Non-catalytic Binding Sites Induce Weaker Long-Range Evolutionary Rate Gradients than Catalytic Sites in Enzymes

Avital Sharir-Ivry and Yu Xia

Department of Bioengineering, *McGill University, Montreal, QC H3A 0E9, Canada*

Correspondence to Yu Xia: *Department of Bioengineering, McGill University, Montreal, QC H3A 0E9, Canada. avital.ivry@mail.mcgill.ca, brandon.xia@mcgill.ca*
https://doi.org/10.1016/j.jmb.2019.07.019

Abstract

Enzymes exhibit a strong long-range evolutionary constraint that extends from their catalytic site and affects even distant sites, where site-specific evolutionary rate increases monotonically with distance. While protein–protein sites in enzymes were previously shown to induce only a weak conservation gradient, a comprehensive relationship between different types of functional sites in proteins and the magnitude of evolutionary rate gradients they induce has yet to be established. Here, we systematically calculate the evolutionary rate (dN/dS) of sites as a function of distance from different types of binding sites in enzymes and other proteins: catalytic sites, non-catalytic ligand binding sites, allosteric binding sites, and protein–protein interaction sites. We show that catalytic sites indeed induce significantly stronger evolutionary rate gradient than all other types of non-catalytic binding sites. In addition, catalytic sites in enzymes with no known allosteric function still induce strong long-range conservation gradients. Notably, the weak long-range conservation gradients induced by non-catalytic binding sites in enzymes is nearly identical in magnitude to those induced by ligand binding sites in non-enzymes. Finally, we show that structural determinants such as local solvent exposure of sites cannot explain the observed difference between catalytic and non-catalytic functional sites. Our results suggest that enzymes and non-enzymes share similar evolutionary constraints only when examined from the perspective of non-catalytic functional sites. Hence, the unique evolutionary rate gradient from catalytic sites in enzymes is likely driven by the optimization of catalysis rather than ligand binding and allosteric functions.

Introduction

Enzymes are key players in metabolic pathways and are crucial for cell function. They bind their chemical reactants and enhance the production rate of chemical products by several orders of magnitude. The catalytic function occurs in the catalytic site, which is usually composed of relatively buried residues in a large cleft [1]. Due to their function, catalytic site residues are highly conserved and even their close vicinity residues are under strong selective pressure where their conservation decreases with distance from the catalytic site [1]. Recently, it was shown that enzymes exhibit a long-range, nearly linear conservation gradient from their catalytic site that extends even up to ~30 Å in distance [2].

Possible confounding factors of the observed long-range conservation gradient in enzymes include various local structural properties known to drive evolutionary rate variation of protein sites [3–12]. The main local structural determinants known are residue solvent accessibility and residue packing [6,8,9] such that residues are generally less conserved with increasing solvent exposure or decreasing packing. Indeed, catalytic sites are generally relatively buried and therefore could induce a conservation gradient based solely on solvent exposure gradients. However, for enzymes where the catalytic site is on the surface, a strong conservation gradient was observed as well [2], demonstrating that distance and solvent exposure contribute independently to the conservation

Table 1. Non-catalytic ligand binding sites induce weaker evolutionary rate gradients than catalytic sites

	Type of binding site	Slope	Evolutionary rate (dN/dS) of binding site residues
Enzymes	Catalytic sites	0.0029 (±0.0001)	0.015 (±0.003)
Enzymes	Non-catalytic ligand-binding site	0.0016 (±0.0002)	0.034 (±0.007)
Non-enzymes	Non-catalytic ligand-binding site	0.0017 (±0.0003)	0.033 (±0.005)

Slope of the linear fit of the average evolutionary rate (dN/dS) *versus* distance from catalytic sites and non-catalytic ligand-binding sites, as well as the average evolutionary rate of binding site residues. In brackets—standard errors.

gradient. Still, given that protein fold usage is very different between enzymes and non-enzymes [13–15], it is possible that other local structural determinants (other than solvent exposure and residue packing) could potentially cause the strong conservation gradient from catalytic sites in enzymes. However, we have previously shown that catalytically inactive pseudoenzymes with nearly identical tertiary structure as enzymes exhibit a significantly reduced conservation gradient, thereby demonstrating that the observed conservation gradient in enzymes cannot be dominated by any backbone-based local structural determinants [16,17].

This study addresses the following outstanding question: Are catalytic sites unique in their abilities to induce such strong long-range conservation gradients? What about other non-catalytic functional sites which share similar functional capacities of binding another molecule as catalytic binding sites? Using off-lattice protein model, it was shown that the requirement to maintain a specific ligand-binding site gives rise to a conservation gradient from the ligand-binding site [18,19]. For protein–protein interaction sites, however, it was shown that interfacial sites induce significantly weaker conservation gradients than catalytic sites in enzymes [2]. Here, we present a data-driven study of the evolutionary rate (dN/dS) of residues as a function of their distance from different types of functional sites in proteins. We consider four types of functional sites: catalytic sites, non-catalytic ligand binding sites, allosteric binding sites, and protein–protein interaction sites. We show that all types of non-catalytic binding sites induce significantly weaker long-range evolutionary rate gradients than catalytic sites in enzymes. Surprisingly, non-catalytic ligand-binding sites in enzymes do not induce significant long-range evolutionary rate gradients. Instead, the weak evolutionary rate gradient induced from non-catalytic ligand-binding sites in enzymes resembles that from ligand-binding sites in non-enzymes. Moreover, we show that catalytic sites in enzymes with no known allosteric function still induce a strong long-range evolutionary

