantly stronger long-range conservation gradient than other types of functional sites with similar degrees of conservation. We conclude that for many different types of functional sites, site-induced conservation gradient can be explained by the percolation of site-specific selective pressure through the rest of the protein via residue-residue contacts. However, catalytic sites in enzymes induce significantly stronger long-range conservation gradient in the rest of the protein than expected from the percolation theory. This is likely due to the unique requirement for the enzyme active site to selectively bind to and stabilize the transition state of the catalyzed chemical reaction(16).

Overall, we show that a more complete understanding of the selective pressure on protein sites can be achieved by integrating the intrinsic measure of site-specific evolutionary conservation with the extrinsic measure of site-induced conservation gradient, with potential implications in protein design, study of disease mutations and better phylogenetic inference(17).

Results

Evolutionary conservation gradient induced from a protein residue in the proteins structure is linearly correlated with the conservation of the residue

We based our study on a dataset of homology-based structural models of the *S. cerevisiae* proteome. This basic dataset contains structural templates from the Protein Data Bank (PDB)(18) mapped to ORFs of *S. cerevisiae* via sequence alignment (see Methods). We had 1274 yeast ORFs with structural models from the PDB for which residue conservation scores are available in ConSurf-DB (19,20). Each residue in the dataset was ranked according to its relative conservation within the protein (the residue's rank of conservation divided by the total number of residues). This normalized conservation rank of a residue ranges from 0 to 1, with higher rank corresponding to higher conservation within the protein. We have also calculated for each residue the Pearson correlation between the conservation scores of all other residues in the protein and their distance from that reference residue. This Pearson correlation between conservation and distance describes the degree of percolation of the evolutionary conservation from a residue throughout the protein tertiary structure, i.e., the 'long-range conservation gradient'. All residues were then binned into 20 equally spaced bins of conservation rank and the average conservation gradient for each bin was calculated.

A clear negative linear trend is observed between the conservation rank of a residue and the strength of the evolutionary conservation gradient induced from it (Fig 1a). The more a residue is conserved within the protein (higher conservation rank), the stronger the evolutionary conservation gradient it induces. The distribution of overall per-protein Pearson correlations

between conservation ranks and conservation gradients (per-protein conservation-percolation trend) is mainly between -0.2 and -0.7, and the average correlation in the dataset is -0.5 (Fig 1b), indicating that the linear relationship between site-specific conservation and site-induced conservation gradient is high for all types of proteins.

We have also examined this relationship between site-specific conservation and site-induced conservation gradient specifically for functional sites. We identified different functional sites in the dataset: catalytic sites, non-catalytic ligand-binding sites (on enzymes and on nonenzymatic proteins), protein-protein interaction sites, and allosteric sites. We identified catalytic sites using the Catalytic Site Atlas(21), ligand-binding sites using BioLip(22), allosteric sites using the Allosteric Database (ASD)(23,24) and protein-protein interaction sites using our previous protocol of identifying protein-protein interfaces in the yeast proteome(5). Relative solvent accessibility (RSA) of residues was calculated and residues were classified as buried if $RSA=0.0$, exposed if $RSA>0.8$ and middle for $0.0<RSA\leq 0.8$. The linear relationship between site-specific conservation rank and site-induced conservation gradient holds regardless of the residue's location within the protein or its functional role (Fig 2). Overall, our results reveal the existence of a conservation-percolation relationship in which higher residue conservation leads to stronger long-range percolation of selective pressure to adjacent sites in the tertiary structure. Furthermore, this relationship holds for residues in different types of functional sites.

Conservation gradient induced from a protein site is linearly correlated with the evolutionary rate of the site

The linear trend between conservation and percolation shown above is based on conservation ranks of residues, which are relative within a protein. In these calculations, we have lumped together residues from different proteins having similar conservation rank, however these residues, being from different proteins, can be under different selective pressure. To address this caveat, we examined whether the conservation-percolation linear trend still holds when conservation is measured in terms of absolute evolutionary rate (dN/dS). The evolutionary rate is calculated for the yeast proteins in *S. cerevisiae* compared with its orthologs in four closely related yeast species (see Methods). Conservation score annotations are transferred from structural homologs onto yeast proteins using sequence alignment. Then, we binned all residues into 100 equally spaced bins of conservation rank and calculated their average evolutionary rate (dN/dS). The average evolutionary rate increases for residues with increasing conservation rank (decreasing conservation) (S1 Fig), showing high correlation between the evolutionary rate of yeast protein sites and the conservation scores of their structural homologs. The conservation-percolation linear trend is shown to still hold here, where conservation is measured in terms of absolute evolutionary rate (Fig 3). The lower the average evolutionary rate of a protein site is, the stronger percolation of evolutionary conservation is induced from it.

