The Many Nuanced Evolutionary Consequences of Duplicated Genes

Ashley I. Teufel,*,[1,2] Mackenzie M. Johnson,[1,2] Jon M. Laurent,[†,2,3] Aashiq H. Kachroo,[‡,2,3] Edward M. Marcotte,[2,3] and Claus O. Wilke[1,2]

[1]Department of Integrative Biology, The University of Texas at Austin, Austin, TX

[2]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, TX

[3]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX

[†]Present address: Department of Biochemistry and Molecular Pharmacology, Institute for Systems Genetics, New York University Langone Health, New York, NY

[‡]Present address: The Department of Biology, Centre for Applied Synthetic Biology, Concordia University, Montreal, QC, Canada

***Corresponding author:** E-mail: ateufel@utexas.edu.

Associate editor: Mary O'Connell

Abstract

Gene duplication is seen as a major source of structural and functional divergence in genome evolution. Under the conventional models of sub or neofunctionalization, functional changes arise in one of the duplicates after duplication. However, we suggest here that the presence of a duplicated gene can result in functional changes to its interacting partners. We explore this hypothesis by in silico evolution of a heterodimer when one member of the interacting pair is duplicated. We examine how a range of selection pressures and protein structures leads to differential patterns of evolutionary divergence. We find that a surprising number of distinct evolutionary trajectories can be observed even in a simple three member system. Further, we observe that selection to correct dosage imbalance can affect the evolution of the initial function in several unexpected ways. For example, if a duplicate is under selective pressure to avoid binding its original binding partner, this can lead to changes in the binding interface of a nonduplicated interacting partner to exclude the duplicate. Hence, independent of the fate of the duplicate, its presence can impact how the original function operates. Additionally, we introduce a conceptual framework to describe how interacting partners cope with dosage imbalance after duplication. Contextualizing our results within this framework reveals that the evolutionary path taken by a duplicate's interacting partners is highly stochastic in nature. Consequently, the fate of duplicate genes may not only be controlled by their own ability to accumulate mutations but also by how interacting partners cope with them.

Key words: **gene duplication, protein evolution, protein–protein interactions, dosage imbalance.**

Introduction

Gene duplication is a major driver of functional divergence (Ohno 1970; Lynch and Conery 2000). Duplicate genes provide an additional source of genetic material that is free from the selective constraints experienced by the original gene copy (Ohno 1970). Whereas freedom from these selective constraints often results in the duplicate losing function and being lost from the genome, some duplicate genes acquire functional changes resulting in their preservation. The long term preservation of a duplicated gene is often attributed to one of two mechanisms. Under the first, known as *neofunctionalization*, the duplicate gene acquires a novel and beneficial functional change leading to the gene's retention. Under the second, called *subfunctionalization*, both the original and the duplicate gene copies acquire partial loss-of-function mutations, requiring their mutual retention to perform the original function (Innan and Kondrashov 2010; Dittmar and Liberles 2011). While the preservation of duplicate genes is rare, it is thought to be more likely to occur if a newly duplicated gene remains in the genome for an extended period of time. This extended window of time allows for an increased opportunity for mutational and selective forces to reshape the function of the duplicate gene (Konrad et al. 2011). Further, these mechanisms are not mutually exclusive. The initial retention of a duplicate could be due to subfunctionalization that eventually leads to neofunctionalization (Roth et al. 2007). When multiple genes are duplicated, selective pressure to maintain dosage balance can extend the initial retention period and serve as mechanism for later sub or neofunctionalization (Teufel et al. 2016). Genome architecture may also play a role in the initial retention of a duplicate, as it may be more difficult to expel duplicates depending on their location (Naseeb et al. 2017).

While a number of theories offer explanations of how specific forces foster the functional diversification of duplicate genes, little attention has been given to the secondary effects duplicates have on the evolution of their interacting partners. Considering that the rate of gene duplication can be similar to or exceed the rate of synonymous substitutions

doi:10.1093/molbev/msy210 Advance Access publication November 14, 2018

(Lipinski et al. 2011), the ability of duplicated genes to affect the evolutionary trajectory of interacting partners may be significant. We hypothesize that the traditional view, under which functional changes occur within the duplicate itself, may not capture the full spectrum of evolutionary outcomes for a system of interacting genes after a duplication.

Here, we explore the evolution of a heterodimeric protein complex when one subunit is duplicated. Even in this simple system of three interacting proteins, there are a number of ways that the proteins could be impacted by the presence of a duplicate. To examine the evolutionary consequences of gene duplication, we simulate protein evolution and assess the stability of its subunits, their ability to bind, and the mechanisms driving functional change. We do so under various selection scenarios for several heterodimeric structures. We find that the presence of a duplicate gene can influence the evolution of each member in a protein–protein interaction. Additionally, we observe a surprising amount of variability in how protein-interaction networks cope with dosage imbalance, and we introduce a framework for describing the impact imbalance can have on interacting partners. Our results highlight that the presence of a duplicate gene can affect the evolution of each of its interacting partners and that this impact depends both on protein structure and stochastic events.

Results

We examine the evolution of a simplified protein-interaction network after a partial duplication. While the traditional view of gene duplication predicts that the duplicate gene will escape selective pressure, we suggest that the presence of a duplicate can impose a new set of selective constraints on its interacting partners. Considering a relatively simple system of a heterodimer whose two subunits we refer to as A and B, we impose a duplication event that results in a redundant copy of the B subunit (denoted as B′) (fig. 1). To examine how protein-interaction networks are impacted by duplication, we implement evolutionary simulations under seven different selection schemes (fig. 1). Each of these selection schemes assumes that selection acts on the stability of each of the proteins, and they make varying assumptions about how fitness depends on binding of the subunits. The selection schemes are chosen to reflect different evolutionary scenarios that have been hypothesized to contribute to duplicate gene divergence. Two of these experiments are control simulations that do not include a duplication event (fig. 1, selection schemes 1 and 2). The five remaining simulations include a duplicate of B, referred to as B′, that may or may not confer a fitness benefit depending on whether it binds to A or not (fig. 1, selection schemes 3–7).

To simulate evolution under each of these selection schemes, we use a recently developed simulation platform for protein evolution (Kachroo et al. 2015; Teufel and Wilke 2017). This platform uses a physics-based model with atom-level resolution to evaluate the effect of individual mutations on protein stability and binding. The simulation works with PDB files representing bound complexes of the proteins of interest. Evolution is simulated by sequentially introducing mutations to random positions in any of the proteins within a PDB file and either accepting or rejecting each mutation based on its effect on fitness. The fitness of a given complex is a function of the individual proteins' predicted stability and the strengths of their respective binding interactions, as defined by the different selection schemes (fig. 1). Thermodynamic stability and binding strengths are calculated with the protein-design software Rosetta (Leaver-Fay et al. 2011) and converted into fitness contributions using a soft-threshold model (Chen and Shakhnovich 2009; Wylie and Shakhnovich 2011; Serohijos et al. 2012). Genes are duplicated by duplicating all the atoms in the original PDB file that correspond to the protein to be duplicated and subsequently keeping track of both the original and the duplicated atoms and treating them as separate polypeptides that can independently bind and be mutated.