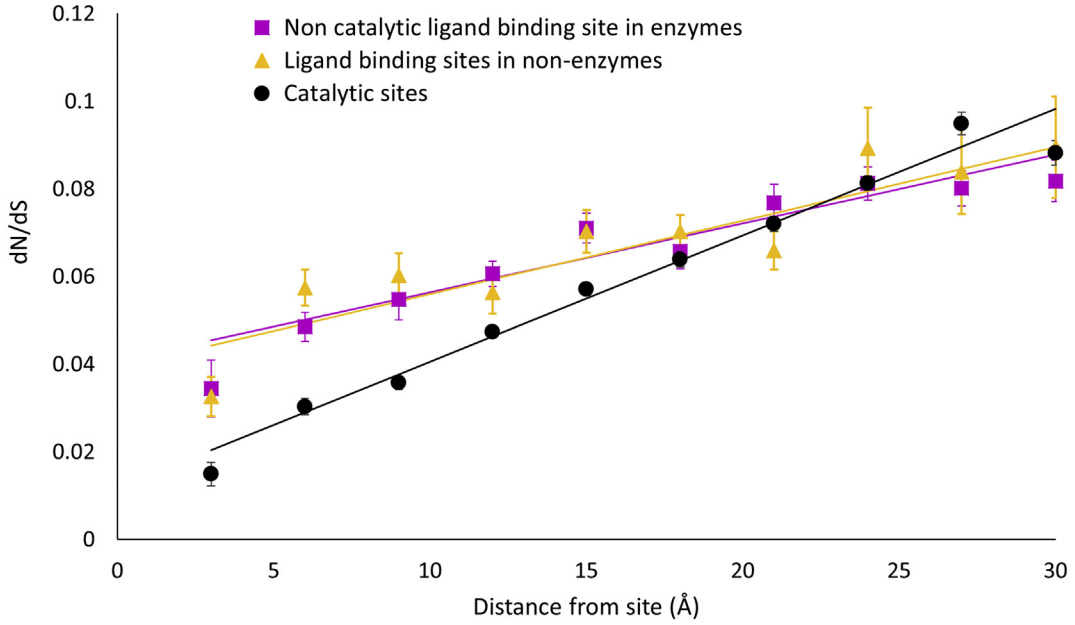

Fig. 1. Non-catalytic ligand binding sites induce weaker evolutionary rate gradients than catalytic sites. Evolutionary rate (dN/dS) as a function of distance from catalytic and non-catalytic ligand-binding sites in enzymes, as well as from ligand-binding sites in non-enzymatic proteins.

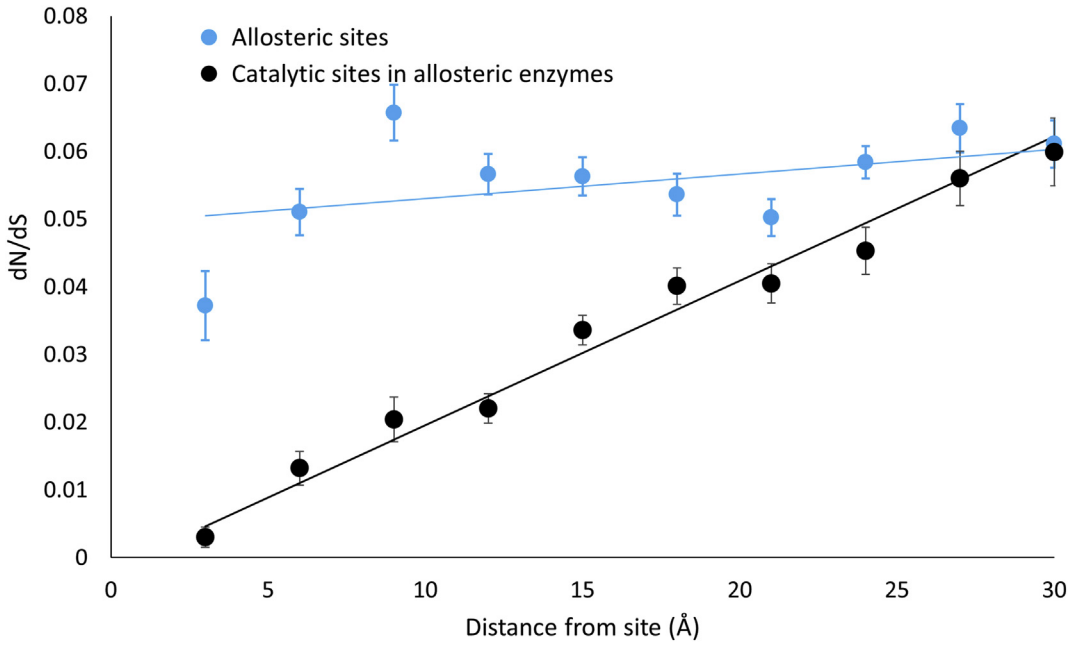

Fig. 2. Allosteric binding sites induce weaker evolutionary rate gradients than catalytic sites. Evolutionary rate (dN/dS) as a function of distance from allosteric binding sites and from catalytic sites in allosteric enzymes.

rate gradient, indicating that allosteric function is not the main driving force of the observed evolutionary rate gradients from catalytic sites. Lastly, we show that solvent exposure gradients cannot explain the differences in magnitude between evolutionary rate gradients induced from catalytic and non-catalytic ligand-binding sites. Taken together, our results suggest that the observed evolutionary rate gradient from catalytic sites in enzymes is primarily driven by the optimization and maintenance of catalytic function rather than ligand-binding or allosteric function.

Results

Non-catalytic ligand-binding sites induce weaker evolutionary rate gradients than catalytic sites

To examine whether evolutionary rate gradients are also induced from non-catalytic ligand-binding sites, we first identified such functional binding sites within enzymes and non-enzymes. As a starting point, we used a data set of 1744 structurally annotated yeast proteins (see Methods). We screened them to find yeast proteins for which the structural model contains a biologically significant ligand bound to it based on the Binding MOAD database [20]. We found 109 ligand-binding sites on 95 enzymes, which do not overlap with the catalytic site on these enzymes, and 71 ligand-binding sites on 65 proteins, which are not known to be enzymes (non-enzymes). On average, the evolutionary rates of the structurally modeled non-enzymes and enzymes in our data sets are not significantly different (0.0687 ± 0.0019 and 0.0697 ± 0.0006, respectively). Ligand-binding residues were identified from BioLip [21], and catalytic site residues in enzymes were identified using the Catalytic Site Atlas [22].

