Non-catalytic binding sites induce weaker long-range evolutionary rate gradients than catalytic sites in enzymes

Avital Sharir-Ivry[1], Yu Xia[1]

[1]Department of Bioengineering, McGill University, Montreal, QC H3A 0E9, Canada

Abstract

Enzymes exhibit strong long-range evolutionary constraint that extends from their catalytic site and affects even distant sites, yet it remains unclear how the biological function of different sites affects the magnitude of evolutionary rate gradients they induce. Here, we calculate the evolutionary rate (dN/dS) of sites as a function of distance from different binding sites: catalytic sites, non-catalytic ligand binding sites, allosteric binding sites and protein-protein interaction sites. We show that catalytic sites induce significantly stronger evolutionary rate gradient than all other functional sites. Notably, weak long-range evolutionary rate gradients from non-catalytic binding sites on enzymes resemble those from ligand binding sites on non-enzymes. Our results suggest that enzymes and non-enzymes share similar evolutionary constraints only when examined from the perspective of non-catalytic functional sites, and that the unique evolutionary rate gradient from catalytic sites 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 (Bartlett, Porter, Borkakoti, & Thornton, 2002). 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 (Bartlett et al., 2002). Recently it was shown that enzymes exhibit a long-range, nearly-linear conservation gradient from their catalytic site that extends even up to ~30Å in distance (Jack, Meyer, Echave, & Wilke, 2016).

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 (Bloom, Drummond, Arnold, & Wilke, 2006; Conant & Stadler, 2009; Echave, Spielman, & Wilke, 2016; Franzosa & Xia, 2009; Franzosa, Xue, & Xia, 2013; Marcos & Echave, 2015; Oberai, Joh, Pettit, & Bowie, 2009; Overington, Donnelly, Johnson, Sali, & Blundell, 1992; Ramsey, Scherrer, Zhou, & Wilke, 2011; Sharir-Ivry & Xia, 2017). The main local structural determinants known are residue solvent accessibility and residue packing (Franzosa & Xia, 2009; Marcos & Echave, 2015; Sharir-Ivry & Xia, 2017) 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 (Jack et al., 2016), demonstrating that distance and solvent exposure contribute independently to the conservation gradient. Still, given that protein fold usage is very different between enzymes and non-enzymes (Day, Soudackov, Čuma, Schmitt, & Voth, 2002; Hegyi & Gerstein, 1999; Martin et al., 1998), 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 (Sharir-Ivry & Xia, 2018).

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 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 (Nelson & Grishin, 2016a, 2016b). For protein-protein interaction sites however, it was shown that interfacial sites induce significantly weaker conservation gradients than catalytic sites in enzymes (Jack et al., 2016). Here, we present a data-driven proteome-level 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 on enzymes do not induce significant long-range evolutionary rate gradients. Instead, the weak evolutionary rate gradient induced from non-catalytic ligand-binding sites on enzymes resembles that from ligand-binding sites on non-enzymes. Moreover, we show that catalytic sites in enzymes with no known allosteric function still induce a strong long-range evolutionary 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 in catalytic and non-catalytic ligand-binding sites. Taken together, our results suggest that the observed

87 evolutionary rate gradient from catalytic sites in enzymes is primarily driven by the optimization and
88 maintenance of catalytic function rather than ligand-binding or allosteric function.
89
90
91
92
93
94
95
96
97
98
99
100
101

102 # Results

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

105 To examine whether evolutionary rate gradients are also induced from non-catalytic ligand-binding sites,
106 we first identified such functional binding sites within enzymes and non-enzymes in the yeast proteome.
107 As a starting point, we used a dataset of the yeast proteome that contains 1744 structurally annotated
108 yeast proteins (see Methods). We screened them to find yeast proteins for which the structural model
109 contains a biologically significant ligand bound to it based on the BioLip database (J. Yang, Roy, & Zhang,
110 2013), excluding ions. We found 486 ligand-binding sites on 240 enzymes and 462 ligand-binding sites on
111 276 proteins which are not known to be enzymes (non-enzymes). Ligand-binding residues were identified
112 from BioLip, and catalytic site residues in enzymes were identified using the Catalytic Site Atlas (Furnham
113 et al., 2014). The ligand-binding sites on enzymes were then subdivided into 250 ligand-binding sites that
114 include catalytic site residues, and 236 ligand-binding sites that are completely separated from the
115 catalytic site. On average, the evolutionary rates of the structurally modeled non-enzymes and enzymes
116 in our datasets are not significantly different (0.0676±0.0009 and 0.0656 ±0.0007 respectively). For each
117 residue, we calculated the distance to the closest ligand-binding residue. Average evolutionary rate
118 (dN/dS) was then calculated for residues over distance bins to obtain the evolutionary rate gradients.

119 The slope of the linear fit for the evolutionary rate gradient from catalytic binding sites is significantly
120 larger than the slope for the evolutionary rate gradient from non-catalytic ligand-binding sites (t-test,
121 $P<0.001$, Table 1 and Figure 1). Surprisingly, the evolutionary rate gradient induced from non-catalytic
122 ligand-binding sites on enzymes is weak and actually resembles the evolutionary rate gradient induced
123 from ligand-binding sites on non-enzymes (t-test, $P=0.17$). The weak evolutionary rate gradient from non-
124 catalytic ligand-binding sites suggests that ligand-binding functionality of catalytic sites is probably not the
125 main determinant of the strong evolutionary rate gradients induced from them.

