Constraints on the expansion of paralogous protein families

Citation: McClune, Conor J. and Michael T. Laub. "Constraints on the expansion of paralogous protein families." Current Biology 30, 10 (May 2020): R460-R464. © 2020 Elsevier Inc

Published Version: http://dx.doi.org/10.1016/j.cub.2020.02.075

Publisher: Elsevier BV

Permanent Link: https://hdl.handle.net/1721.1/131144

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Constraints on the expansion of paralogous protein families

Conor J. McClune[1,2], Michael T. Laub[1,3,*]

[1] Dept. of Biology; Massachusetts Institute of Technology; Cambridge, MA 02139

[2] Dept. of Chemical Engineering & ChEM-H Institute; Stanford University, Stanford, CA 94305

[3] Howard Hughes Medical Institute; Massachusetts Institute of Technology; Cambridge, MA 02139

* corresponding author: laub@mit.edu

Abstract

Duplication and divergence is a major mechanism by which new proteins and functions emerge in biology. Consequently, most organisms, in all domains of life, have genomes that encode large paralogous families of proteins. For recently duplicated pathways to acquire different, independent functions, the two paralogs must acquire mutations that effectively insulate them from one another. For instance, paralogous signaling proteins must acquire mutations that endow them with different interaction specificities such that they can participate in different signaling pathways without disruptive cross-talk. Although duplicated genes undoubtedly shape each other's evolution as they diverge and attain new functions, it is less clear how other paralogs impact or constrain gene duplication. Does the establishment of a new pathway by duplication and divergence require the system-wide optimization of all paralogs? The answer has profound implications for molecular evolution and our ability to engineer biological systems. Here, we discuss models, experiments, and approaches for tackling this question, and for understanding how new proteins and pathways are born.

New proteins frequently arise from the duplication and divergence of existing genes. This powerful and common evolutionary process has produced genomes in all domains of life that encode many paralogous proteins, which provide cells with diverse functions. In particular, signaling proteins are often members of very large paralogous protein families, or are comprised of paralogous domains, endowing cells with sophisticated regulatory capabilities. The use of duplication and divergence to create new signaling pathways is likely advantageous; it is presumably faster and easier to modify an existing type of signaling pathway than to invent a completely new form of signaling. However, the expansion of a cell's signaling repertoire through duplication requires that cells maintain the interaction specificity of proteins that are, necessarily, closely related at the sequence and structural levels. If different signaling pathways are to faithfully convey different signals within a cell, but involve paralogous proteins, evolution must select against any unwanted pathway cross-talk. The two paralogs produced by a given gene duplication event undoubtedly constrain each other's evolutionary trajectories; they must acquire mutations to diverge from one another as they attain new, independent functions. But do other paralogs in a cell influence or constrain recent duplicates? Is the establishment of interaction specificity a system-wide, or global, optimization problem or only a local one involving the recent duplicates? Does it depend on the size of a paralogous family? Are there limits on how large a paralogous protein family can become before cross-talk prevents further evolutionary expansion? In this Essay, we discuss studies on a range of paralogous protein families that are beginning to shed light on these questions.

Sequence spaces

To begin, we must introduce the concept of sequence space. Evolutionary biologists have long used the idea of a sequence space to conceptualize the set of possible sequences, either of amino acids or nucleic acids, and the trajectories that evolution can take as those sequences evolve. Consider a protein, P, of length 100 amino acids. There is a vast sequence space of size 20^{100}, containing all possible amino-acid sequences of length 100 of which protein P is just one. As protein P accumulates substitutions during evolution, it moves through this sequence space. As first proposed by John Maynard Smith in 1970, one can consider 'trajectories' through sequence space [1], *i.e.* how can proteins accumulate substitutions that change its sequence, but without going through a non-functional intermediate (Fig. 1A). Sequence space-based models have since been routinely used to map the relationship between a set of sequences and a functional output, such as protein folding [2, 3], RNA structure [4], and enzyme efficiency [5].