For each residue, we calculated the distance to the closest ligand-binding residue. Average evolutionary rate (dN/dS) was then calculated for residues over

Table 2. Allosteric sites induce weaker evolutionary rate gradients than catalytic sites and catalytic sites in non-allosteric enzymes induce strong evolutionary rate gradients

	Type of functional site	Slope	Evolutionary rate (dN/dS) of functional site residues
Allosteric enzymes	Catalytic sites	0.0021 (±0.0001)	0.003 (±0.002)
Allosteric enzymes	Allosteric sites	0.0004 (±0.0002)	0.037 (±0.005)
Non-allosteric enzymes	Catalytic sites	0.0030 (±0.0002)	0.029 (±0.006)

Slope of the linear fit of the average evolutionary rate (dN/dS) *versus* distance from allosteric sites and catalytic sites in allosteric enzymes and non-allosteric enzymes, as well as the average evolutionary rate of the binding site residues. In brackets—standard errors.

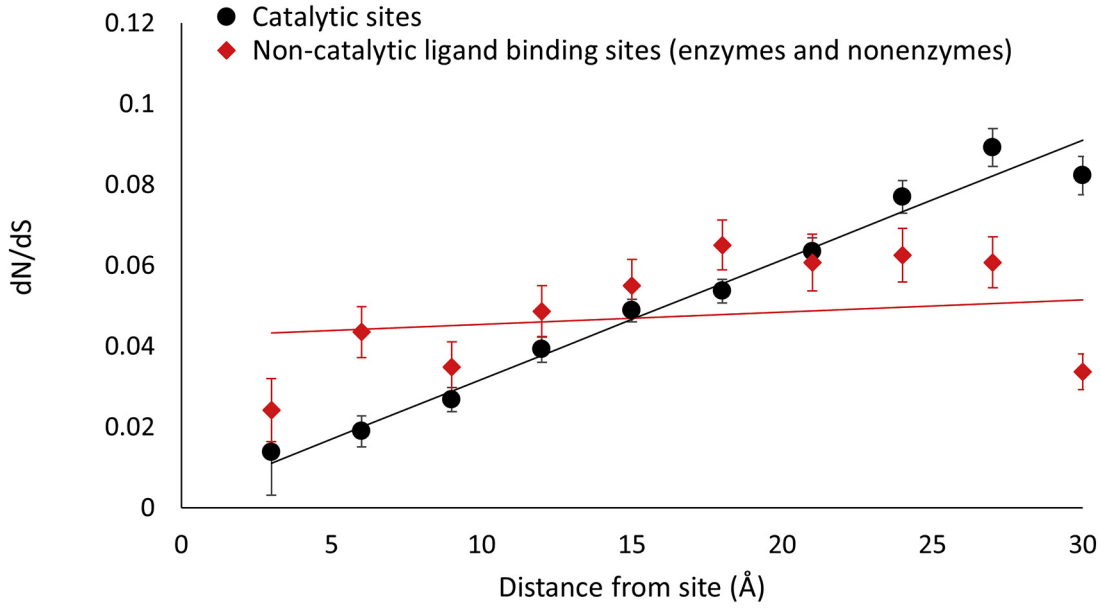

Fig. 3. Yeast-annotated non-catalytic ligand binding sites induce weaker evolutionary rate gradients than yeast-annotated catalytic sites. Evolutionary rate (dN/dS) as a function of distance from catalytic and non-catalytic ligand-binding sites (including non-catalytic binding sites in enzymes as well as in non-enzymatic proteins) for the subset of *S. cerevisiae* proteins known to contain the respective functional binding sites.

distance bins to obtain the evolutionary rate gradients.

The slope of the linear fit for the evolutionary rate gradient from catalytic sites is significantly larger than the slope for the evolutionary rate gradient from non-catalytic ligand-binding sites (*t*-test, $P < 0.001$; Table 1 and Fig. 1). Surprisingly, the evolutionary rate gradient induced from non-catalytic ligand-binding sites in enzymes is weak and actually resembles the evolutionary rate gradient induced from ligand-binding sites in non-enzymes. The weak evolutionary rate gradient from non-catalytic ligand-binding sites suggests that ligand-binding functionality of catalytic sites is probably not the main determinant of the strong evolutionary rate gradients induced from them.

In those enzymes with both catalytic and non-catalytic binding sites, the average distance of residues from the catalytic site is higher than 15 Å throughout the different distance bins from the non-catalytic binding site (Fig. S1), suggesting that the relationship between evolutionary rate and distance from one functional site is not strongly affected from the presence of the second functional site.

In addition, we have also calculated the fraction of residues for which a non-synonymous mutation exists (mutation that leads to a change in amino acid residue) as a function of the distance from the functional sites. The mutability of sites increases with distance from catalytic sites, while this increase is shallower from non-catalytic ligand binding sites in enzymes and non-enzymes (Fig. S2 in the Supplementary Material).

Yeast-annotated non-catalytic ligand-binding sites induce weaker evolutionary rate gradients than catalytic sites

The annotations of functional sites were made according to the homology-based structural models

Table 3. Yeast-annotated non-catalytic ligand binding sites induce weaker evolutionary rate gradients than yeast-annotated catalytic sites

Type of binding site	Slope	Evolutionary rate (dN/dS) of binding site residues
Catalytic sites	0.0030 (±0.0002)	0.014 (±0.011)
Non-catalytic ligand binding sites	0.0003 (±0.0005)	0.024 (±0.008)

Slope of the linear fit of the average evolutionary rate (dN/dS) *versus* distance from catalytic sites and non-catalytic ligand-binding sites, as well as the average evolutionary rate of binding site residues for yeast proteins known to contain functional binding sites. In brackets—standard errors.

of the *Saccharomyces cerevisiae* proteins. These models are usually based on solved structures of homologous proteins in other species. In order to rule out the confounding factor that some of our homology-based annotations of functional sites are false positives in yeast, we examined a subset of yeast proteins that are also known to have the relevant binding functionality in yeast. The results for these yeast proteins with high-confidence annotations in Fig. 2 and Table 2 clearly show a significant difference between the strong evolutionary rate gradients induced from catalytic sites in enzymes and all other non-catalytic binding sites. Due to the significantly smaller size of the yeast-annotated data set, we grouped non-catalytic ligand binding sites in enzymes and non-enzymes together. Overall, our results are robust to possible false positives of our homology-based functional binding site annotations for yeast proteins.