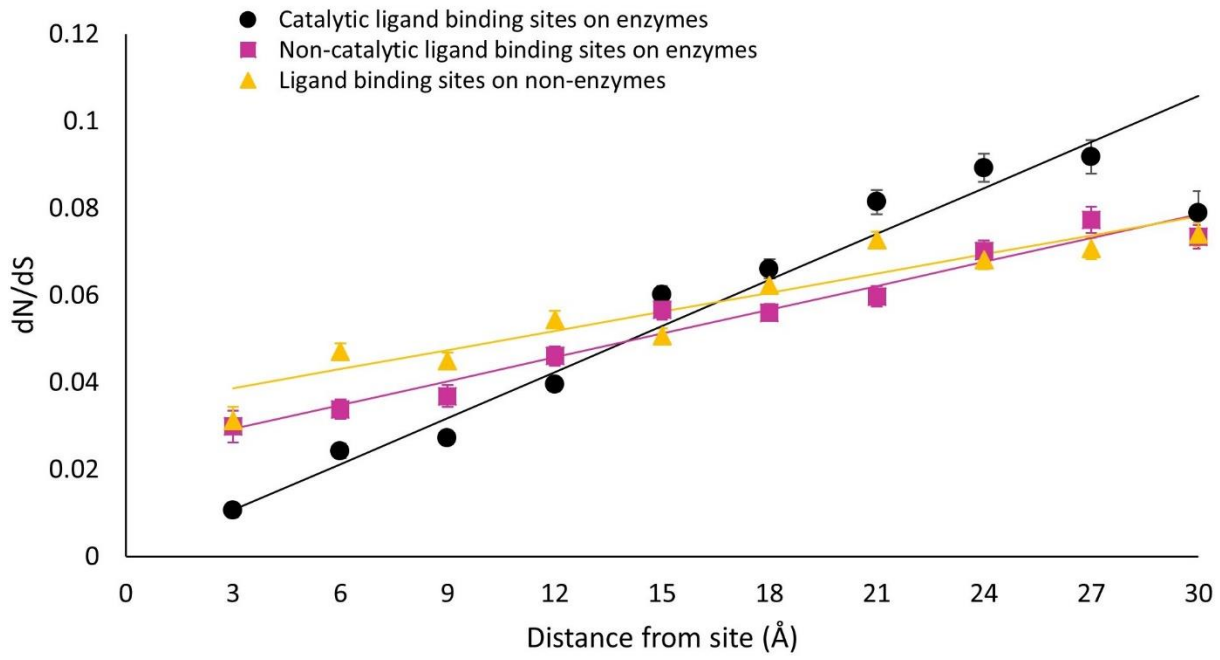


Figure 1. Evolutionary rate (dN/dS) as a function of distance from catalytic and non-catalytic ligand-binding sites in enzymes, as well as from ligand-binding sites in non-enzymatic proteins.

Table 1. 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.

	Type of binding site	**Slope**	**Eevolutionary rate (dN/dS) of binding site residues**
Enzymes	Catalytic ligand-binding sites	0.0035 (±0.0003)	0.011 (±0.002)
Enzymes	Non-catalytic ligand-binding site	0.0018 (±0.0001)	0.030 (±0.004)
Non-enzymes	Non-catalytic ligand-binding site	0.0015 (±0.0002)	0.0312 (±0.003)

- In brackets – standard errors

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

139 Next, we focused on the special case of allosteric binding sites. We collected those yeast proteins that
140 have a structural model which is known to have an allosteric function with its allosteric site residues
141 annotated in the allosteric database (ASD) (Shen et al., 2016). We found 190 allosteric proteins, of which
142 108 are known enzymes with known catalytic site residues. All residues were binned according to their
143 distance from the closest allosteric binding residue as well as from the closest catalytic residue in the case
144 of enzymes. Average evolutionary rate (dN/dS) was calculated for the residues in each distance bin.

145 The slope of the evolutionary rate gradient induced from allosteric binding sites is significantly smaller
146 than that induced from catalytic sites (t-test, $P<0.001$, Figure 2 and Table 2). Allosteric binding sites in
147 enzymes modulate the activity of catalytic sites in that the catalytic site is shifting into its active
148 conformation upon a binding event in the distant allosteric binding site (Liu & Nussinov, 2016; Stadtman,
149 1966). Despite this long-range interaction between catalytic sites and allosteric binding sites, our results
150 suggest that allosteric function is not the main determinant of the long-range evolutionary rate gradient
151 induced from catalytic sites in enzymes.