Our main study system is a heterodimer consisting of a small ubiquitin-like protein complexed with a peptide (PDB ID: 2EKE, Duda et al. 2007). We refer to the peptide as A and the ubiquitin-like protein as B, and hence duplicate the ubiquitin-like protein in the majority of our simulations. However, to examine how our results are affected by the choice of the protein complex, and to investigate the robustness of our findings, we repeat a subset of our experiments (fig. 1, selection schemes 3, 5, and 6) assuming that A is the duplicate protein (denoted as A′). Further, we repeat this subset of experiments with the heterodimer protein structure of antifungal protein KP6 (PDB ID: 4GVB, Allen et al. 2013).

Stability of Protein Interfaces and Structures

To assess how the presence of a duplicate gene influences the evolution of protein interfaces, we measure the stability of the A–B interface over the course of the simulation (fig. 2). The resulting interface stability can be described by three general dynamic patterns. Selection schemes that do not select for binding (selection schemes 2 and 7) result in a destabilization of the protein interface. Most selection schemes that do select for binding (selection schemes 1 and 3–5) consistently maintain interface stability. However, selection scheme 6, which corresponds to deleterious dosage imbalance, causes initial destabilization of the interface followed by a recovery of stability (fig. 2, dark blue line).

Whereas the dynamics of selection schemes 1–5 and 7 are expected, the destabilization of the A–B interface under deleterious dosage imbalance (selection scheme 6) is unexpected and surprising. This observation demonstrates that selection to avoid binding B′ can result in a change to the A–B interface to avoid A–B′ binding. Hence, selection for a functional change of B′ impacts how the A–B interface evolves, forcing a temporary reduction in the stability of the A–B interface. We also examine the stability of the A–B′ interface and find that it displays similar dynamics to the stability of the A–B interface under most of the selection scenarios (supplementary fig. S1, Supplementary Material online). However, under the selection scenario to avoid binding B′, the A–B′ interface

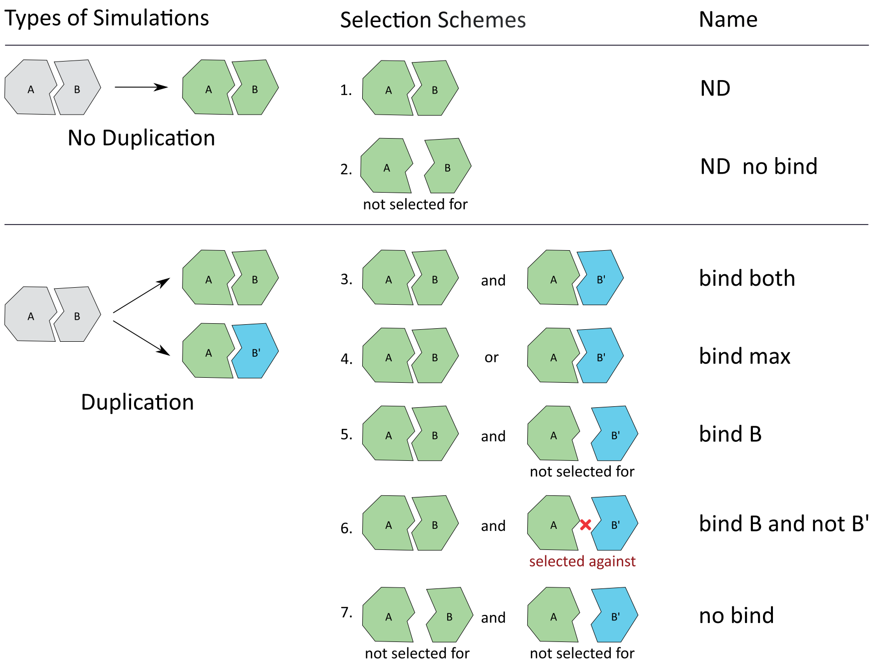

Fig. 1. Selection schemes. We simulate evolution under seven different selection schemes, five of which include the duplication of protein B, denoted as B′. All simulations assume selection for the stability of A, B, and B′ if applicable. Two simulations assume that a duplication event does not occur. Selection scheme 1 (no duplication, ND) describes a scenario without duplication where the stability of the A–B interface is included in the fitness function. Selection scheme 2 (ND no bind) describes a scenario without duplication where the stability of the A–B interface is not included in the fitness function. To examine how the duplication of a subunit affects evolutionary dynamics, we consider five additional selection schemes. Selection scheme 3 (bind both) describes a scenario where both duplicates (B and B′) need to bind A, and the stability of both the A–B and the A–B′ interface is included in the fitness function. This type of selection pressure could occur in a situation where increased dosage of B is beneficial. Selection scheme 4 (bind max) describes a competition scenario where the stability of only one interface, that of the maximum stability of binding for either the A–B or the A–B′ interface, is included in the fitness function. Selection scheme 5 (bind B) describes the process of B′ nonfunctionalization, and the stability of only the A–B interface is included in the fitness function. Selection scheme 6 (bind B and not B′) describes diversifying selection, and the stability of the A–B interface is included in the fitness function whereas the stability of the A–B′ interface is used as a fitness penalty. This sort of selection scenario mimics that of dosage imbalance, where an excessive amount of unbound B′ is harmful. Selection scheme 7 (no bind) describes a control duplication experiment where the stability of neither the A–B nor the A–B′ interface is included in the fitness function.

quickly destabilizes (supplementary fig. S1, Supplementary Material online, dark blue line), as one would expect.

We measure the stability of the duplicate protein, B′, to assess if any of our selection schemes have other effects on the evolving protein (supplementary fig. S2, Supplementary Material online). We find that most of our selection scenarios (selection schemes 3–5, 7) result in a consistent level of stability. However, under deleterious dosage imbalance (selection scheme 6), we find that B′ is destabilized early on in the simulation (supplementary fig. S2, Supplementary Material online, dark blue line). It appears that B′ is unable to fully recover from this initial destabilization during the remainder of the simulation. These findings demonstrate that selection for diversifying functionality, such as the loss of binding ability, may initially cause a destabilization of a protein's structure. After functional changes occur, the stability of the structure may then be refined. We obtain similar results for interface and duplicate stability when we duplicate A instead of B (supplementary figs. S3 and S4, Supplementary Material online) and when we simulate using the antifungal protein (supplementary figs. S5 and S6, Supplementary Material online).

Retention of Ancestral Binding

To further examine how the presence of a duplicate affects the co-evolution of the A–B binding interface, we compare the functionality of the evolved interface to its ancestral function. To compare our evolved and ancestral functions, we assess if our evolved A protein can functionally bind the ancestral B protein. We consider binding to be nonfunctional if the thermodynamic stability of binding does not exceed a protein-specific minimal threshold (see Materials and Methods). The percentage of simulations where A retains its ability to bind the ancestral B is given in supplementary figure S7A, Supplementary Material online for each selection scenario. Generally, it appears that selection for A–B binding, as implemented in selection schemes 3–5, results in a co-evolutionary process that maintains the ability to bind to an ancestral partner, consistent with prior work in a nonduplicated context (Kachroo et al. 2015).