Allosteric binding sites induce weaker long-range evolutionary rate gradients than catalytic sites

Next, we focused on the special case of allosteric binding sites. We collected those yeast proteins that have a structural model, which is known to have an allosteric function with its allosteric site residues annotated in the allosteric database (ASD) [23]. We found 153 allosteric proteins, of which 81 are known enzymes with known catalytic site residues. All residues were binned according to their distance from the closest allosteric binding residue as well as from the closest catalytic residue in the case of enzymes. Average evolutionary rate (dN/dS) was calculated for the residues in each distance bin.

The slope of the evolutionary rate gradient induced from allosteric binding sites is significantly smaller than that induced from catalytic sites (*t*-test, $P < 0.001$; Fig. 3 and Table 3). Allosteric binding sites in enzymes modulate the activity of catalytic sites in that the catalytic site is shifting into an alternative conformation upon a binding event in the distant allosteric binding site [24,25]. Despite this long-range interaction between catalytic sites and allosteric binding sites, our results suggest that allosteric function is not the main determinant of the long-range evolutionary rate gradient induced from catalytic sites in enzymes.

Catalytic sites in non-allosteric enzymes induce strong evolutionary rate gradients

For comparison, we also identified "non-allosteric enzymes" in our data set as those proteins with an enzyme structural model with known catalytic residues but with no known allosteric function (219 proteins). Notably, similar to allosteric enzymes, catalytic sites in non-allosteric enzymes also exhibit a strong evolutionary rate gradient that extends to distant sites (Fig. 4 and Table 3). The existence of evolutionary rate gradient from catalytic sites appears to be independent of the existence of allosteric function in the enzyme. This result further supports

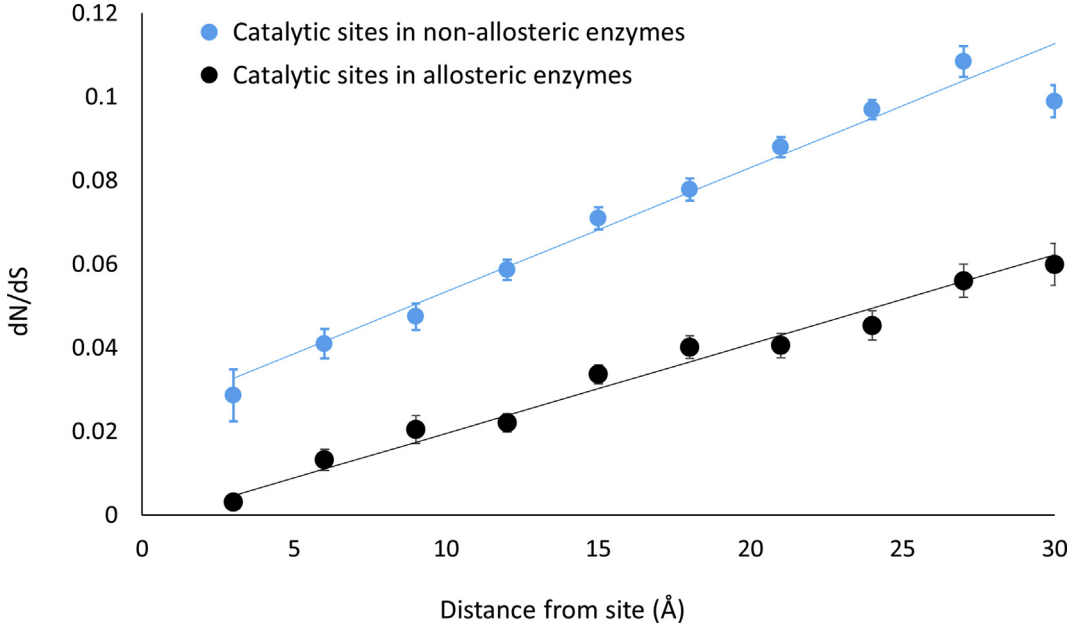

Fig. 4. Catalytic sites in non-allosteric enzymes induce strong evolutionary rate gradients. Evolutionary rate (dN/dS) as a function of distance from catalytic sites in allosteric and non-allosteric enzymes.