The analogy in Fig. 1A is, however, somewhat misleading. As Maynard Smith noted, there are individual substitutions that will retain the same function of a protein and are accessible through a series of neutral or nearly neutral mutations. Thus, a better depiction of proteins in sequence space is as a cloud or set of sequences that harbor a given function (although there could be a quantitative measure of function, for simplicity we classify sequences here as functional or not). In the context of signaling proteins, we can think of a cloud or niche in sequence space as representing the set of sequences that enable a signaling protein to bind with a given interaction partner (Fig. 1B). For a family of paralogous signaling proteins, each member has a niche in sequence space, and if they are insulated (do not engage in cross-talk) these niches should be mostly non-overlapping. The process of duplication and divergence can then be reframed as a question of how two niches in sequence space arise post-duplication from one niche (Fig. 1C). And the question of how other paralogs constrain a duplication event can be posed as a question of whether the niches of those other paralogs are close enough in sequence space to constrain the movement of recent duplicates.

As a concrete example, consider a protein kinase P that phosphorylates a substrate S. If a kinase-substrate pair duplicate, the two paralogous systems can diverge to become insulated, forming pairs P1-S1 and P2-S2. Although some paralogous proteins become insulated through temporal or spatial compartmentalization, many proteins operate in the same cellular spaces at the same time such that specificity must be enforced at the level of molecular recognition [6]. From the perspective of the kinases, the two niches derived from a duplication of P can diverge in one of two general ways. The first, called subfunctionalization (Fig. 1C, top), involves P1 and P2 becoming more specific, such that each only phosphorylates a subset of the substrates that the parent protein P previously recognized. P1 and P2 effectively subdivide the niche of P. The second, called neofunctionalization (Fig. 1C, bottom), involves P1 or P2 (or both) diverging in specificity such that one or both niches is displaced in sequence space relative to that of the ancestral P. For these neofunctionalization events, what, if anything, constrains movement through sequence space? In particular, do the niches of other paralogs constrain movement? Such system-wide constraints may influence the relative evolutionary likelihoods of subfunctionalization and neofunctionalization strategies. Thus, the critical question becomes: how crowded is sequence space?

The crowdedness of sequence space reflects the difficulty, or the evolutionary challenge, of generating a new paralog with sufficiently different interaction specificity to operate independently

of all paralogs. There are two extremes (Fig. 1D). Paralogs may be sparsely distributed, meaning that sequence space is large enough and sufficiently unoccupied by extant complexes that new proteins can arise, and change over time, with a low probability of interference or cross-talk arising. In this extreme, recently duplicated paralogs move in sequence space with little constraint until insulated from each other. In the opposite extreme, sequence space is crowded, implying that the establishment of new pathways is difficult because multiple paralogs must be changed, or be repositioned in sequence space, to avoid cross-talk. In such a regime, the evolutionary expansion of signaling pathways requires some degree of global or family-wide optimization to ensure insulation between all pathways.

How crowded is sequence space?

It is difficult to surmise which of these extremes is more likely *a priori*. On one hand, the size of a protein sequence space is extremely large. If the interaction specificity of a protein is determined by N interfacial residues, the full set of interface possibilities is $\sim 20^N$, meaning that sequence space should be sparsely occupied, even for a protein family containing hundreds of paralogs. However, paralogs may not be uniformly distributed across sequence space, leading to local crowding. Further, the size of N can be quite limited, often less than 10 and as small as 3. Many paralogous families of protein complexes, including some kinase-substrate pairs [7], bacterial toxin-antitoxin systems [8], Ig domains [9], and protocadherins [10], rely primarily on such limited sets of coevolving interface residues for interaction specificity. Moreover, protein interfaces can be highly degenerate, such that many substitutions in those interface residues have no significant impact on binding or protein function [11]. Additionally, as a protein family grows, the number of non-cognate interactions that must be avoided grows quadratically.