152

153

154

155

156

157

158

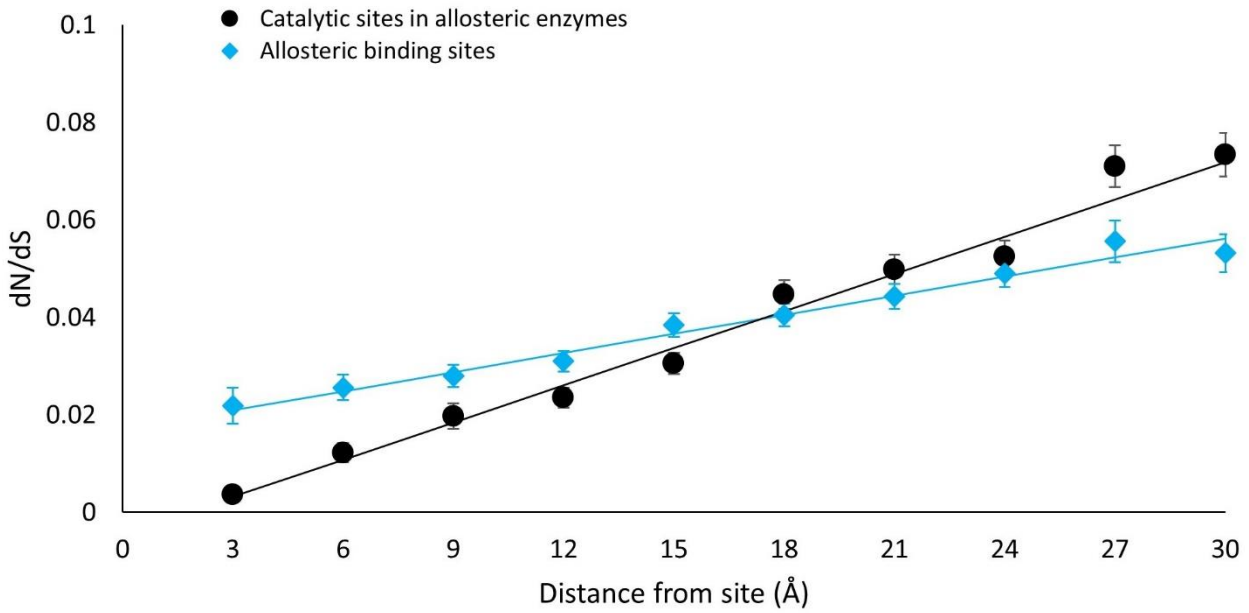


Figure 2. Evolutionary rate (dN/dS) as a function of distance from allosteric binding sites and from catalytic sites in allosteric enzymes.

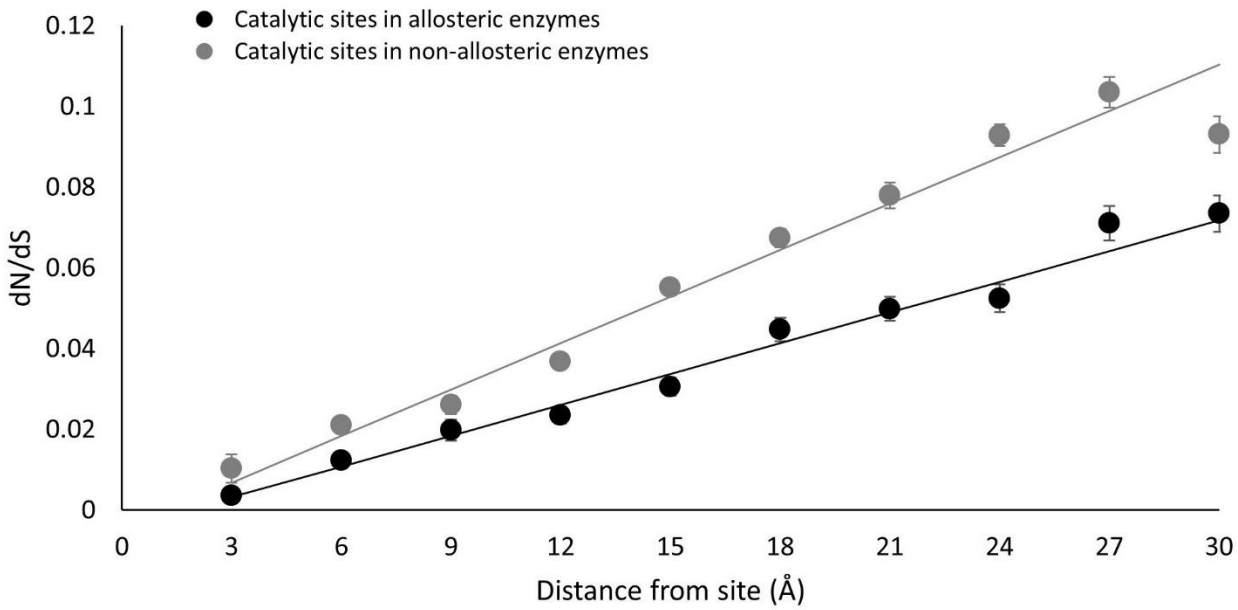
Table 2. Slope of the linear fit of the average evolutionary rate (dN/dS) versus distance from allosteric binding sites and catalytic sites in allosteric enzymes, as well as the average evolutionary rate of binding site residues.

	Binding site type	**Slope**	**Evolutionary rate (dN/dS) of binding site residues**
Allosteric enzymes	Catalytic site	0.0025 (±0.0001)	0.004 (±0.002)
Allosteric enzymes	Allosteric binding sites	0.0013 (<0.0001)	0.022 (±0.004)
Non-allosteric enzymes	Catalytic site	0.0038 (±0.0002)	0.010 (±0.004)

- In brackets – standard errors

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

173 For comparison, we also identified 'non-allosteric enzymes' in our dataset as those proteins with an
174 enzyme structural model with known catalytic residues but with no known allosteric function (219
175 proteins). Notably, similar to allosteric enzymes, catalytic sites in non-allosteric enzymes also exhibit a
176 strong evolutionary rate gradient that extends to distant sites (Figure 3 and Table 2). The existence of
177 evolutionary rate gradient from catalytic sites appears to be independent of the existence of allosteric
178 function in the enzyme. This result further supports the hypothesis that the allosteric function is not the
179 main determinant of evolutionary rate gradients induced from catalytic site in enzymes.