However, when selection acts to avoid binding the duplicate B′ (selection scheme 6), there is a sharp decrease in the ability of A to bind the ancestral partner (supplementary fig. S7A, Supplementary Material online, dark blue line). This indicates that the A subunit changes substantially in order to avoid binding B′, such that only around 20% of the A

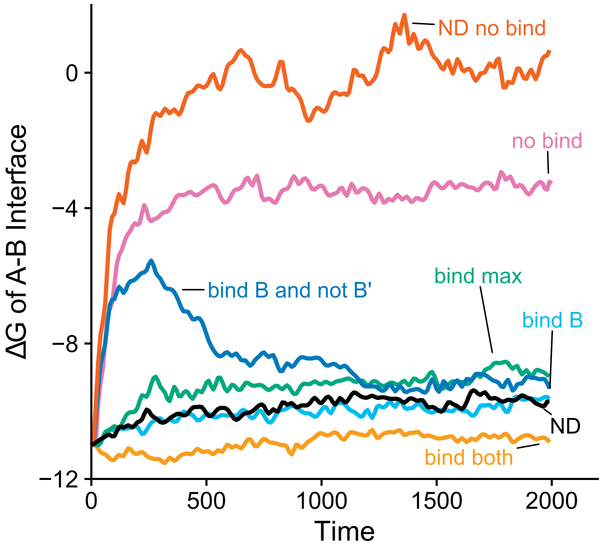

Fig. 2. Stability of the A–B interface versus time, for all seven selection schemes of figure 1. ΔG values are averaged over replicate simulations. Under selection schemes that do not reward binding, the interface stability tends to decay rapidly (more positive ΔG indicates less stable binding). In contrast, selection schemes that do reward binding tend to maintain the interface stability throughout. One exception to this pattern is the bind B and not B′ case, which first shows rapid destabilization of the A–B interface, followed by subsequent regaining of binding stability.

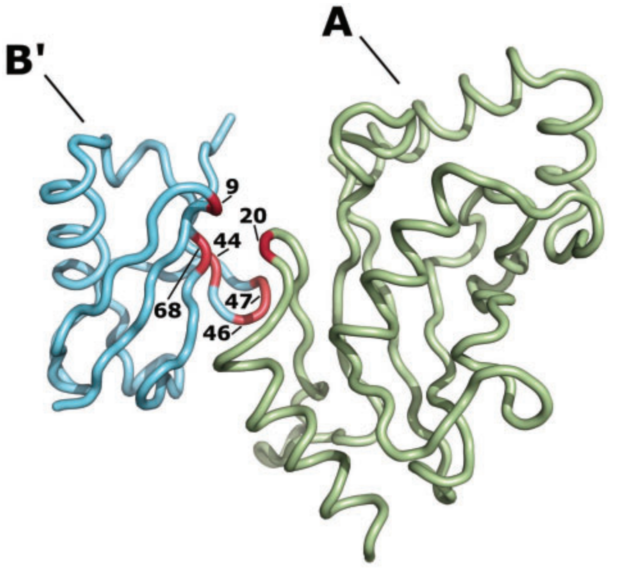

Fig. 3. Significantly differing sites after adaptation to different selection schemes. The highlighted sites differ significantly in their amino acid composition between the bind both and the bind B and not B′ simulations. These sites include site 20 in protein A and sites 9, 44, 46, 47, and 68 in protein B′. No sites were found to significantly differ in their amino acid composition in protein B.

proteins are able to bind the ancestral B by the end of the simulation. This result shows that a functional change can occur in a nonduplicated interacting partner in response to the presence of a duplicate.

We also evaluate how our duplicated proteins, B and B′, functionally diverge based on their ability to bind to the ancestral A protein (supplementary fig. S7*B* and *C*, Supplementary Material online). The ability of the B (supplementary fig. S7*B* and Supplementary Material online) and B′ (supplementary fig. S7*C*, Supplementary Material online) proteins to bind the ancestral A protein decreases more sharply than observed for the A protein. This suggests that the duplicated proteins diverge in function more so than the A protein. Notably, both duplicates similarly retain the ability to bind the ancestral A protein under most selection scenarios, with the exception of selection to avoid binding B′ (supplementary fig. S7*C*, Supplementary Material online, dark blue line).

Pathways of Adaptation to Dosage Imbalance

Considering the destabilization of the A–B interface and of the B′ structure observed when dosage imbalance is selected against (fig. 2, supplementary fig. S2, Supplementary Material online, dark blue lines), as well as the evidence that the A subunit appears to undergo functional change (supplementary fig. S7*A*, Supplementary Material online, dark blue line), we assess how deleterious binding is escaped on a biochemical level. We use a chi-square test to determine which sites have differing amino acid compositions between the bind both (selection scheme 3) and bind B and not B′ (selection scheme 6) experiments at every 10th substitution for the first 250 substitutions.

We find that the amino acid compositions of site 20 in A and 46 in B′ differ significantly ($\alpha = 0.05$, Benjamini and Hochberg 1995 corrected). The number of generations before these sites display significant differences in their amino acid composition also differs. Site 20 in A displays a significant difference by generation 70, whereas site 46 in B′ displays a significant difference by generation 80. This difference in timing suggests that diversifying changes first occur in A and then in the duplicate B′. Further, the loss of A–B′ binding occurs via modifications to site 20 in A and 46 in B′.

To investigate if other downstream effects occur after functional change, we repeat the same chi-square analysis at every 100th substitution. Other sites in protein B′ (9, 44, 47, and 68) also display differing amino acid compositions between the two experiments, though these sites do not differ significantly until after 350 generations. These changes are most likely secondary effects that occur after A has escaped binding to B′. Notably, each of the sites that differ in their amino acid compositions are located in the binding interface (fig. 3).

To examine the properties of sites that significantly differ in their amino acid composition, we assess how residues at these sites contribute to binding using a "stickiness" scale (Levy et al. 2012), which describes the propensity of amino acids to be in protein–protein interfaces. The distributions of amino acid stickiness of site 20 in A and site 46 in B′ is shown for the first 250 substitutions (supplementary fig. S11, Supplementary Material online). Under the bind both selection scheme, the relative stickiness of both of these residues is maintained across time, indicating that maintaining the initial distribution of stickiness is important for the retention of binding. Further, it appears that site 20 initially becomes less sticky to avoid binding B′, suggesting that increasing the propensity for amino acids that are less likely to be involved in protein–protein interactions is crucial to avoid B′ binding. Site 46 in B′ (supplementary fig. S11*B*, Supplementary Material online) appears to increase its propensity for interface residues while under selection to avoid binding B′, suggesting that perturbing the original composition of site stickiness is crucial to avoid B′ binding.

Additionally, we examine the long-term dynamics of sites with differing amino acid compositions at every 100th generation, and observe similar dynamics in terms of the stickiness (fig. 4) and mass of these residues (supplementary fig. S12, Supplementary Material online). Most notably, while site 20 in protein A initially decreases in its stickiness (fig. 4*A*) and mass (supplementary fig. S12, Supplementary Material online), the site later regains some of its stickiness and mass. Other sites in the B′ subunit also display differing distributions in terms of their stickiness (fig. 4). The persistence of differing amino acid compositions at sites in both the nonduplicated interacting partner and the deleterious duplicate throughout the simulation suggests that both the A and the B′ subunit undergo diversifying changes.