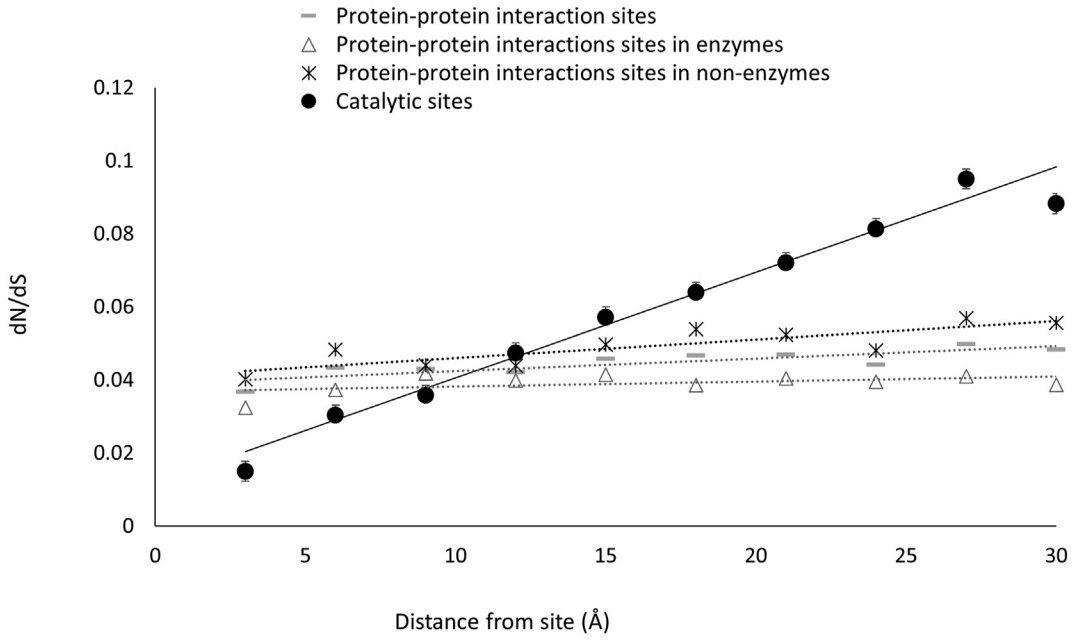

Fig. 5. Protein–protein interaction sites in enzymes or non-enzymes induce weaker long-range evolutionary rate gradients than catalytic sites. Evolutionary rate (dN/dS) as a function of distance from catalytic sites and from protein–protein interaction sites in enzymes and non-enzymes.

the hypothesis that the allosteric function is not the main determinant of evolutionary rate gradients induced from catalytic site in enzymes.

Protein–protein interaction sites induce weaker long-range evolutionary rate gradients than catalytic sites

Using relative conservation scores, it was previously shown that protein–protein interactions sites in enzymes induce only a minor conservation gradient, which is significantly weaker than that from catalytic sites [2]. Here, we studied residue evolutionary rate (dN/dS) as a function of distance from protein–protein interaction sites in yeast. We screened the structurally annotated data set of yeast proteins to find proteins that are known to interact with other proteins (see Methods). We found 436 interfacial sites for 231 proteins. Interfacial residues were identified as those residues that have different solvent accessibility when in complex compared to when the interacting partner is deleted. Residues were binned according to their distance from interfacial residues, and average dN/dS was calculated for each bin. The slope of the evolutionary rate gradient induced from interfacial sites is indeed significantly smaller than that induced from catalytic sites (*t*-test, $P < 0.001$; Fig. 5 and Table 4), either in enzymes or in non-enzymes.

In this study, we lumped together different types of protein interaction sites whether the interaction partner is a structured protein or a disordered one. It was shown that there are different types of protein disorder with different evolutionary conservation patterns that are related to different functions [26], and it would therefore be interesting in

Table 4. Protein–protein interaction sites in enzymes or non-enzymes induce weaker long-range evolutionary rate gradients than catalytic sites

Protein–protein interaction sites on:	Slope	Evolutionary rate (dN/dS) of binding site residues
All proteins	0.0004 (±0.0001)	0.037 (±0.001)
Enzymes	0.0002 (±0.0001)	0.032 (±0.002)
Non-enzymes	0.0005 (±0.0001)	0.040 (±0.002)

Slope of the linear fit of the average evolutionary rate (dN/dS) *versus* distance from protein–protein interaction sites, as well as the average evolutionary rate of interfacial residues. In brackets—standard errors.

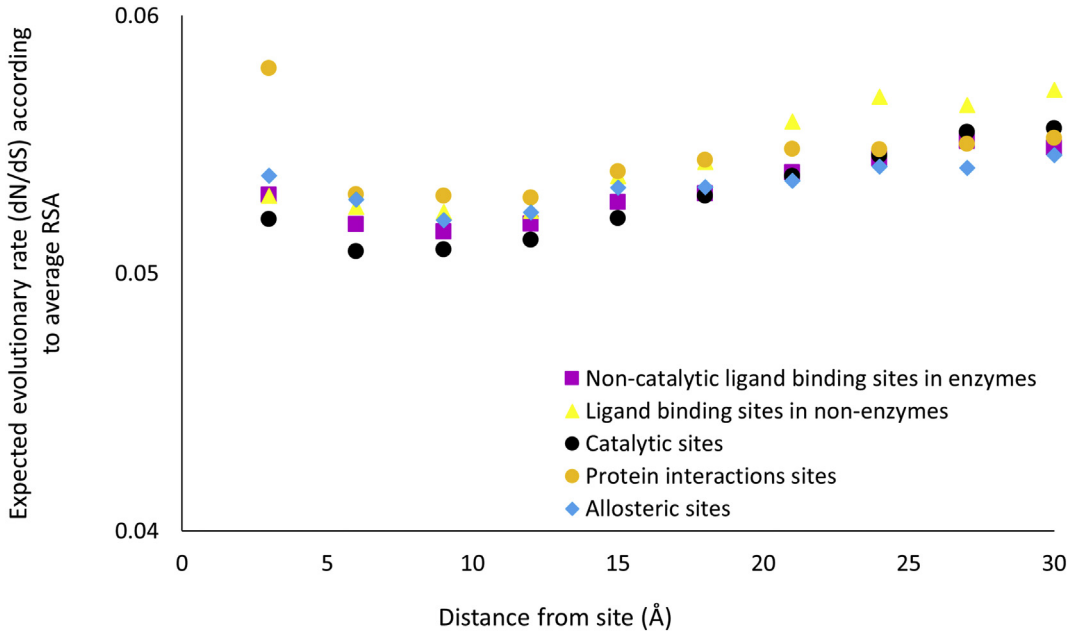

Fig. 6. Evolutionary rate based on solvent exposure alone does not show an increase with distance from catalytic and non-catalytic binding sites. Expected dN/dS according to average RSA of residues as a function of distance from binding site.

future studies to further explore whether the type of interaction site affects the conservation gradient induced from it.