One of the first studies to quantitatively assess the sequence space of macromolecular interactions examined transcription factor-DNA binding specificity [12]. This bioinformatic study found that the number of transcription factors in a given family is typically bounded and that the maximum size of a family correlated with the number of DNA bases recognized by members of that family. This correlation led to the provocative suggestion, supported by considerations of coding theory, that the size of a transcription factor family is limited by a need to avoid cross-talk between members of that family [12, 13]. In other words, the sequence space relevant to the DNA-binding properties of a family of transcription factors may become crowded enough to prevent further

expansion through duplication. However, regulatory complexity and additional specificity can arise through other mechanisms, such as through homodimeric or heterodimeric interactions [14], combinatorial interactions with other transcription factors, kinetic mechanisms, or spatial sequestration [15]. Importantly, transcription factor-DNA binding specificity may be fundamentally different than that of protein-protein interaction specificity due to the limited conformational diversity of DNA relative to proteins.

Some have sought to determine whether cross-talk represents an important evolutionary constraint on protein-protein interactions by measuring the extent of cross-talk exhibited by paralogous proteins from different genomes. If the sequence space relevant to the interaction specificity of a protein is crowded and the insulation of paralogs has been achieved through genome- or system-wide optimization, then a paralog from a different organism (not subjected to the same optimization) may show significant cross-talk in a new genomic context. A study of SH3 domains, abundant peptide-binding modules in eukaryotic signaling pathways, argued that the limited cross-talk between paralogs in *Saccharomyces cerevisiae* was the product of a global optimization of binding [16]. This study found that Pbs2, the peptide substrate for the *S. cerevisiae* SH3 domain Sho1, did not bind any of the other 27 SH3 domains in the *S. cerevisiae* genome, but did strongly bind several SH3 domains from metazoans. From such observations, the authors concluded that selection against cross-talk optimized the full set of *S. cerevisiae* SH3 domains to avoid each other's substrates, but did not endow SH3-peptides pairs with sufficient specificity to be insulated in the context of a different genome. This model suggests that sequence space is, at least for SH3 paralogs in *S. cerevisiae*, densely occupied, or crowded. However, this interpretation of the SH3 data is heavily dependent on the evolutionary relationships between the SH3 domains examined. Indeed, most of the metazoan SH3 domains that bound Pbs2 are, in fact, closely-related homologs of *S. cerevisiae* Sho1, the cognate partner of Pbs2 [16]. This sequence homology makes the observed cross-compatibility less surprising and undermines the notion of a system-wide optimization. Instead, the results support a model in which the ancestral Sho1 SH3 domain expanded through a series of duplication events in the metazoan lineage, producing a set of SH3 domains that are relatively close in sequence space such that each retains some ability to interact with Pbs2 (the local optimization scenario in Fig. 1E). This study highlights the challenge of avoiding the bias of evolutionary history when using extant proteins to investigate sequence space dynamics.

Another approach to assessing sequence space crowdedness is to use high-throughput assays to make systematic measurements of specificity. In one such effort, protein microarrays were used to quantify the binding specificity of 217 genome-encoded peptides to 157 mouse PDZ domains [17], another signaling domain with dozens of paralogs per genome. The resulting data were fed into linear regression models of specificity and principle component analysis was used to reduce the dimensionality of each PDZ's specificity to three parameters. The predictive power of this simple linear model underscores the independent additivity of different specificity residues in PDZ-binding peptides. As with SH3 domains, the authors concluded that specificity resulted from a system-wide optimization of these paralogous domains because mouse PDZ domains spread out uniformly across their three principle component axes. However, this spread is likely the outcome of dimensional reduction. Principle component analysis, by definition, selects eigenvectors that maximally capture the variance of the data. This study further exemplifies the limits of using extant proteins to assess the evolutionary potential and consequences of adding new, orthogonal proteins to a genome. Without sampling outside extant sequences, it is difficult to fully assess the capacity and occupancy of a sequence space governing paralogous protein interactions.

New approaches to probing sequence space

Two new approaches are now making it feasible to more directly probe sequence space crowding by asking how easily one can either select for or rationally engineer proteins with specificity orthogonal to extant paralogs. If sequence space is crowded, it should be difficult to introduce new proteins that are insulated from extant paralogs, *i.e.* do not cross-talk to the partner proteins of those paralogs. However, if extant paralogs are more sparsely distributed in sequence space, it should be easy to produce new, insulated proteins.