Catalytic sites induce stronger conservation gradients than predicted by the conservation-percolation trend

The annotation of functional and structural site residues was transferred to yeast proteins from their structural models using sequence alignment. The average conservation gradient induced from them is shown in Fig 3. As expected, functional residues evolve more slowly (low dN/dS) than other residues. The average conservation gradients induced from ligand binding sites and protein-protein interaction sites can all be predicted reasonably well from the linear conservation-percolation trend. Protein-protein interaction sites induce slightly weaker conservation gradient than expected, possibly due to their tendency to rewire during evolution. Exposed non-functional residues induce slightly stronger conservation gradient than expected, possibly due to hidden functional sites on protein surfaces. Remarkably, catalytic sites are the only functional sites with the most significant deviation from the linear conservation-percolation trend. The average conservation gradient induced from catalytic site residues is significantly stronger than expected by the linear conservation-percolation trend. This can also be seen from the stronger conservation gradients from catalytic sites compared with other sites with similar conservation rank (Fig 2). These results suggest that the strong long-range conservation gradient induced from catalytic residues cannot be solely attributed to the percolation of the strong selective pressure on them.

Catalytic site residues induce stronger conservation gradients than non-catalytic functional site residues with similar evolutionary rates

We have shown that on average, catalytic site residues induce stronger conservation gradients than expected by their average evolutionary rate. Next, we further reinforce this result by comparing the conservation gradient induced from subsets of catalytic and non-catalytic site residues with similar evolutionary rates. We have sampled 1,000 random subsets of residues from each type of functional site. For each such subset (represented by a circle in Fig 4), we calculated the average evolutionary rate and average conservation gradient. As expected, catalytic site residue subsets span the lowest dN/dS values (x-axis), followed by non-catalytic ligand binding sites and allosteric sites, protein-protein interaction sites, ligand binding sites in non-enzymes and finally buried residues. The linear trend between evolutionary rate and conservation gradient holds for each of these functional site types (Fig 4). This result indicates the robustness of the conservation-percolation linear trend regardless of the functional or structural role of the residues. Interestingly, residue subsets from different functional sites with similar evolutionary rates induce conservation gradients with different magnitudes. In particular, catalytic sites induce significantly stronger long-range conservation gradients than all other non-catalytic sites with similar evolutionary rates, including non-catalytic ligand binding sites (Fig 4). Finally, protein-protein interaction site residues induce lower conservation gradients than most other functional site residues with similar evolutionary rates, possibly due to the tendency for protein-protein interactions to rewire during evolution.

Catalytic site residues often induce stronger conservation gradients than more conserved non-catalytic functional site residues within the same protein

We have shown that conservation gradients induced by functional site residues in general correlate linearly with the evolutionary rates of these functional site residues. We have also shown that catalytic site residues are special in that they induce the strongest conservation gradients among subsets of different functional site residues with similar average evolutionary rates. However, our analyses have so far been carried out by grouping together functional site residues from different proteins. In this section, we perform a stringent, per-protein analysis by focusing on multi-functional proteins with at least two distinct functional sites and comparing the evolutionary properties of different functional site residues within the same protein. Some proteins in our dataset contain both catalytic sites as well as non-catalytic functional sites, whereas other proteins in our dataset contain two distinct non-catalytic functional sites. We found that for the majority of the functional site residue pairs within the same multi-functional protein, the more conserved functional site residue indeed induces stronger conservation gradient (86% if the more conserved residue is catalytic, Fig 5A; 67% if both residues are non-catalytic, Fig 5B). These results agree with the hypothesis that induced conservation gradients are largely driven by the percolation of selective pressure acting on functional sites.

Remarkably, in cases where the catalytic site residue is less conserved than the non-catalytic functional site residue within the same multi-functional protein, the catalytic site residue still induces stronger conservation gradient than the non-catalytic functional site for many of these cases (discordance of 49%, Fig 5C). This large discordance shown in cases where the lower

conservation residue is a catalytic site residue, is significantly higher compared to the cases where the lower conservation residue is a non-catalytic site residue (binomial test $P < 0.01$). These results clearly show that within the same protein, less conserved catalytic site residues often induce stronger conservation gradient than the more conserved non-catalytic site. Therefore, the strong conservation gradients from catalytic sites cannot be entirely explained by the percolation of the strong selective pressure acting on the catalytic sites.