To investigate how deleterious binding is lost when A′ is the duplicate subunit, we again measure the stickiness (supplementary fig. S13, Supplementary Material online) and mass (supplementary fig. S14, Supplementary Material online) of residues found to have significantly different distributions of amino acids across time. However, only sites in A′ appear to have continuous differences across time in their amino acids compositions (supplementary fig. S13*A* and *C*, Supplementary Material online), indicating that diversifying changes occur in the A′ subunit. When examining the stickiness (supplementary fig. S15, Supplementary Material online) and mass (supplementary fig. S16, Supplementary Material online) of sites that significantly differ in simulations of the antifungal protein complex we find that sites in both B and B′ significantly differ across time. Interestingly, the presences of the deleterious duplicate B′ appears to influence the mass of residues at a site in B (supplementary fig. S16*B*, Supplementary Material online), despite the fact that these two proteins do not directly interact. We examine why changes towards smaller amino acids at this site occur by substituting a glycine into this position of the protein structure used to initialize these simulations. We find that this substitution increases the A–B binding stability ($\Delta\Delta G = -1.06$); however, it decreases the stability of the B protein ($\Delta\Delta G = 5.48$). Hence, changes to this site appear to be a compensatory mechanism to stabilize the A–B interface during selection to avoid binding B′, but at the cost of structural destabilization. Upon inspection of the stability of B, we do indeed observe a slight destabilization early on in the simulation (supplementary fig. S17, Supplementary Material online, blue line).

We further examine how these functionally diversifying changes occur by tracking the location of the first 250 substitutions relative to the binding interface. We find that selection to avoid binding B′ results in a preference for changes to interface residues in the A and B′ proteins beyond what is observed in the bind both selection scheme (supplementary fig. S8, Supplementary Material online). We obtain similar results when we duplicate A instead of B (supplementary fig. S9, Supplementary Material online) and when we simulate using the antifungal protein structure (supplementary fig. S10, Supplementary Material online). The combination of these results, along with our analysis of critical interface sites, demonstrates that deleterious dosage imbalance results in an increased rate of interface substitutions as well as changes to the amino acid composition of critical interface sites.

As each simulation of deleterious dosage imbalance (selection scheme 6) initialized with a different structure results in diversifying changes in different subunits, we conclude that the mechanisms of functional differentiation are dependent on a protein's structure. In our systems of three interacting proteins we have observed three different outcomes, though five other combinations are theoretically possible as well (fig. 5). In our initial simulation, we have found that functional changes occur in the duplicate itself and in the nonduplicated interacting partner (option 5 in fig. 5). When duplicating the A protein, we find that the changes essential for escaping deleterious binding occur in the duplicate itself (option 1 in fig. 5). Finally, in the simulations initialized with the antifungal structure, we observe that changes occur in the interfaces of both of the duplicated proteins (option 4 in fig. 5). Notably, in each of our simulations the deleterious duplicate displays functional changes.

Beyond Sub and Neofunctionalization: The Impact of Functional Change on Interacting Partners

To quantify the impact that deleterious dosage imbalance has on its interacting partners, we examine how the interacting partners functionally diversify in response to the duplicate's presence. Functional change is assessed by measuring an evolved protein's ability to bind extant and ancestral partners (see Materials and Methods for details). To describe how a deleterious duplicate's interacting partners cope with its presence, we define functionalization pathways in terms of a protein's ability to bind extant and ancestral partners (table 1). When a protein loses the ability to bind its current partner and does not later regain this ability, we refer to this process as *defunctionalization*. When a protein is found to consistently bind its current interacting partner throughout the simulation, but loses the ability to bind its ancestral partner at any point, we refer to this process as *isofunctionalization*. When a protein loses the ability to bind its current partner and then regains the ability later, we refer to this process as *refunctionalization*. After refunctionalization, three distinct fates are possible. If binding to the ancestral partner is permanently lost after refunctionalization, we refer to this process as *hard isofunctionalization*. If binding to the ancestral partner is regained at any point after refunctionalization, we refer to this process as *soft isofunctionalization*. Finally, if binding to the ancestral partner is retained across evolution after