Evolutionary rate gradient variations among different functional sites cannot be explained by solvent exposure gradients

We examined whether local structural determinants such as solvent exposure gradients are responsible for the observed difference in magnitude of evolutionary rate gradients induced from catalytic and non-catalytic binding sites. Site-specific solvent exposure and packing are known to be the strongest structural determinants to correlate with protein site evolutionary rate. We calculated the relative solvent accessibility (RSA) of each residue and then the expected linear trend of dN/dS *versus* RSA over all the proteins in the data set. We then used the calculated slope and intercept of this dN/dS *versus* RSA relationship (Fig. S3 in the Supplementary Material) to plot the expected evolutionary rate at each distance bin according to the average RSA of the residues at that bin. As shown in Fig. 6, there is no significant expected increase of evolutionary rate with distance based on solvent exposure alone for any of the data sets of functional binding sites. We therefore conclude that solvent exposure patterns cannot explain the differences in evolutionary rate gradients induced from catalytic and non-catalytic binding sites.

Discussion

Strong long-range rate gradient is observed from catalytic sites in enzymes where there is a monotonic, nearly linear increase in selective pressure on residues with their proximity to the catalytic site. We studied here functional non-catalytic ligand-binding sites in enzymes and non-enzymes to examine whether they induce similar evolutionary rate gradients as enzymatic catalytic sites. For this purpose, we used structurally annotated yeast proteins as a model to systematically study site-specific evolutionary rates as a function of distance from functional binding sites. We show that non-catalytic binding sites, even in enzymes, induce significantly weaker evolutionary rate gradients than catalytic sites.

Previous studies show that evolutionary rate gradients in enzymes cannot be explained by local structural properties of the protein [2,16], suggesting that the gradient is driven by functional determinants. A recent biophysical model of enzyme evolution that incorporates enzymatic activity constraint along with stability constraints better explains the evolutionary rate gradient from catalytic sites [27]. Enzymatic function is very complex, involving both substrate binding and accelerating the production rate of chemical products by decreasing the free energy barrier of the chemical reaction. By showing that non-catalytic ligand-binding sites induce significantly weaker evolutionary rate gradients than catalytic

sites, we can rule out several hypotheses on the origin of the long-range evolutionary rate gradient. The first hypothesis that can be ruled out is that the ligand-binding function of the catalytic site is the main determinant of the long-range rate gradient. While ligand binding is part of the function of the catalytic site, it appears improbable that the specific binding of the substrate is the main determinant of the long-range selection pressure in enzymes.

The second hypothesis that can be ruled out is that the long-range evolutionary rate gradient from catalytic sites is mainly driven by the allosteric function of the enzyme. Allosteric function exhibits a long-range effect in which binding of an effector molecule at a distant allosteric site shifts the catalytic site into alternative conformation [24,25]. It is reasonable to expect that the need to maintain proper long-range allosteric function can explain the observed long-range evolutionary rate gradient from the catalytic site. On the contrary, we have shown that allosteric binding sites induce a significantly weaker evolutionary rate gradient than catalytic sites. Moreover, a strong evolutionary rate gradient exists even in enzymes that are not known to be allosteric. These results suggest that optimizing the allosteric function of the enzyme is not the main determinant of the observed evolutionary rate gradient.

Finally, we show that for all the functional sites we have studied here, there is no significant solvent exposure gradients extending from them. Hence, the observed evolutionary rate gradient variations among different functional sites cannot be explained by known local structural determinants such as solvent exposure gradients.

We are therefore left with two plausible hypotheses on the origin of the unique evolutionary rate gradient from catalytic sites. The first hypothesis is that strongly conserved sites in general induce stronger long-range evolutionary rate gradient. Here, the evolutionary rate gradient is a result of a percolating effect of penetrating conservation [19,28,29]. Thus, the higher the selective pressure on a site is, the stronger the percolation of evolutionary rates will be from it. The second hypothesis is that evolutionary rate gradients are driven by the chemistry of the enzyme where the main determinant of the increasing selective pressure is to optimize the catalytic power of the enzyme by specifically stabilizing the reaction transition state [30]. Indeed, a chemistry-centered view on the evolution of new catalytic functions in enzymes suggests that new function in an enzyme can evolve from the ability of the enzyme to stabilize the transition state of the new chemical reaction (rather than just binding of the new ligand) [31–33]. These two hypotheses are both consistent with the observations in this study, and they are not mutually exclusive. Further work is needed to critically examine the validity and applicability of these hypotheses in terms of explaining the origin of the long-range evolutionary rate gradient observed in enzymes.

Methods

Structural annotation of yeast proteins

We based our study on a data set of structural homologs of yeast proteins. This data set was created using gapped BLAST [34] searches between protein subunit sequences with solved structure from the Protein Data Bank [35] and 5861 translated open reading frames (ORFs) of the yeast *S. cerevisiae* [36]. We kept those ORF–subunit pairs in which 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 relative yeast species *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, and *Saccharomyces pombe*. This way, 1755 yeast ORFs were mapped to a homolog in the PDB. The site-specific structural features of the chosen PDB homolog were transferred to the query proteins according to the sequence alignment.

Annotation of ligand-binding sites and catalytic sites

The structural subunits were screened to find those with an identified biologically relevant ligand-binding site according to Binding MOAD [20,37], which is a database of biologically significant protein-ligand binding in the PDB. Binding MOAD filters out cases in which the ligands are crystallographic additives, buffers, salts, metals, and covalently linked ligands. For each ORF with ligand-bound structural models, the model that produced the lowest alignment *E*-value with the ORF was chosen. Ligand-binding residues were identified using BioLip [21] database. For all other yeast ORFs, the subunit that produced the lowest alignment *E*-value with the ORF was chosen as the structural homolog. Yeast ORF was classified as an enzyme if its paired structural subunit has a catalytic site annotated in the Catalytic Site Atlas [22].

A total of 398 enzymes with annotated catalytic site were identified from which 95 enzymes have additional 109 ligand-binding sites that do not overlap with their catalytic sites. To find cases of ligand binding sites in non-enzymes, we identified the proteins for which the structural model is not assigned with an EC number and in addition is not

part of sequence cluster of at least 70% identity with other sequences in the PDB [38] for which an EC number is assigned. Sixty-five proteins with 71 ligand binding sites were identified as non-enzymes.