Discussion

In this paper we have shown a linear relationship between two measures of evolutionary importance of protein sites under purifying selection. These are the degree of evolutionary conservation of the site itself, as well as the percolation of evolutionary conservation induced from the site via neighboring residues in the protein tertiary structure. Despite major differences between these two measures, we have shown that the linear relationship between the two measures is universal as it holds for different types of proteins as well as for different types of functional sites in proteins. However, catalytic sites in enzymes are the principal exception to this rule. We have shown here that catalytic sites in enzymes induce significantly stronger conservation gradients in the rest of the protein than expected from the degree of conservation of the site alone. Catalytic sites have a unique and complex functionality as they both bind a substrate as well as reduce the free energy barrier required for a chemical reaction to occur. These catalytic sites were shown to be under stronger selective pressure compared with other functional sites such as protein-protein interaction sites and ligand binding sites(13,15). It was also shown that they induce a significantly stronger long-range evolutionary rate gradient than other functional sites. One hypothesis regarding the origin of the strong conservation gradient

from catalytic sites is that it is simply due to the strong selective pressure acting on these sites percolating through the rest of the protein via residue-residue contacts. However, we have shown here that the strong selective pressure acting on catalytic sites cannot entirely explain the strong conservation gradients induced from catalytic sites.

The main determinant of the long-range conservation gradient from catalytic sites is still not completely understood. Local structural constraints (such as residue burial and packing) are usually potential contributors as they are known to generally have a significant effect on residue evolutionary rate(1,2,5,10). However, it was shown that local structural constraints are not the main determinants of long-range conservation gradients in enzymes(13,25,26). Moreover, the fact that non-catalytic ligand binding sites and allosteric binding sites induce significantly weaker evolutionary rate gradients implies that the ligand binding and allosteric function are not the main determinants of the long-range conservation gradients in enzymes either(15). We are therefore left with the hypothesis that the uniquely strong conservation gradient in enzymes is imposed by the special requirement for catalytic sites to differentially bind the transition state of a chemical reaction rather than the reactants or products with very similar properties (16), a function which is unique to catalytic sites compared with other functional binding sites. A recent physical model of residue evolutionary rates in enzymes introduced an activation term in addition to a stability term and showed improved predictive ability(27). The model attributed the activity to the free energy required to transform a distorted catalytic site upon mutation to its native conformation. The improved ability of the model supports the hypothesis that the main determinant of the observed long-range conservation gradient in enzymes is a functional rather than structural constraint. Overall, our results suggest that the stringent requirement for the

catalytic site to differentially bind to and stabilize the transition state of the catalyzed chemical reaction imposes extensive evolutionary constraints on a large portion of the enzyme beyond just the catalytic site, all of which play key roles in maintaining the catalytic function.

The results in this paper provide a complete, quantitative picture of evolutionary conservation patterns within proteins induced by functional sites. The linear relationship between the degree of conservation of a protein functional site and the induced conservation gradient in the rest of the protein suggests that the strong selective pressure acting on the functional site percolates through the rest of the protein via residue-residue contacts. Our results also clearly show the unique evolutionary behavior of enzymes in which the catalytic site induces significantly stronger evolutionary constraints on their surroundings than can be explained by the percolation theory alone. Moreover, our results emphasize that catalysis requires the participation of a much larger set of residues than just the few key catalytic residues. Finally, our proposed measure of the evolutionary importance of protein sites can assist in better phylogenetic inference, functional site prediction and protein design.