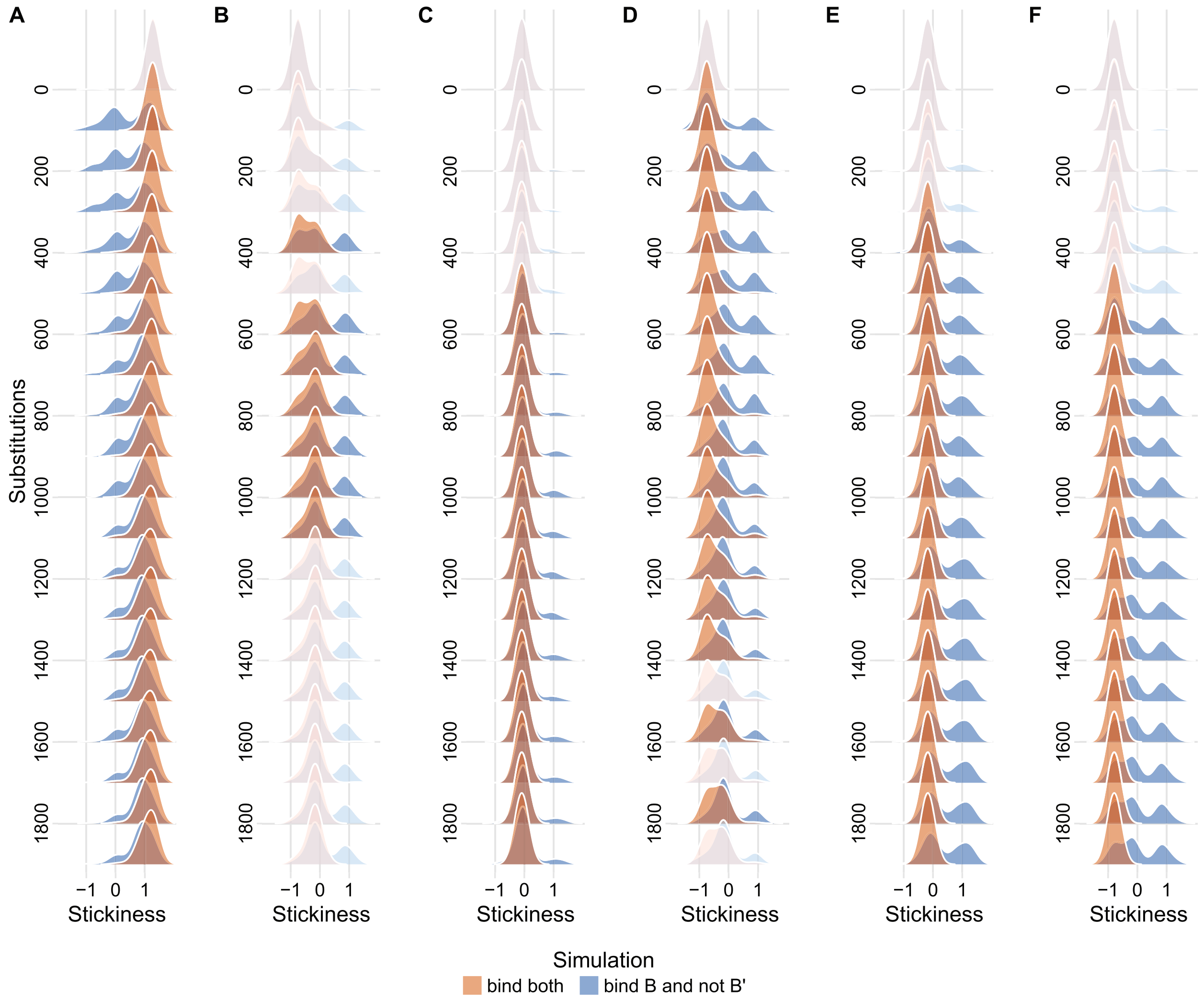

Fig. 4. Distribution of stickiness for sites with differing amino acid compositions, shown at every 100th generation. Grayed-out distributions indicate time points at which sites do not display significant differences. (*A*) Site 20 in protein A. Whereas the stickiness of site 20 in protein A in the bind B and not B′ scenario is initially reduced, some stickiness is regained and by the end of the simulation the distributions of residue stickiness are similar for the two selection schemes. This observation suggests that once dosage imbalance is escaped, the A–B interface is refined and reoptimized. (*B*) Site 9 in protein B′. This site displays a slight shift toward more sticky residues under selection to avoid binding B′, though a larger portion of this distribution still resembles the bind both scenarios. However, distributions differ significantly only from generation 400 to generation 1,100, and this transient behavior may be related to the restabilization of the B′ structure. (*C*) Site 44 in protein B′. For this site, the differing amino acid composition between the two selection schemes does not appear to be reflected in the distribution of amino acid stickiness. (*D*) Site 46 in protein B′. The dynamics at this site suggest that selection to avoid binding B′ initially shifts the distribution towards stickier residues. Interestingly, this shift also occurs under the bind both selection scheme, just later in time. It appears that selection to avoid binding B′ results in an accelerated exploration of sequence space. (*E*) Site 47 in protein B′. (*F*) Site 68 in protein B′. Sites 47 and 68 increase in stickiness under selection to avoid binding B′. However, this effect sets in only around generation 500, indicating that these changes are related to the restabilization of B′.

refunctionalization, we refer to this process as a *functional reacquisition*.

For the case of selection to avoid binding B′ (selective scheme 6), we compare the fractions of simulations resulting in each functionalization pathway from the perspective of the nonduplicated interacting partner (fig. 6*A*) and the nondeleterious duplicate (fig. 6*B*), for simulations initialized with different structures. Notably, each structure results in distinct fractions of functionalization pathways, suggesting that protein structure plays a role in how functional change is achieved. Further, the variation of diversification mechanisms displayed by each different protein structure suggest that stochastic events shape how functional change is achieved in replicate simulations. For each protein, some functional pathways appear to be more common than others. For instance, when simulating with the SUMO ubiquitin-like protein complex (PDB ID: 2EKE, Duda et al. 2007), we observe a larger fraction of functional reacquisition than defunctionalization events. The differences in these fractions implies that some functionalization pathways are more common than others.

For the nonduplicated interacting partner (fig. 6*A*), only a small fraction of our simulations escape dosage imbalance without the temporary loss of the ability to bind the extant partner (isofunctionalization). This indicates that avoiding

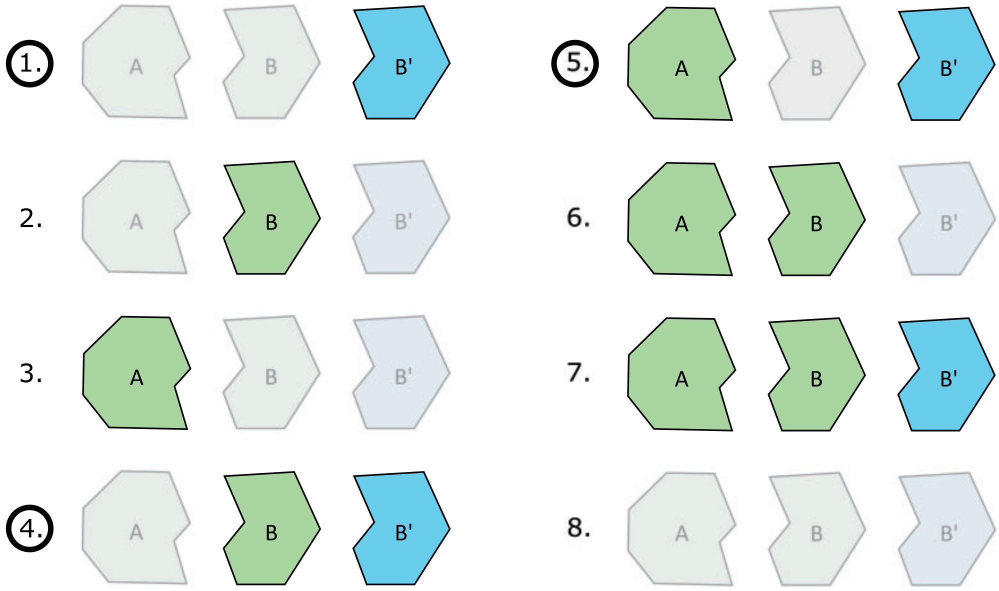

Fig. 5. All possible ways in which the proteins in a three-member network could adapt after duplication. Grayed-out structures indicate no adaptation and structures in full color indicate diversifying change. In options 1 and 2, only one of the duplicated interacting partners acquires diversifying changes. Option 3 describes a situation where only the nonduplicated interacting partner accumulates diversifying changes. Option 4 describes the case where both of the duplicated genes acquire diversifying changes. In options 5 and 6, the nonduplicated partner and one of the duplicate partners accumulate diversifying changes. Finally, in option 7 all proteins and in option 8 none of the proteins acquire diversifying changes. Black circles denote options that we observe in this study.

Table 1. Definitions of Functionalization Pathways.

Functionalization Type	Bind Current Partner	Bind Ancestral Partner
Defunctionalization	Lost and never regained	N/A
Isofunctionalization	Never lost	Lost at any point
Refunctionalization: hard isofunctionalization	Lost and later regained	Permanently lost
Refunctionalization: soft isofunctionalization	Lost and later regained	Intermittently lost
Refunctionalization: reacquisition	Lost and later regained	Lost and later permanently regained

Note.—Each pathway is defined based on the ability a subunit has to bind its current and/or ancestral binding partner.

binding a deleterious duplicate while maintaining binding to another duplicate is possible, though this outcome is relatively rare. Most of the nonduplicated interacting partners undergo refunctionalization, a process that reflects the destabilization of the interface observed early on in each of the simulations (fig. 2, supplementary figs. S3 and S5, Supplementary Material online). After refunctionalization occurs, the nonduplicated partner's interface can be reshaped in several different ways. In the case of the simulations initialized with the ubiquitin-like protein complex (PDB: 2EKE) and the antifungal protein (PDB: 4GVB), this reshaping often results in a protein interface that is still able to perform the ancestral function (reacquisition), though this is observed less often when A is the duplicated subunit (fig. 6A). In each of these simulations, we also observe a notable fraction of cases where binding to the ancestral partner is entirely or intermittently lost after refunctionalization. This result indicates that, in a substantial portion of our simulations, the interface of the nonduplicated interacting partner functionally diversifies from that of the ancestral interface. This finding reinforces the idea that the presence of a deleterious duplicate can result in lasting functional change of the nonduplicated partner.