The first new approach is primarily computational. For instance, Rosetta-based tools are allowing the targeted re-engineering of the specificities of known protein complexes [18-20]. As these tools mature and predictive power grows, *in silico* methods offer a means of both designing protein interactions and assessing their global insulation in a cellular context. One recent effort used Rosetta to design orthogonal heterodimeric coiled-coils [18]. This computational study produced 16 different heterodimers (32 chains) that were shown *in vitro* to be mostly orthogonal, *i.e.* selectively bound their cognate partner as designed, with 6 heterodimers confirmed as mutually orthogonal *in vivo*. The ability to design these non-natural heterodimers hints that nature may not

have fully sampled or explored the full set of coiled-coil interfaces, despite the prevalence of coiled-coils in cells. These successful design efforts portend more comprehensive sampling and assessments *in silico* of the sequence spaces relevant to some protein interactions.

The second approach for probing the density of paralogs in sequence space involves screening large libraries of protein pairs for desired properties [21-23]. For a given protein-protein interaction, one can generate libraries in which the key interface residues in each protein are randomized. With an appropriate selection, functional variant pairs can be identified and compared to extant sequences or even assayed for cross-talk with extant proteins. One recent study measured the compatibility of the 1007 possible single mutants in Fos with 1000 possible single mutants of Jun [23]. The authors identified pairs in which each single mutant alone disrupts the Fos-Jun interaction, but the pair are functional together, *i.e.* the two detrimental mutations compensate for each other. Such variant pairs are effectively orthogonal to the wild-type Fos-Jun pair and suggest, like the computational studies, that novel coiled-coils can be built. However, whether the mutant pairs are fully insulated from the wild-type Fos-Jun pair or other homologous complexes *in vivo* was not reported.

The study of Fos and Jun also only assessed single mutants in the interacting proteins, a significant restriction on the sampling of sequence space. We recently examined a different protein-protein interaction, but used large, diverse libraries in which 5-6 residues on each protein were fully randomized [21]. Our study focused on the two-component signaling pathway PhoQ-PhoP from *E. coli*. Most bacteria, including *E. coli*, harbor dozens of paralogous two-component signaling pathways, each comprised of a sensor histidine kinase and its cognate response regulator, with individual pathways typically well insulated from each other [24]. From large libraries, we selected for PhoQ-PhoP variant pairs that could productively interact, such that the PhoQ kinase variant phosphorylated the variant PhoP response regulator. The PhoQ and PhoP variants selected were then tested for their insulation from the parental PhoQ-PhoP system, as well as the other 27 canonical two-component signaling pathways present in the native host, *E. coli*. Approximately 40% of the selected variant pairs were insulated from the parental proteins and none of these showed any evidence, *in vitro* or *in vivo*, for cross-talk to the 27 other two-component pathways. Furthermore, we identified many sets of up to six PhoP*-PhoQ* variant pairs that were all mutually orthogonal. These observations support the notion of a sparsely occupied sequence space such that many new pathways can be introduced without the global optimization of paralogs.

There is, however, some evidence that two-component paralogs can interfere with each other during the process of duplication and divergence. An earlier study provided evidence that the NtrB-NtrC pathway in alpha-proteobacteria was likely duplicated to produce the NtrX-NtrY pathway [25]. As the duplicated pathways diverged from each other, cross-talk to another paralogous pathway, PhoR-PhoB, arose. This likely selected for changes, conserved throughout the alpha-proteobacteria, in the PhoR-PhoB system that reduce this cross-talk. Taken together, these studies suggest that recently duplicated pathways constrain each other and occasionally are constrained by another pathway. However, this example of paralog interference likely reflects a clustering, or crowding, of a subset of paralogs within a limited region of sequence space. In other words, changes in interaction specificity may stem from or reflect local constraints in sequence space, but not require a global or system-wide optimization of specificities. In line with this idea, we propose that the positions of extant paralogs in sequence space reflect their evolutionary history (Fig. 1E). Duplication events lead to local clusters of paralogs that have moved apart in sequence space as they were selected to be insulated, yielding the observed differences in their specificity residues. However, once insulated, further movement in sequence space would presumably arise only from neutral changes in the specificity residues of individual systems.