Methods

Protein dataset collection and functional site annotations

The current dataset is based on a dataset of structural homologs of yeast proteins(15). The dataset was created first by using gapped BLAST(28) searches between protein subunit sequences with solved structure from the Protein Data Bank(18) and 5,861 translated open reading frames (ORFs) of the yeast *Saccharomyces cerevisiae*(29). The ORF–subunit pairs were chosen such that both the subunit sequence and the ORF sequence had coverage of $\geq 50\%$ in the alignment and E-value $< 10^{-5}$ and could be paired with their orthologs in four other closely-related yeast species *S. paradoxus*, *S. mikatae*, *S. bayanus* and *S. pombe*. This way, 1,555 yeast ORFs were mapped to homologs in the PDB. The procedure included the following steps:

First, if one of the homology-based structural models of a yeast ORF had an annotated allosteric site, this model was chosen. For 115 yeast proteins the structural model was identified with a known allosteric site as well as pre-calculated conservation scores in ConSurf-DB(19). For all other yeast ORFs, if they had structural models with known ligand binding sites that do not overlap with catalytic sites and the closest catalytic site residue is at least 6 Å away (non-catalytic ligand binding site), the model with the lowest E-value out of them was chosen. Overall, 54 non-enzymes with 57 ligand binding sites and 41 enzymes with 47 non-catalytic ligand-binding sites were part of the dataset. For all other yeast ORFs, the structural model with the lowest E-value was chosen. In this manner, 1064 more ORFs for which the best structural model had pre-calculated conservation scores in ConSurf-DB were added to the dataset. Overall, 1,274 ORF-subunit pairs were included in the study. Out of them, 148 yeast proteins were identified with 280 protein-protein interactions sites and 348 proteins were identified with catalytic sites. Full

list of yeast proteins, their structural models along with identified functional sites can be found in Supplementary Table S1.

Functional site annotations

Allosteric sites within the structural subunits were found and their residues annotated using the Allosteric Database (ASD)(23,24). Biologically-relevant ligand-binding sites were found using Binding MOAD (Mother of all Databases)(30,31), a database of biologically significant protein-ligand binding in the PDB. Using MOAD, sites with bound crystallographic additives, buffers, salts, metals and sites with covalently linked ligands are excluded. Ligand-binding residues were identified using the BioLip(22) database. Catalytic sites within the protein structural subunits were found using the Catalytic Site Atlas(21). In order to find proteins that participate in protein-protein interactions in our dataset, we identified structural subunits where each subunit is both in physical contact with another subunit and the corresponding modelled ORFs are reported as interacting by at least one physical experiment in the BioGRID (32,33). Our dataset contains 148 proteins with 280 protein-protein interaction interfaces. Interfacial residues were identified as residues with different solvent accessibility values when in complex compared to when the interacting partner is manually deleted from the tertiary structure. Distances between residues were calculated as distances between their respective C α atoms. All functional site annotations of the chosen structural subunits were transferred to the yeast ORF sequence according to the sequence alignment.

Evolutionary conservation and rate calculations

In this study we calculated both evolutionary conservation scores for the residues of the structural subunits as well as average absolute evolutionary rates (dN/dS) for the yeast ORFs. Evolutionary conservation scores were taken from ConSurf-DB(19), which is a database of pre-calculated conservation scores of residues in proteins with known structures in the Protein Data Bank (PDB). ConSurf-DB conservation scores are based on collected sequence homologs of the PDB structure and using the Rate4Site algorithm(34).

To calculate the average evolutionary rates (dN/dS) for residues of *S. cerevisiae*, we first used the orthology assignment of the protein-coding genes of *S. cerevisiae* with four other closely-related yeast species (*S. paradoxus*, *S. mikatae*, *S. bayanus*, and *S. pombe*), according to the Fungal Orthogroup Repository(35). We then aligned the ORFs using MAFFT(36). Then, evolutionary rates were calculated using the program codeml within the PAML software package(37). The tree was specified as (((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*), *S. bayanus*), *S. pombe*). Codon frequencies were assumed equal (CodonFreq=0) and other parameters in codeml were left to their default values. The codon alignments can be found in Supplementary S2 File.

Statistical analysis

1000 random sample of 250 residues were collected from each type of functional site residues for Fig 4.

Estimated standard errors in our measurements of conservation gradients (Pearson correlations) and of dN/dS values were done using 50 rounds of bootstrap resampling.

Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council of Canada (grant numbers RGPIN-2019-05952, RGPAS-2019-00012), Canada Foundation for Innovation (grant number JELF-33732), and Canada Research Chairs program.

Figure Legends

Figure 1. Long-range conservation gradient induced from a protein residue is linearly correlated with its conservation within the protein. (a) Violin plots and respective average of long-range conservation gradient from a residue as a function of conservation rank for all residues in the dataset binned into 20 equally spaced bins of conservation rank. (b) Distribution of per-protein Pearson correlation between residues' conservation ranks and conservation gradients.