180

181 **Figure 3.** Evolutionary rate (dN/dS) as a function of distance from catalytic sites in allosteric and non-
182 allosteric enzymes.

183

184

185

186 **Protein-protein interaction sites induce weaker long-range evolutionary rate gradients than**
187 **catalytic sites**

188 Using relative conservation scores, it was previously shown that protein-protein interactions sites on
189 enzymes induce only a minor conservation gradient which is significantly weaker than that from catalytic
190 sites (Jack et al., 2016). Here, we studied residue evolutionary rate (dN/dS) as a function of distance from
191 protein-protein interaction sites in the yeast proteome. We screened the structurally-annotated yeast
192 proteome dataset to find proteins that are known to interact with other proteins (see Methods). We found
193 459 interfacial sites for 250 proteins. Interfacial residues were identified as those residues that that have
194 different solvent accessibility when in complex compared to when the interacting partner is deleted.
195 Residues were binned according to their distance from interfacial residues, and average dN/dS was
196 calculated for each bin. The slope of the evolutionary rate gradient induced from interfacial sites is indeed
197 significantly smaller than that induced from catalytic sites (t-test, P<0.001, Figure 4 and Table 3), either
198 on enzymes or on non-enzymes.

199

200 **Table 3**. Slope of the linear fit of the average evolutionary rate (dN/dS) versus distance from protein-
201 protein interaction sites, as well as the average evolutionary rate of interfacial residues.

Protein-protein interaction sites on:	**Slope**	**Evolutionary rate (dN/dS) of binding site residues**
All proteins	0.0004 (±<0.0003)	0.039 (±0.001)
Non-enzymes	0.0008 (±<0.0002)	0.043 (±0.002)
Enzymes	-0.0001 (±<0.0002)	0.031 (±0.002)

202 • In brackets – standard errors

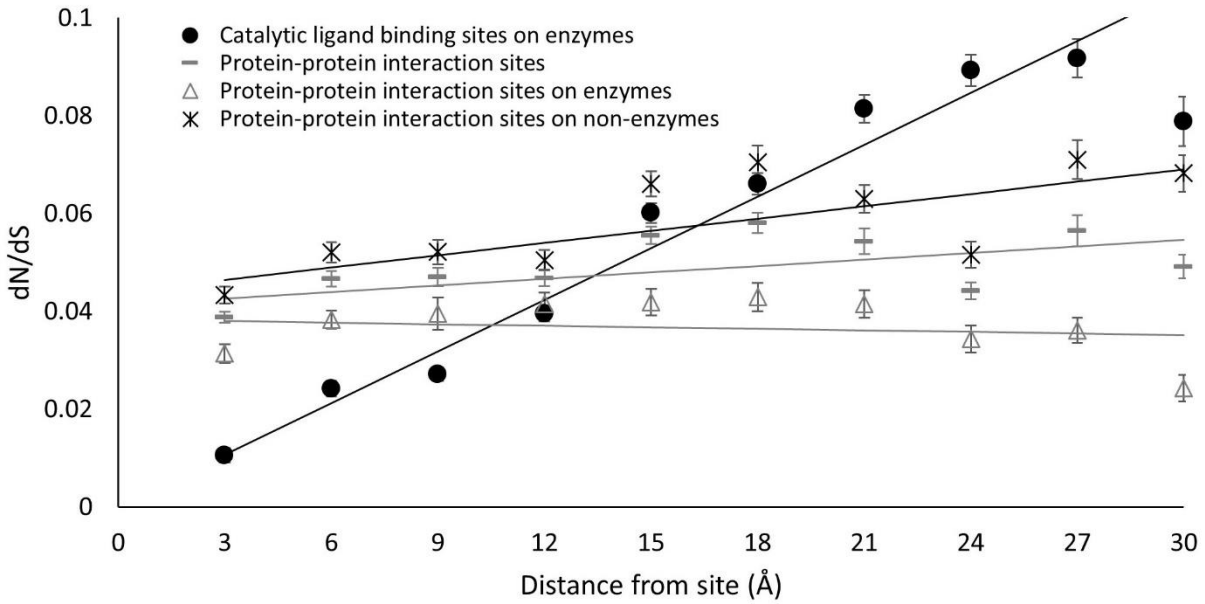
203

204

205

206

207



208

209 **Figure 4.** Evolutionary rate (dN/dS) as a function of distance from catalytic binding sites and from protein-
210 protein interaction sites on enzymes and non-enzymes.