Examining the functionalization pathways of the nondeleterious duplicate (fig. 6B), it appears that a larger fraction of nondeleterious duplicates undergoes isofunctionalization than does their nonduplicated partner, implying that functional maintenance is more prominent for the nondeleterious duplicate. We also observe that most of the nondeleterious duplicates undergo a phase of refunctionalization; however, this pathway often results in reacquisition of ancestral binding. Notably, in only a few simulations do we see permanent loss of the ability to bind an ancestral interacting partner after refunctionalization (hard isofunctionalization), whereas a few waver on this ability (soft isofunctionalization). A comparison of the nonduplicate and nondeleterious duplicate fractions of functionalization pathways (fig. 6A and B) suggests that the effect of deleterious dosage imbalance affects the nonduplicated interacting partner more substantially than it does the nondeleterious duplicate.

Discussion

We have examined how protein-interaction networks evolve when one member of an interaction network is duplicated. Under a number of different selection schemes, we find that how selection acts on duplicates can impact how their

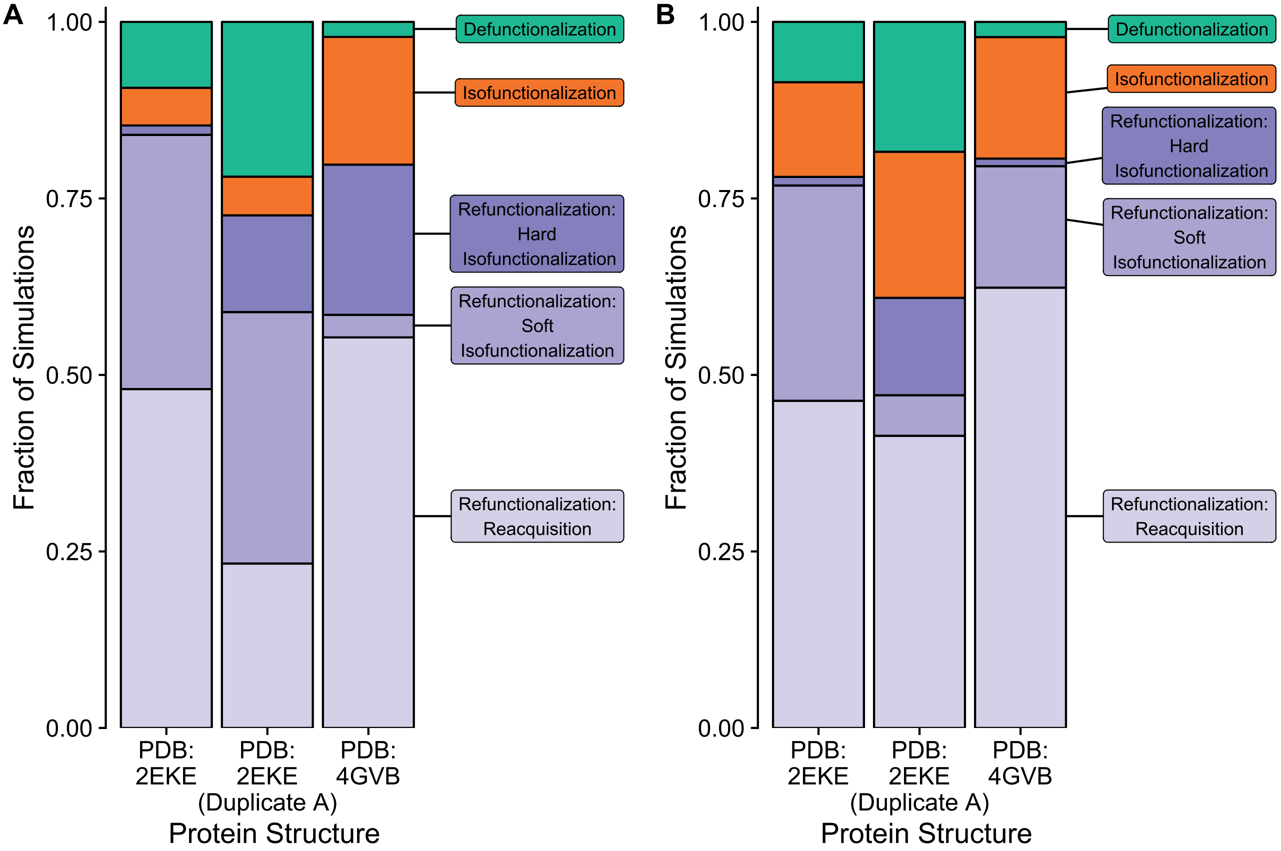

Fig. 6. Comparison of functionalization pathways under dosage imbalance. Each bar represents an experiment initialized with a different starting structure. (*A*) Fraction of functionalization pathways of the nonduplicated interacting partner. (*B*) Fraction of functionalization pathways of the duplicated interacting partner that is under selection to maintain binding. Data are not shown for the deleterious duplicate because it defunctionalizes in all replicates.

interacting partners evolve. Most notably, we find that escaping deleterious dosage imbalance can be achieved through several mechanisms, summarized in figure 5, though five other mechanisms we have not observed are also theoretically possible. Considering that a different combination of changes to interacting partners is observed under each type of simulation instantiated with a different structure, this finding suggests that protein structure plays a substantial role in the location of diversifying changes. Interestingly, in two of our experiments, we find that changes to the original binding interface occur due to the presence of the deleterious duplicate. This functional modification to avoid or cope with dosage imbalance can occur either through changes in the nonduplicated partner or through changes in the nondeleterious duplicate. Further, we find selection to correct dosage imbalance can impact the mass and stickiness of some critical interface resides. The combination of these results indicates that the fate and functionality of a duplicate gene and their interacting partners is in part dictated by protein structure.

We also find that dosage imbalance can destabilize a protein, effectively pushing it into a fitness valley. Once dosage balance is restored, this gene is then free to explore mutational space like any other duplicate, but starting from a different starting position on the fitness landscape. From our simulations, we find that once dosage balance had been restored, the duplicate is able to recover some of its stability by climbing up a different fitness peak. Though this fitness peak is suboptimal, the fact that a novel area of mutational space is explored is notable. The exploration of distance mutational space suggests that dosage balance may not only act as a transition state to subsequent neo or subfunctionalization (Teufel et al. 2016), but actually promote the appearance of these functional changes. In fact, other studies have noted that the evolution of promiscuity, an important step towards functional change, is often due to protein-destabilizing mutations, with additional refinement of novel functionality leading to structural restabilization (Tokuriki and Tawfik 2009; Sikosek and Chan 2014; Dellus-Gur et al. 2015; Petrie et al. 2018).