Figure 2. Residues in functional and structural sites exhibit linear correlation between conservation and conservation gradient. Violin plots of conservation gradient of residues as a function of their conservation rank within a protein for residues in different types of structural and functional sites.

Figure 3. Conservation gradient induced from residues is linearly (negatively) correlated with their evolutionary rate (dN/dS), with catalytic site residues inducing stronger conservation gradient than expected by the linear trend. Average conservation gradient (calculated as the average Pearson correlation between conservation of residues and their distance from a site) as a function of average evolutionary rate (dN/dS) for all yeast protein residues binned according to their annotated conservation rank into 100 equally spaced bins.

Figure 4. Catalytic site residues induce stronger conservation gradients than non-catalytic functional site residues with similar evolutionary rates. Average conservation gradients of subsets of functional residues. Each circle represents a subset of residues, colored by different types of functional sites.

Figure 5. Catalytic site residues often induce stronger conservation gradients than more conserved non-catalytic functional site residues within the same protein. Within the same protein, (A) more conserved catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, $P < 0.001$); (B) more conserved non-catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, $P < 0.001$); (C) less conserved catalytic site residues often induce stronger conservation gradient than more conserved non-catalytic site residues (discordance for half of the cases, binomial test, $P = 0.13$). Functional site residue pairs for which the ordering of residue conservation agrees with the ordering of induced conservation gradient (concordance) are marked in blue. Functional site residue pairs for which the ordering of residue conservation disagrees with the ordering of induced conservation gradient (discordance) are marked in orange.

References

1. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol.* 2009 Oct 16;10(10):709–20.
2. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet.* 2016;17(2):109–21.
3. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* 1992 Feb;1(2):216–26.
4. Conant GC, Stadler PF. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol Biol Evol.* 2009 May 1;26(5):1155–61.
5. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol.* 2009;26(10):2387–95.
6. Franzosa EA, Xue R, Xia Y. Quantitative residue-level structure-evolution relationships in the yeast membrane proteome. *Genome Biol Evol.* 2013 Jan 17;5(4):734–44.
7. Sharir-Ivry A, Xia Y. The impact of native state switching on protein sequence evolution. *Mol Biol Evol.* 2017 Feb;34(6):1378–90.
8. Hamelryck T. An amino acid has two sides: A new 2D measure provides a different view of solvent exposure. *Proteins.* 2005 Feb 1;59(1):38–48.
9. Yeh SW, Huang TT, Liu JW, Yu SH, Shih CH, Hwang JK, et al. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed Res Int.* 2014;572409.
10. Marcos ML, Echave J. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ.* 2015 Apr 23;3:e911.
11. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. *J Mol Biol.* 2002 Nov 15;324(1):105–21.
12. Tóth-Petróczy A, Tawfik DS. Slow protein evolutionary rates are dictated by surface-core association. *Proc Natl Acad Sci U S A.* 2011 Jul 5;108(27):11151–6.
13. Jack BR, Meyer AG, Echave J, Wilke CO. Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biol.* 2016 May 3;14(5):e1002452.
14. Nelson ED, Grishin N V. Evolution of off-lattice model proteins under ligand binding constraints. *Phys Rev E.* 2016 Aug 15;94(2):022410.
15. Sharir-Ivry A, Xia Y. Non-catalytic Binding Sites Induce Weaker Long-Range Evolutionary Rate Gradients than Catalytic Sites in Enzymes. *J Mol Biol.* 2019;431(19):3860–70.