211

212 **Evolutionary rate gradient variations among different functional sites cannot be explained by**
213 **local structural determinants**

214 We examined whether local structural determinants such as solvent exposure gradients are responsible
215 for the observed difference in magnitude of evolutionary rate gradients induced from catalytic and non-
216 catalytic binding sites. We calculated the relative solvent accessibility (RSA) of each residue and then the
217 expected linear trend of dN/dS versus RSA for each protein dataset. We then used the calculated slope
218 and intercept of this dN/dS versus RSA relationship (fig. S1 and table S1 in the Supplementary Material)
219 to plot the expected evolutionary rate at each distance bin according to the average RSA of the residues
220 at that bin. As shown in Figure 5, there is no significant expected increase of evolutionary rate with
221 distance based on solvent exposure alone for any of the datasets of functional binding sites. We therefore
222 conclude that solvent exposure patterns cannot explain the differences in evolutionary rate gradients
223 induced from catalytic and non-catalytic binding sites.

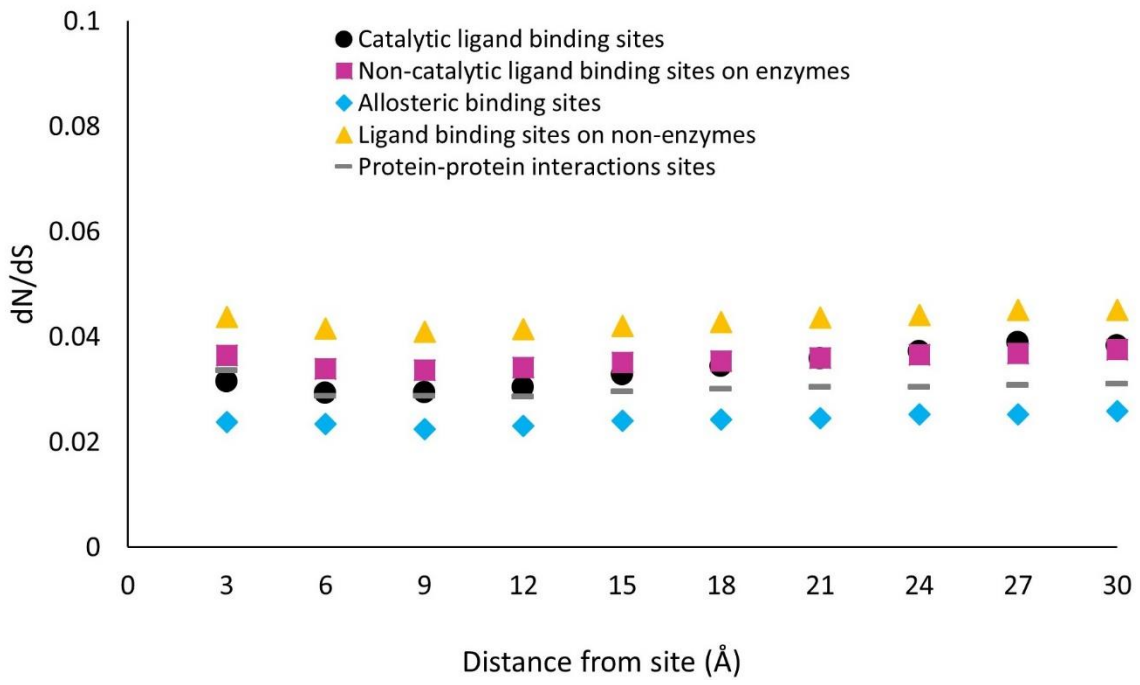


Figure 5. Expected dN/dS according to average relative solvent accessibility (RSA) of residues as a function of distance from binding site.

236 ## Discussion

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

244 Previous studies show that evolutionary rate gradients in enzymes cannot be explained by local structural
245 properties of the protein (Jack et al., 2016; Sharir-Ivry & Xia, 2018), suggesting that the gradient is driven
246 by functional determinants. A recent biophysical model of enzyme evolution that incorporates enzymatic
247 activity constraint along with stability constraints better explains the evolutionary rate gradient from
248 catalytic sites (Echave, 2019). Enzymatic function is very complex, involving both substrate binding and
249 accelerating the production rate of chemical products by decreasing the free energy barrier of the
250 chemical reaction. By showing that non-catalytic ligand-binding sites induce significantly weaker
251 evolutionary rate gradients than catalytic sites, we can rule out several hypotheses on the origin of the
252 long-range evolutionary rate gradient. The first hypothesis that can be ruled out is that the ligand-binding
253 function of the catalytic site is the main determinant of the long-range rate gradient. While ligand binding
254 is part of the function of the catalytic site, it appears improbable that the specific binding of the substrate
255 is the main determinant of the long-range selection pressure in enzymes.

256 The second hypothesis that can be ruled out is that the long-range evolutionary rate gradient from
257 catalytic sites is mainly driven by the allosteric function of the enzyme. Allosteric function exhibit a long-
258 range effect in which binding of an effector molecule at a distant allosteric site shifts the catalytic site into

259 an active conformation (Liu & Nussinov, 2016; Stadtman, 1966). It is reasonable to expect that the need
260 to maintain proper long-range allosteric function can explain the observed long-range evolutionary rate
261 gradient from the catalytic site. On the contrary, we have shown that allosteric binding sites induce a
262 significantly weaker evolutionary rate gradient than catalytic sites. Moreover, a strong evolutionary rate
263 gradient exists even in enzymes which are not known to be allosteric. These results suggest that optimizing
264 the allosteric function of the enzyme is not the main determinant of the observed evolutionary rate
265 gradient.