Further, we find that how interacting partners cope with dosage imbalance is stochastic in nature, and we introduce several terms to describe how dosage imbalance affects the evolution of interacting partners. The stochasticity of our experiments suggests that mechanisms not traditionally considered may describe how duplicated genes diverge and how networks of interacting proteins cope with dosage imbalance. For example, the loss of functionality, such as the ability to bind, is often attributed to deleterious mutations in a duplicate gene. While many duplicate genes may in fact lose functionality this way, our results imply that loss of function can also be achieved by, or result in changes to, the original interaction. Additionally, the terms introduced here, *defunctionalization*, *isofunctionalization*, and *refunctionalization*, to categorize the ways in which a duplicated gene's interacting partners respond or adapt to the duplication, offer a conceptual framework for describing the consequences of gene duplication in a larger interaction network.

We would like to mention that the phrase isofunctionalization is often used in a different context to refer to nonhomologous enzyme isoforms (Omelchenko et al. 2010). Here, we use the term in a congruent fashion to denote the loss of ancestral function while maintaining extant function in the context of protein evolution. The underlying idea behind the concept remains the same; simply, a set of proteins or enzymes perform an equivalent function, they just differ in how they do it. The concept of refunctionalization has also been previously introduced, though it has been used as blanket term for a functional change (Beerhues and Liu 2009; Vecchi 2012). This concept does differ from our use of the term, which we use to describe the temporary loss of function in evolutionary time.

While the prevalence of any of these mechanism across genome evolution is unknown, our findings suggest that the process of duplicate gene divergence may be more complex than previously appreciated. Further, our results demonstrate that the presence of a duplicated gene can shape how duplicated and nonduplicated interacting partners evolve, independent of the fate of the redundant gene. In fact, the duplicated gene may ultimately be lost over the course of evolution, but its presence may have a lasting impact on the evolution of other members in its interaction network.

Granted, these hypotheses of how functional changes occur in duplicates are based solely on simulations. The prevalence of remodeling at the protein–protein interface, to cope with dosage imbalance, in naturally evolving genomes is unknown. Our observations could be a special case associated with smaller proteins with one interacting partner. Additionally, our simulations are based solely on metrics of protein thermodynamics and do not consider the role of expression levels. Changes in the expression level of one or more of the interacting proteins may produce very different evolutionary outcomes (Zhang et al. 2008; Zhang and Shakhnovich 2008). Exploring the evolution of more complex protein networks, where either the entire network or substantial portions of interacting partners are duplicated, in greater biophysical detail would be ideal. However, a single replicate simulation of just three interacting partners can take up to two weeks to run on a 2.20GHz Intel Xeon processing core. In order to run 100 replicates of each simulation, we make use of 100 of these processing cores simultaneously. Simulating a larger system would also require running the simulations for longer, to reach mutation–selection balance. Hence, computational constraints limit the scope of this study. Our observations could also be affected by how the evolutionary process was simulated. The accelerated origin-fixation model used here changes the order in which substitutions are accepted across evolutionary time. However, this reordering was shown to have only a minor influence when compared with evolutionary experiments which did not use this accelerated model (Teufel and Wilke 2017). In fact, this model has been shown to generate realistic variation in alignments of protein sequences (Jiang et al. 2018).

Even though our study is based on a simplified protein-interaction network with only three partners, it still generates novel hypotheses about how duplicated genes diverge and how protein interfaces evolve. Further, it seems unlikely that more complex natural systems would display less variation than observed in our small simulated system. Our study suggests that, even with just three interacting proteins, a wealth of different evolutionary pathways are possible. Additionally, our results demonstrate that duplicated proteins can have long lasting effects on how interacting partners evolve, and these effects are a function of both stochastic events and protein structure. In total, our findings suggest a structurally aware and network-wide perspective is essential to understanding the many fates and consequences of gene duplication.

Materials and Methods

We construct a simulation of protein evolution with the use of an accelerated origin-fixation model (Teufel and Wilke 2017). The simulation is initialized with a small ubiquitin-like protein complexed with a peptide it binds as the resident genotype (PDB: 2EKE, Duda et al. 2007). This protein complex has two subunits, a peptide and the ubiquitin-like protein, which we refer to as A and B, respectively. This naming convention is arbitrary and introduced for the sake of simplicity. In selection schemes that include a duplication event, we copy the PDB information associated with the duplicated subunit. We relabel the copy, append the copy to the PDB file, and renumber the PDB file with the renumber_pdb.py script provided in Rosetta (Leaver-Fay et al. 2011). This results in a PDB file that contains three subunits, which is used to instantiate the simulation. The simulation first parses this file and stores each of the subunits (A, B, B′) as separate entities. This allows for each of the subunits to be considered independently and for the construction of each of the binding combinations (A–B, A–B′) from the subunits. At each step in the simulation, a novel genotype is created by mutating a random single amino acid to a nonresident amino acid. The mutated subunit is then locally repacked 5 Å around the mutation. The stability of each subunit ($\Delta G_{\text{subunit}}$) and the stability of binding ($\Delta G_{\text{binding}}$) are evaluated with Rosetta's all-atom score function (Rohl et al. 2004; Leaver-Fay et al. 2011).

To convert protein stability and binding into fitness, we use a soft-threshold model. This model assumes that the protein's fitness is given by the fraction of proteins in the ground state in thermodynamic equilibrium (Chen and Shakhnovich 2009; Wylie and Shakhnovich 2011; Serohijos et al. 2012). This assumption results in a sigmoidal fitness function (specifically, the Fermi function), where very stable proteins have a fitness of one and fitness declines as stability passes through a threshold value. We calculate the fitness of contributions of stability and binding as

$$f_i = \frac{1}{e^{\beta(\Delta G_i - \Delta G_{\text{thresh}})} + 1} \quad (1)$$

where β is the inverse temperature, ΔG_i is the structural stability or stability of the binding interface for protein i, and ΔG_{thresh} is the threshold at which the protein has lost 50% of its stability. The ΔG_{thresh} parameter controls the extent to which a given stability value implies that a protein or protein binding is stable. This parameter could be tuned for each fitness component to reflect the stringency of selection for that component. For example, setting ΔG_{thresh} of binding to a low value and ΔG_{thresh} of stability to a high value would result in selection for very stable binding interfaces and less stable protein structures. Hence, the contribution that each component has on fitness can be modified through changes in ΔG_{thresh}. To combine the fitness contributions of binding and stability into a single fitness measure, we log-transform fitness as

$$x_i = \log(f_i) = -\log[e^{\beta(\Delta G_i - \Delta G_{\text{thresh}})} + 1], \quad (2)$$

and sum over the fitness contributions of binding and stability (Kachroo et al. 2015). When penalizing binding, we use subtraction rather than addition of the fitness contribution of binding. Using this metric of fitness allow us to express the probability that the mutant genotype will replace the resident genotype as

$$\pi(i \rightarrow j) \approx \begin{cases} 1 & \text{for } x_j > x_i \\ e^{-2N_e(x_i - x_j)} & \text{otherwise}, \end{cases} \quad (3)$$

where N_e is the effective population size (Teufel and Wilke 2017).