- 423 16. Warshel A, Sharma PK, Kato M, Xiang Y, Liu H, Olsson MHM. Electrostatic basis for
424 enzyme catalysis. *Chem Rev.* 2006 Aug;106(8):3210–35.
- 425 17. Yang Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol*
426 *Evol.* 1996 Sep 1;11(9):367–72.
- 427 18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data
428 Bank. *Nucleic Acids Res.* 2000 Jan 1;28(1):235–42.
- 429 19. Goldenberg O, Erez E, Nimrod G, Ben-Tal N. The ConSurf-DB: pre-calculated evolutionary
430 conservation profiles of protein structures. *Nucleic Acids Res.* 2009 Jan;37(Database
431 issue):D323–7.
- 432 20. Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, et al. ConSurf: Using
433 Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Isr J Chem.* 2013
434 Apr 1;53(3–4):199–206.
- 435 21. Furnham N, Holliday GL, De Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM. The
436 Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes.
437 *Nucleic Acids Res.* 2014 Jan;42(D1):D485–9.
- 438 22. Yang J, Roy A, Zhang Y. BioLiP: A semi-manually curated database for biologically relevant
439 ligand-protein interactions. *Nucleic Acids Res.* 2013 Jan;41(D1):D1096–1103.
- 440 23. Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, et al. ASD: a comprehensive database of
441 allosteric proteins and modulators. *Nucleic Acids Res.* 2011 Jan 1;39(Database):D663–9.
- 442 24. Shen Q, Wang G, Li S, Liu X, Lu S, Chen Z, et al. ASD v3.0: unraveling allosteric regulation
443 with structural mechanisms and biological networks. *Nucleic Acids Res.* 2016 Jan
444 4;44(D1):D527–35.
- 445 25. Sharir-Ivry A, Xia Y. Nature of Long-Range Evolutionary Constraint in Enzymes: Insights
446 from Comparison to Pseudoenzymes with Similar Structures. *Mol Biol Evol.* 2018 Nov
447 1;35(11):2597–606.
- 448 26. Sharir-Ivry A, Xia Y. Using Pseudoenzymes to Probe Evolutionary Design Principles of
449 Enzymes. *Evol Bioinforma.* 2019 Jan 13;15:117693431985593.
- 450 27. Echave J. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with
451 Selection on Stability and Activity. *Mol Biol Evol.* 2019;36(3):613–20.
- 452 28. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
453 PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*
454 1997 Sep 1;25(17):3389–402.
- 455 29. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, et al. SGD: Saccharomyces
456 Genome Database. *Nucleic Acids Res.* 1998 Jan 1;26(1):73–9.
- 457 30. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All
458 Databases). *Proteins Struct Funct Bioinforma.* 2005 Jun 21;60(3):333–40.

31. Ahmed A, Smith RD, Clark JJ, Dunbar JB, Carlson HA. Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Res.* 2015 Jan 28;43(D1):D465–9.
32. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006 Jan 1;34(Database issue):D535-9.
33. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 2017;45(D1):D369–79.
34. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics.* 2002 Jul 1;18(Suppl 1):S71–7.
35. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature.* 2007 Sep 6;449(7158):54–61.
36. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol.* 2013 Apr;30(4):772–80.
37. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007 Aug 1;24(8):1586–91.

Figures

Figure 1. Long-range conservation gradient induced from a protein residue is linearly correlated with its conservation within the protein. (a) Violin plots and respective average of long-range conservation gradient from a residue as a function of conservation rank for all residues in the dataset binned into 20 equally spaced bins of conservation rank. (b) Distribution of per-protein Pearson correlation between residues' conservation ranks and conservation gradients.

Figure 2. Residues in functional and structural sites exhibit linear correlation between conservation and conservation gradient. Violin plots of conservation gradient of residues as a function of their conservation rank within a protein for residues in different types of structural and functional sites.

Fig 3. Conservation gradient induced from residues is linearly (negatively) correlated with their evolutionary rate (dN/dS), with catalytic site residues inducing stronger conservation gradient than expected by the linear trend. Average conservation gradient (calculated as the average Pearson correlation between conservation of residues and their distance from a site) as a function of the average evolutionary rate (dN/dS) for all yeast protein residues binned according to their annotated conservation rank into 100 equally spaced bins (black) as well as the average conservation gradients of different types of functional sites.

Figure 4. Catalytic site residues induce stronger conservation gradients than non-catalytic functional site residues with similar evolutionary rates. Average conservation gradients of subsets of functional residues. Each circle represents a subset of residues, colored by different types of functional sites.

Fig 5. Catalytic site residues often induce stronger conservation gradients than more conserved non-catalytic functional site residues within the same protein. Within the same protein, (A) more conserved catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, $P < 0.001$); (B) more conserved non-catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, $P < 0.001$); (C) less conserved catalytic site residues often induce stronger conservation gradient than more conserved non-catalytic site residues (discordance for half of the cases, binomial test, $P = 0.13$). Functional site residue pairs for which the ordering of residue conservation agrees with the ordering of induced conservation gradient (concordance) are marked in blue. Functional site residue pairs for which the ordering of residue conservation disagrees with the ordering of induced conservation gradient (discordance) are marked in orange.

Conservation gradient