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

270 We are therefore left with two plausible hypotheses on the origin of the unique evolutionary rate gradient
271 from catalytic sites. The first hypothesis is that strongly conserved sites in general induce stronger long-
272 range evolutionary rate gradient. Here, the evolutionary rate gradient is a result of a percolating effect of
273 penetrating conservation (Nelson & Grishin, 2016b; Rajasekaran, Suresh, Gopi, Raman, & Naganathan,
274 2017; Tóth-Petróczy & Tawfik, 2011). Thus, the higher the selective pressure on a site is, the stronger the
275 percolation of evolutionary rates will be from it. The second hypothesis is that evolutionary rate gradients
276 are driven by the chemistry of the enzyme where the main determinant of the increasing selective
277 pressure is to optimize the catalytic power of the enzyme by specifically stabilizing the reaction transition
278 state (Warshel et al., 2006). Indeed, a chemistry-centered view on the evolution of new catalytic functions
279 in enzymes suggests that new function in an enzyme can evolve from the ability of the enzyme to stabilize
280 the transition state of the new chemical reaction (rather than just binding of the new ligand) (Gerlt &
281 Babbitt, 2001; Khersonsky & Tawfik, 2010; Luo, van Loo, & Kamerlin, 2012). These two hypotheses are
282 both consistent with the observations in this study, and they are not mutually exclusive. Further work is

283 needed to critically examine the validity and applicability of these hypotheses in terms of explaining the
284 origin of the long-range evolutionary rate gradient observed in enzymes.
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301

302 # Methods

303 **Structural annotation of yeast proteins**

304 We based our study on a dataset of structural homologs for the yeast proteome. This dataset was created
305 using gapped BLAST (Altschul et al., 1997) searches between protein subunit sequences with solved
306 structure from the Protein Data Bank (Berman et al., 2000) and 5,861 translated open reading frames
307 (ORFs) of the yeast *Saccharomyces cerevisiae* (Cherry et al., 1998). We kept those ORF–subunit pairs in
308 which both the subunit sequence and the ORF sequence had coverage of ≥50% in the alignment and E-
309 value $<10^{-5}$ and could be paired with their orthologs in three other closely related yeast species *S.*
310 *paradoxus*, *S. mikatae*, and *S. Bayanus*. The model with highest percent of similar residues in the
311 alignment was chosen (where two residues are considered similar if the substitution score is positive).
312 This way, 1744 yeast ORFs were mapped to a homolog in the PDB. The site-specific structural features of
313 the PDB homolog were transferred to the query proteins according to the sequence alignment.

314 *Annotation of ligand-binding and catalytic sites*

315 The structural subunits were screened to find those with an identified biologically-relevant ligand-binding
316 site according to the BioLip protein function database (J. Yang et al., 2013). We excluded binding sites
317 where the ligand is an ion. Yeast ORF was classified as an enzyme if its paired structural subunit has a
318 known EC number. Catalytic sites of protein models were identified using the Catalytic Site Atlas (Furnham
319 et al., 2014). 240 enzymes with 486 ligand-binding sites were identified. We then subdivided the ligand-
320 binding sites on yeast enzymatic proteins into those with an overlap between binding site and catalytic
321 site residues (250 catalytic ligand-binding sites) and those where none of the binding site residues is a
322 catalytic residue (236 non-catalytic ligand-binding sites). 276 proteins with 462 ligand-binding sites are
323 not assigned with an EC number and were therefore termed as 'non-enzymes'. The distance of the Cα
324 atom of each residue in these proteins to the closest Cα of the binding site residues and catalytic site

325 residues was calculated. Residues were binned according these distances up until 30Å into equally spaced
326 bins of 3Å in size. The first bin essentially contained all the ligand-binding residues and only them.

327 *Allosteric proteins*

328 Using the dataset of proteins with known allosteric function from the allosteric database (ASD) (Shen et
329 al., 2016), we screened the yeast ORFs by means of their structural subunit to those that are known to
330 have an allosteric function and their allosteric sites residues are known. 196 yeast ORFs matched this
331 criterion, and out of them we kept 190 proteins for which the structural alignment contained the allosteric
332 binding sites. Within the dataset, 108 proteins are identified as enzymes with an EC number and with
333 known catalytic site from the Catalytic Site Atlas (Furnham et al., 2014). Other proteins in our dataset
334 which have an EC number assigned and their catalytic site residues are identified, but whose structural
335 models are not known to have an allosteric function (219 proteins), were classified as 'non-allosteric'
336 enzymes. The distance of the Cα atom of each residue to the closest Cα atom of allosteric site residue and
337 closest Cα atom of catalytic site residue was calculated. Residues were binned according these distances
338 up until 30Å into equally spaced bins of 3Å in size.

339 *Protein interfaces*

340 We screened the protein complexes from which the best ORF-subunit were derived to those that are in
341 physical contact with another ORF-subunit pair (not necessarily the pair with highest similarity) and were
342 reported as interacting with the other ORF-subunit pair by at least one physical experiment in the BioGRID
343 (Chatr-Aryamontri et al., 2017; Stark et al., 2006). Our dataset contained 250 yeast proteins with 459
344 interfaces. Interfacial residues were identified as the residues which have different solvent accessibility
345 values when in complex compared to when the interaction partner is manually deleted.

346 All of the optimal ORF-subunit pairs and the corresponding subunit binding site residues are listed in
347 supplementary table S2 in the Supplementary Material).