We have also created a subset of enzymes and non-enzymes that contains only yeast proteins which are known to contain the relevant functional site that included 93 enzymes with annotated catalytic sites and 43 proteins with 48 non-catalytic ligand binding sites.

The distance of the Cα atom of each residue in these proteins to the closest Cα of the binding site residues and catalytic site residues was calculated. Residues were binned according these distances up until 30 Å into equally spaced bins of 3 Å in size. The first bin essentially contained all the ligand-binding residues and only them.

Annotation of allosteric proteins

Using the data set of proteins with known allosteric function from the allosteric database (ASD) [23], we screened the yeast ORFs by means of their structural subunit to those that are known to have an allosteric function and their allosteric sites residues are known. One hundred fifty-six yeast ORFs matched this criterion, and out of them, we kept 153 proteins for which the structural alignment contained the allosteric binding sites. Within the data set, 81 proteins are identified as enzymes with an EC number and with known catalytic site from the Catalytic Site Atlas [22]. Other proteins in our data set that have an EC number assigned and their catalytic site residues are identified, but whose structural models are not known to have an allosteric function (219 proteins), were classified as "non-allosteric" enzymes. The distance of the Cα atom of each residue to the closest Cα atom of allosteric site residue and closest Cα atom of catalytic site residue was calculated. Residues were binned according these distances up until 30 Å into equally spaced bins of 3 Å in size.

Protein interfaces

We screened the protein complexes from which the best ORF–subunit were derived to those that are in physical contact with a different ORF–subunit pair (not necessarily the pair with highest similarity) and were reported as interacting with the other ORF–subunit pair by at least one physical experiment in the BioGRID [39,40]. Our data set contained 231 yeast proteins with 436 interfaces. Interfacial residues were identified as residues that have different solvent accessibility values when in complex compared to when the interaction partner is manually deleted.

All of the optimal ORF–subunit pairs and the corresponding subunit binding site residues in this study are listed in Supplementary Table S1 in the Supplementary Material and their alignments in Supplementary File S1.

Solvent exposure annotation

Solvent-accessible surface area (SASA) for the residues was calculated using the DSSP program [41] with hydrogen atoms excluded. SASA values were normalized by a set of reference values that correspond to the maximal ASA for each residue to account for differences in the sizes and empirical SASA distributions of the 20 different amino acid residue types [42]. This procedure produced the RSA, with values ranging from 0 (for a completely buried residue) to 1 (for a residue fully exposed to the solvent).

Evolutionary rate calculations

Orthology assignment between protein-coding genes of *S. cerevisiae* and four other related species *S. paradoxus*, *S. mikatae*, *S. bayanus*, and *S. pombe* was taken from the Fungal Orthogroup Repository [43]. We then aligned the orthologous proteins using MAFFT [44]. The aligned residues were binned according to their structural annotations as explained above, and their average evolutionary rate was calculated. The measure we used to calculate evolutionary rate is dN/dS, defined as the number of non-synonymous substitutions per non-synonymous site divided by the number of synonymous substitutions per synonymous site [45]. dN/dS is a popular measure of the direction and strength of selection acting on protein-coding genes. Assuming that synonymous substitutions are neutral, then dN/dS is less than 1 for purifying or negative selection, greater than 1 for Darwinian or positive selection, and equal to 1 for neutral evolution. dN/dS values were calculated over the multiple sequence alignments using the program codeml within the PAML software package [46]. 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 sequence alignments can be found in Supplementary File S2 in the Supplementary Material.

Statistical analysis

For each bin, we estimated the standard error in our measurements of dN/dS using 50 rounds of bootstrap resampling. We used a weighted least squares regression to fit dN/dS *versus* distance where the standard errors of dN/dS in each bin were considered such that distance bins with small dN/dS estimation errors receive greater weight in the line fitting process. We used two-tailed *t*-tests to

compute the significance of the difference between slopes.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.07.019.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (Grant Nos. RGPIN-2019-05952 and RGPAS-2019-00012 to Y.X.), Canada Foundation for Innovation (Grant Nos. JELF-33732 and IF-33122 to Y.X.), and Canada Research Chairs program (to Y.X.).

Received 29 March 2019;
Received in revised form 26 June 2019;
Accepted 11 July 2019
Available online 17 July 2019

Keywords:
Ligand binding sites;
protein–protein interaction sites;
allosteric sites;
evolutionary rate (dN/dS);
enzyme evolution

Abbreviations used:
RSA, relative solvent accessibility; ORF, open reading frame; SASA, solvent-accessible surface area.