At each step in the simulation, the probability of replacement of the resident genotype with a mutant genotype is evaluated with equation (3). Our simulations assume that $\beta = 1$ and $N_e = 1$. These parameters were chosen to decrease the run-time of the simulations, though changes to some of these parameters, such as N_e, can affect the amount of variance observed in $\Delta\Delta G$ (Teufel and Wilke 2017). Each ΔG_{thresh} component is set to half of the initial stability or stability of binding. Setting these threshold values relative to the starting values insures that each component can contribute equivalently to fitness. A burn-in phase is run for 1,000 substitutions to ensure that steady-state sampling behavior is being exhibited.

To quantify the functionality of binding we set a functionality threshold ($\Delta G_{\text{functional}}$), and we consider ΔG of binding in excess of this threshold as nonfunctional. We derive the threshold value by setting the left-hand side of equation (3) equal to 10^{-4} and solving for ΔG_i. Assuming the state of the system is at ΔG_{thresh}, we find

$$\Delta G_{\text{functional}} = \Delta G_i = \frac{\log[2^{1-2/N_e} \times 25^{-1/N_e} - 1]}{\beta} + \Delta G_{\text{thresh}} \quad (4)$$

Seven different selection mechanisms are implemented, and we simulate 100 replicates of each of these conditions for 2,000 substitutions each. These simulations all assume that selection acts on the stability of each subunit and they differ in how selection pressure on binding is imposed. Figure 1 illustrates each of the selection schemes we simulate. Two of these simulations are controls and do not include a duplication event (selection schemes 1 and 2). To examine how the duplication of a subunit affects evolutionary dynamics, we carry out five more sets of simulations (selection schemes 3–7). In these simulations, we assume that the B subunit is duplicated, and we refer to the duplicate protein as B′. A subset of these experiments are also repeated assuming that A is the duplicated protein (so that we have A and A′) rather than B. We run simulations were A′ is the duplicated subunit for selection schemes 3, 5, and 6. We also simulate the evolution of an antifungal protein (PDB: 4GVB, Allen et al. 2013) under selection schemes 3, 5, and 6 to examine if our findings are specific to a particular protein structure. $G_{\text{functional}}$ and the ΔG_{thresh} values are also recalculated for this system.

When analyzing how functional change occurs under each of these different selection scenarios, we compare the location of substitutions, both in terms of which subunits they occur in and their position relative to the binding interface, between the bind both (selection scheme 3) and bind B and not B′ (selection scheme 6) experiments. We choose these two experiments for comparison because they both include five terms in their fitness functions. This choice is made because the inclusion of other scenarios with fewer terms in the fitness function can affect the point of mutation–selection balance. Interface residues are considered to be those within 8 Å of the binding interface. The software package SPIDDER (Porollo and Meller 2006) is used to determine these residues from the structure used to initialize each of our experiments.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Code

The software, results, and analysis tools are available at https://github.com/a-teufel/Protein_Duplication.

Acknowledgments

This study was supported by the National Institutes of Health grant R01 GM088344 to C.O.W. Additional support was provided by the National Science Foundation Cooperative Agreement no. DBI-0939454 (BEACON Center) and grants to E.M.M. from the National Institutes of Health, National Science Foundation, and Welch Foundation (F1515).

References

Allen A, Chatt E, Smith TJ. 2013. The atomic structure of the virally encoded antifungal protein, kp6. *J Mol Biol.* 425(3):609–621.

Beerhues L, Liu B. 2009. Biosynthesis of biphenyls and benzophenones – evolution of benzoic acid-specific type III polyketide synthases in plants. *Phytochemistry* 70(15–16):1719–1727.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B (Methodol).* 57(1):289–300.

Chen P, Shakhnovich EI. 2009. Lethal mutagenesis in viruses and bacteria. *Genetics* 183(2):639–650.

Dellus-Gur E, Elias M, Caselli E, Prati F, Salverda ML, De Visser JAG, Fraser JS, Tawfik DS. 2015. Negative epistasis and evolvability in tem-1 β-lactamasethe thin line between an enzyme's conformational freedom and disorder. *J Mol Biol*. 427(14):2396–2409.

Dittmar K, Liberles D. 2011. Evolution after gene duplication. Hoboken, NJ: John Wiley & Sons.

Duda DM, van Waardenburg RC, Borg LA, McGarity S, Nourse A, Waddell MB, Bjornsti M-A, Schulman BA. 2007. Structure of a sumo-binding-motif mimic bound to smt3p–ubc9p: conservation of a non-covalent ubiquitin-like protein-e2 complex as a platform for selective interactions within a sumo pathway. *J Mol Biol*. 369(3):619–630.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*. 11(2):97–108.

Jiang Q, Teufel AI, Jackson EL, Wilke CO. 2018. Beyond thermodynamic constraints: evolutionary sampling generates realistic protein sequence variation. *Genetics* 208(4):1387–1395.

Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. 2015. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* 348(6237):921–925.

Konrad A, Teufel AI, Grahnen JA, Liberles DA. 2011. Toward a general model for the evolutionary dynamics of gene duplicates. *Genome Biol Evol*. 3:1197–1209.

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, et al. 2011. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*. 487:545.

Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. *Proc Natl Acad Sci U S A*. 109(50):20461–20466.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in caenorhabditis elegans. *Curr Biol*. 21(4):306–310.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.

Naseeb S, Ames RM, Delneri D, Lovell SC. 2017. Rapid functional and evolutionary changes follow gene duplication in yeast. *Proc R Soc B*. 284(1861):20171393.

Ohno S. 1970. Evolution by gene duplication. Berlin Heidelberg: Springer-Verlag.

Omelchenko MV, Galperin MY, Wolf YI, Koonin EV. 2010. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol Direct*. 5(1):31.

Petrie KL, Palmer ND, Johnson DT, Medina SJ, Yan SJ, Li V, Burmeister AR, Meyer JR. 2018. Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science* 359(6383):1542–1545.

Porollo A, Meller J. 2006. Prediction-based fingerprints of protein–protein interactions. *Proteins: Struct Funct Bioinform*. 66(3):630–645.

Rohl CA, Strauss CE, Misura KM, Baker D. 2004. Protein structure prediction using rosetta. *Methods Enzymol*. 383:66–93.

Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D, Liberles DA. 2007. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B: Mol Dev Evol*. 308(1):58–73.

Serohijos AW, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep*. 2(2):249–256.

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface*. 11(100):20140419.

Teufel AI, Liu L, Liberles DA. 2016. Models for gene duplication when dosage balance works as a transition state to subsequent neo-or subfunctionalization. *BMC Evol Biol*. 16(1):45.

Teufel AI, Wilke CO. 2017. Accelerated simulation of evolutionary trajectories in origin-fixation models. *J R Soc Interface*. 14(127):20160906.

Tokuriki N, Tawfik DS. 2009. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*. 19(5):596–604.

Vecchi D. 2012. Taking biology seriously: neo-darwinism and its many challenges. In: Evolution 2.0. Berlin Heidelberg: Springer. p. 225–247.

Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc Natl Acad Sci U S A*. 108(24):9916–9921.

Zhang J, Maslov S, Shakhnovich EI. 2008. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Mol Syst Biol*. 4(1):210.

Zhang J, Shakhnovich EI. 2008. Sensitivity-dependent model of protein–protein interaction networks. *Phys Biol*. 5(3):036011.