348 **dN/dS calculations**

349 We used previous alignments made between the translated ORFs from *S. cerevisiae* and their orthologs
350 in three other closely related species *S. paradoxus*, *S. mikatae*, and *S. Bayanus* (Franzosa & Xia, 2009). The
351 aligned codons were binned according to their distance from different binding sites as explained above
352 and then concatenated. dN/dS values were calculated over the multiple sequence alignments using the
353 program codeml within the PAML software package (Z. Yang, 2007). As our four yeast species are closely
354 related, we opted to calculate a single dN/dS value for the entire tree (using model=0), which we specified
355 as ((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*, *S. bayanus*). Codon frequencies were assumed equal
356 (CodonFreq=0) and other parameters in codeml were left to their default values. The codon alignments
357 and control file parameters can be found in supplementary files S3 and S4 in the Supplementary Material.

358 **Statistical analysis**

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

364

365

366

367

368

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Bartlett, G. J., Porter, C. T., Borkakoti, N., & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, *324*(1), 105–121. https://doi.org/10.1016/S0022-2836(02)01036-7

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bloom, J. D., Drummond, D. A., Arnold, F. H., & Wilke, C. O. (2006). Structural determinants of the rate of protein evolution in yeast. *Molecular Biology and Evolution*, *23*(9), 1751–1761. https://doi.org/10.1093/molbev/msl040

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., … Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, *45*(D1), D369–D379. https://doi.org/10.1093/nar/gkw1102

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., & Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, *26*(1), 73–79. https://doi.org/10.1093/nar/26.1.73

Conant, G. C., & Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular Biology and Evolution*, *26*(5), 1155–1161. https://doi.org/10.1093/molbev/msp031

Day, T. J. F., Soudackov, A. V., Čuma, M., Schmitt, U. W., & Voth, G. a. (2002). A second generation multistate empirical valence bond model for proton transport in aqueous systems. *The Journal of*

Chemical Physics, *117*(12), 5839–5849. https://doi.org/10.1063/1.1497157

Echave, J. (2019). Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msy244

Echave, J., Spielman, S. J., & Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, *17*(2), 109–121. https://doi.org/10.1038/nrg.2015.18

Franzosa, E. A., & Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular Biology and Evolution*, *26*(10), 2387–2395. https://doi.org/10.1093/molbev/msp146

Franzosa, E. A., Xue, R., & Xia, Y. (2013). Quantitative residue-level structure-evolution relationships in the yeast membrane proteome. *Genome Biology and Evolution*, *5*(4), 734–744. https://doi.org/10.1093/gbe/evt039

Furnham, N., Holliday, G. L., De Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., & Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Research*, *42*(D1), D485–D489. https://doi.org/10.1093/nar/gkt1243

Gerlt, J. A., & Babbitt, P. C. (2001). Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. *Annual Review of Biochemistry*, *70*(1), 209–246. https://doi.org/10.1146/annurev.biochem.70.1.209

Hegyi, H., & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology*, *288*(1), 147–164. https://doi.org/10.1006/JMBI.1999.2661

Jack, B. R., Meyer, A. G., Echave, J., & Wilke, C. O. (2016). Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biology*, *14*(5), e1002452.

https://doi.org/10.1371/journal.pbio.1002452

Khersonsky, O., & Tawfik, D. S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual Review of Biochemistry*, *79*(1), 471–505. https://doi.org/10.1146/annurev-biochem-030409-143718

Liu, J., & Nussinov, R. (2016). Allostery: An overview of its history, concepts, methods, and applications. *PLoS Computational Biology*, *12*(6), e1004966. https://doi.org/10.1371/journal.pcbi.1004966

Luo, J., van Loo, B., & Kamerlin, S. C. L. (2012). Catalytic promiscuity in *Pseudomonas aeruginosa* arylsulfatase as an example of chemistry-driven protein evolution. *FEBS Letters*, *586*(11), 1622–1630. https://doi.org/10.1016/j.febslet.2012.04.012

Marcos, M. L., & Echave, J. (2015). Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ*, *3*, e911. https://doi.org/10.7717/peerj.911

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C., & Thornton, J. M. (1998). Protein folds and functions. *Structure (London, England : 1993)*, *6*(7), 875–884. https://doi.org/10.1016/S0969-2126(98)00089-6

Nelson, E. D., & Grishin, N. V. (2016a). Evolution of off-lattice model proteins under ligand binding constraints. *Physical Review E*, *94*(2), 022410. https://doi.org/10.1103/PhysRevE.94.022410

Nelson, E. D., & Grishin, N. V. (2016b). Long-Range Epistasis Mediated by Structural Change in a Model of Ligand Binding Proteins. *PLOS ONE*, *11*(11), e0166739. https://doi.org/10.1371/journal.pone.0166739

Oberai, A., Joh, N. H., Pettit, F. K., & Bowie, J. U. (2009). Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17747–17750.

https://doi.org/10.1073/pnas.0906390106

Overington, J., Donnelly, D., Johnson, M. S., Sali, A., & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Science : A Publication of the Protein Society*, *1*(2), 216–226. https://doi.org/10.1002/pro.5560010203

Rajasekaran, N., Suresh, S., Gopi, S., Raman, K., & Naganathan, A. N. (2017). A General Mechanism for the Propagation of Mutational Effects in Proteins. *Biochemistry*, *56*(1), 294–305. https://doi.org/10.1021/acs.biochem.6b00798