References

[1] G.J. Bartlett, C.T. Porter, N. Borkakoti, J.M. Thornton, Analysis of catalytic residues in enzyme active sites, J. Mol. Biol. 324 (2002) 105–121.
[2] B.R. Jack, A.G. Meyer, J. Echave, C.O. Wilke, Functional sites induce long-range evolutionary constraints in enzymes, PLoS Biol. 14 (2016), e1002452.
[3] E.A. Franzosa, R. Xue, Y. Xia, Quantitative residue-level structure-evolution relationships in the yeast membrane proteome, Genome Biol. Evol. 5 (2013) 734–744.
[4] J.D. Bloom, D.A. Drummond, F.H. Arnold, C.O. Wilke, Structural determinants of the rate of protein evolution in yeast, Mol. Biol. Evol. 23 (2006) 1751–1761.
[5] J. Echave, S.J. Spielman, C.O. Wilke, Causes of evolutionary rate variation among protein sites, Nat. Rev. Genet. 17 (2016) 109–121.
[6] M.L. Marcos, J. Echave, Too packed to change: side-chain packing and site-specific substitution rates in protein evolution, PeerJ 3 (2015), e911.
[7] A. Oberai, N.H. Joh, F.K. Pettit, J.U. Bowie, Structural imperatives impose diverse evolutionary constraints on helical membrane proteins, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 17747–17750.
[8] A. Sharir-Ivry, Y. Xia, The impact of native state switching on protein sequence evolution, Mol. Biol. Evol. 34 (2017) 1378–1390.
[9] E.A. Franzosa, Y. Xia, Structural determinants of protein evolution are context-sensitive at the residue level, Mol. Biol. Evol. 26 (2009) 2387–2395.
[10] J. Overington, D. Donnelly, M.S. Johnson, A. Sali, T.L. Blundell, Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds, Protein Sci. 1 (1992) 216–226.
[11] G.C. Conant, P.F. Stadler, Solvent exposure imparts similar selective pressures across a range of yeast proteins, Mol. Biol. Evol. 26 (2009) 1155–1161.
[12] D.C. Ramsey, M.P. Scherrer, T. Zhou, C.O. Wilke, The relationship between relative solvent accessibility and evolutionary rate in protein evolution, Genetics 188 (2011) 479–488.
[13] H. Hegyi, M. Gerstein, The relationship between protein structure and function: a comprehensive survey with application to the yeast genome, J. Mol. Biol. 288 (1999) 147–164.
[14] T.J.F. Day, A.V. Soudackov, M. Čuma, U.W. Schmitt, G.a. Voth, A second generation multistate empirical valence bond model for proton transport in aqueous systems, J. Chem. Phys. 117 (2002) 5839–5849.
[15] A.C. Martin, et al., Protein folds and functions, Structure 6 (1998) 875–884.
[16] A. Sharir-Ivry, Y. Xia, Nature of long-range evolutionary constraint in enzymes: insights from comparison to pseudoenzymes with similar structures, Mol. Biol. Evol. 35 (2018) 2597–2606.
[17] Sharir-Ivry, A. & Xia, Y. (2019). Using pseudoenzymes to probe evolutionary design principles of enzymes. Evol. Bioinforma. 15, 117693431985593.
[18] E.D. Nelson, N.V. Grishin, Evolution of off-lattice model proteins under ligand binding constraints, Phys. Rev. E 94 (2016), 022410.
[19] E.D. Nelson, N.V. Grishin, Long-range epistasis mediated by structural change in a model of ligand binding proteins, PLoS One 11 (2016), e0166739.
[20] L. Hu, M.L. Benson, R.D. Smith, M.G. Lerner, H.A. Carlson, Binding MOAD (Mother Of All Databases), Proteins Struct. Funct. Bioinforma. 60 (2005) 333–340.
[21] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, Nucleic Acids Res. 41 (2013) D1096–D1103.
[22] N. Furnham, et al., The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes, Nucleic Acids Res. 42 (2014) D485–D489.
[23] Q. Shen, et al., ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks, Nucleic Acids Res. 44 (2016) D527–D535.
[24] E.R. Stadtman, Allosteric regulation of enzyme activity, Adv. Enzymol. Relat. Areas Mol. Biol. 28 (1966) 41–154.
[25] J. Liu, R. Nussinov, Allostery: an overview of its history, concepts, methods, and applications, PLoS Comput. Biol. 12 (2016), e1004966.
[26] J. Bellay, et al., Bringing order to protein disorder through comparative genomics and genetic interactions, Genome Biol. 12 (2011) R14.
[27] J. Echave, Beyond stability constraints: a biophysical model of enzyme evolution with selection on stability and activity, Mol. Biol. Evol. 36 (2019) 613–620.
[28] A. Tóth-Petróczy, D.S. Tawfik, Slow protein evolutionary rates are dictated by surface–core association, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 11151–11156.
[29] N. Rajasekaran, S. Suresh, S. Gopi, K. Raman, A.N. Naganathan, A general mechanism for the propagation of mutational effects in proteins, Biochemistry 56 (2017) 294–305.

[30] A. Warshel, et al., Electrostatic basis for enzyme catalysis, Chem. Rev. 106 (2006) 3210–3235.
[31] J.A. Gerlt, P.C. Babbitt, Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies, Annu. Rev. Biochem. 70 (2001) 209–246.
[32] O. Khersonsky, D.S. Tawfik, Enzyme promiscuity: a mechanistic and evolutionary perspective, Annu. Rev. Biochem. 79 (2010) 471–505.
[33] J. Luo, B. van Loo, S.C.L. Kamerlin, Catalytic promiscuity in *Pseudomonas aeruginosa* arylsulfatase as an example of chemistry-driven protein evolution, FEBS Lett. 586 (2012) 1622–1630.
[34] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[35] H.M. Berman, et al., The Protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[36] J.M. Cherry, et al., SGD: *Saccharomyces* Genome Database, Nucleic Acids Res. 26 (1998) 73–79.
[37] A. Ahmed, R.D. Smith, J.J. Clark, J.B. Dunbar, H.A. Carlson, Recent improvements to binding MOAD: a resource for protein–ligand binding affinities and structures, Nucleic Acids Res. 43 (2015) D465–D469.
[38] M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nat. Biotechnol. 35 (2017) 1026–1028.
[39] C. Stark, et al., BioGRID: a general repository for interaction datasets, Nucleic Acids Res. 34 (2006) D535–D539.
[40] A. Chatr-Aryamontri, et al., The BioGRID interaction database: 2017 update, Nucleic Acids Res. 45 (2017) D369–D379.
[41] R.P. Joosten, et al., A series of PDB related databases for everyday needs, Nucleic Acids Res. 39 (2011) D411–D419.
[42] Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, Claus O. Wilke, Maximum allowed solvent accessibilites of residues in proteins, PLoS One 8 (2013), e80635.
[43] I. Wapinski, A. Pfeffer, N. Friedman, A. Regev, Natural history and evolutionary principles of gene duplication in fungi, Nature 449 (2007) 54–61.
[44] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.
[45] L.D. Hurst, The Ka/Ks ratio: diagnosing the form of sequence evolution, Trends Genet. 18 (2002) 486–487.
[46] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Mol. Biol. Evol. 24 (2007) 1586–1591.