Ramsey, D. C., Scherrer, M. P., Zhou, T., & Wilke, C. O. (2011). The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, *188*(2), 479–488. https://doi.org/10.1534/genetics.111.128025

Sharir-Ivry, A., & Xia, Y. (2017). The impact of native state switching on protein sequence evolution. *Molecular Biology and Evolution*, *34*(6), 1378–1390. https://doi.org/10.1093/molbev/msx071

Sharir-Ivry, A., & Xia, Y. (2018). Nature of Long-Range Evolutionary Constraint in Enzymes: Insights from Comparison to Pseudoenzymes with Similar Structures. *Molecular Biology and Evolution*, *35*(11), 2597–2606. https://doi.org/10.1093/molbev/msy177

Shen, Q., Wang, G., Li, S., Liu, X., Lu, S., Chen, Z., … Zhang, J. (2016). ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Research*, *44*(D1), D527–D535. https://doi.org/10.1093/nar/gkv902

Stadtman, E. R. (1966). Allosteric regulation of enzyme activity. *Advances in Enzymology and Related Areas of Molecular Biology*, *28*, 41–154.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, *34*(Database issue), D535-9. https://doi.org/10.1093/nar/gkj109

Tóth-Petróczy, A., & Tawfik, D. S. (2011). Slow protein evolutionary rates are dictated by surface-core association. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(27), 11151–11156. https://doi.org/10.1073/pnas.1015994108

Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., & Olsson, M. H. M. (2006). Electrostatic basis for enzyme catalysis. *Chemical Reviews*, *106*(8), 3210–3235. https://doi.org/10.1021/cr0503106

Yang, J., Roy, A., & Zhang, Y. (2013). BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, *41*(D1), D1096-1103. https://doi.org/10.1093/nar/gks966

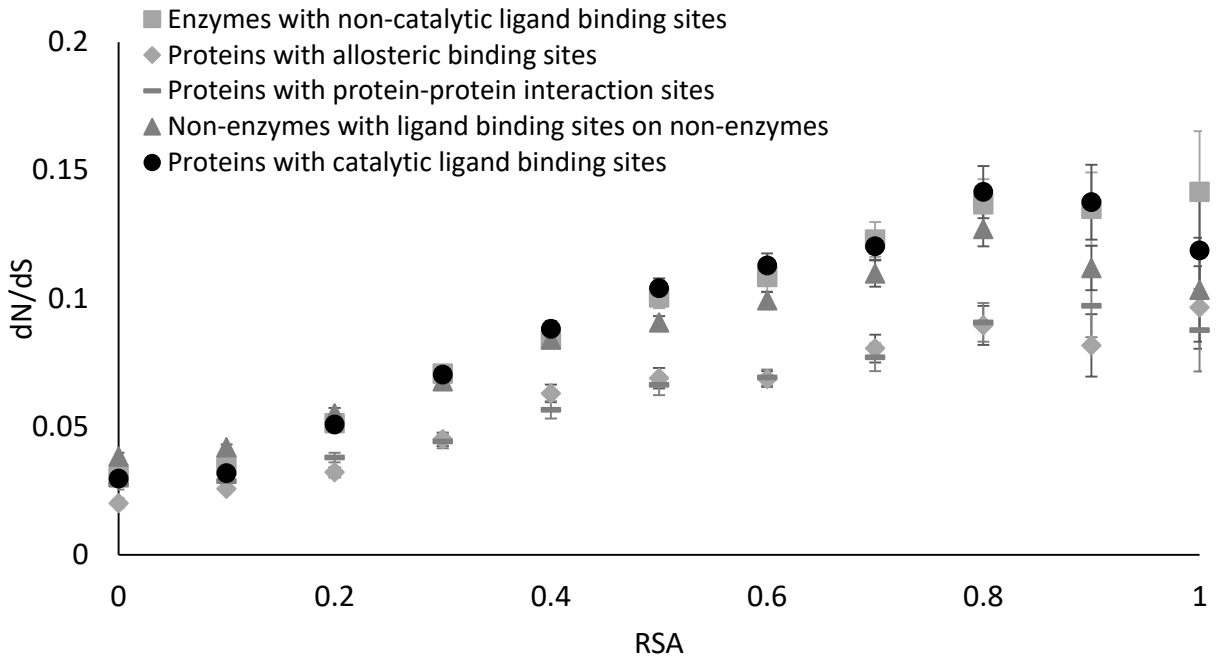
Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

482

483 # Supplementary material

484

485 .



486

487 **Figure S1. Average evolutionary rate (dN/dS) as a function of relative solvent accessibility (RSA).**

488

489 **Table S1. Slope and intercept of the linear fit of the average evolutionary rate (dN/dS) versus**
490 **RSA.**

491

	Slope	Intercept
Proteins with catalytic binding sites	0.145 (±0.011)	0.023 (±0.003)
Enzymes with non-catalytic ligand binding sites	0.136 (±0.008)	0.026 (±0.002)
Non-enzymes with ligand binding sites	0.109 (±0.006)	0.034 (±0.002)
Proteins with allosteric binding sites	0.089 (±0.006)	0.018 (±0.002)
Proteins with protein-protein interaction sites	0.078 (±0.004)	0.023 (±0.001)

492

493

494