Implications of a sparsely occupied sequence space

For two-component signaling proteins, the relatively sparse overall occupancy of sequence space implies that there is ample room for new pathways to be introduced, either through additional duplication-divergence events or via lateral transfer. Like many bacterial genes, those encoding two-component signaling pathways often move horizontally and thus could, in principle, cross-talk with the other systems present in a recipient cell. However, the overall sparsity of sequence space would make cross-talk unlikely. For new pathways that arise through duplication, they may be locally constrained by a limited number of closely related paralogs, but given the overall sparsity, the establishment of new pathways through duplication should still be relatively easy from the standpoint of avoiding cross-talk. However, important questions remain. For instance, what is the nature of the mutations that must arise for two pathways to become insulated? As noted, insulation can arise through either neofunctionalization of one, or both, systems, or through subfunctionalization of a promiscuous ancestor. Elucidating which mechanisms produced the insulated paralogs observed in extant genomes will require careful phylogenetic studies and the characterization of ancestral protein states.

How two-component signaling proteins occupy the sequence space relevant to their interaction specificity may differ from that of other paralogous interactions. The high dimensionality of a protein-protein interaction, likely dependent on the number of residues making contacts, is critical to the total number of different specificities a common protein backbone can encode. This may differ for interactions of different sizes and structures. For instance, interactions between proteins and short linear peptides, as occurs for PDZ domains and other common domains of eukaryotic signaling pathways, may occupy smaller sequence spaces as unstructured peptides can encode fewer specificities than the three dimensional surface of a globular domain. If this difference is substantial, it may be more difficult to evolve or build new protein-peptide interactions orthogonal to each other and endogenous paralogs.

As methods improve for probing the sequence spaces relevant to protein interactions of all kinds, what do we stand to learn and what are the implications? For one, a deeper understanding of the structure and organization of sequence spaces will provide new insights into the evolution of new biological functions. Improved methods for systematically mapping or computationally predicting the binding partners of extant proteins promises a richer look at how paralogs fit into sequence space [26-28]. Are some regions of sequence space more densely occupied than others? How does, or how will, this constrain evolution and the creation of new pathways? Are there paralogous protein families other than DNA-binding proteins that have exhausted sequence space, or hit an upper limit on the number of paralogs that can exist without substantial cross-talk? How robust is paralog insulation? What are the tradeoffs between robustness and evolutionary flexibility?

Better methods for probing sequence spaces also promises to impact efforts to engineer biological systems. Improvements in experimental and computational studies of sequence space will permit the *de novo* design of large sets of protein interactions orthogonal to each other and to endogenous homologs. As synthetic biology transitions from a focus on transcriptional circuits to faster protein-based circuits [29, 30], knowing or predicting the expected cross-talk between endogenous proteins and newly introduced pathways will be essential. To date, most engineered circuits are constructed from naturally evolved protein domains moved to new genomic contexts. Analogous to horizontal gene transfer, functionality in a new cell host requires there be little interference between the introduced components and any endogenous homologs. In general, it should be possible and even relatively easy to prevent cross-talk. Sequence spaces are large and duplication seems to have driven the expansion of paralogs in relatively limited regions of these spaces. This

overall sparsity of extant paralogs leaves ample room for the introduction of new pathways, in both rational engineering efforts and during evolution.

Acknowledgements

We thank I. Frumkin and D. Ghose for their comments on the essay. M.T.L. is an Investigator of the Howard Hughes Medical Institute. This work was also supported by a grant from the Office of Naval Research (N000141310074) to M.T.L.

References

1. Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature *225*, 563-564.

2. Cordes, M.H., Davidson, A.R., and Sauer, R.T. (1996). Sequence space, folding and protein design. Curr Opin Struct Biol *6*, 3-10.

3. Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? Biophys J *73*, 2393-2403.

4. Schuster, P., Fontana, W., Stadler, P.F., and Hofacker, I.L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci *255*, 279-284.

5. Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol *10*, 866-876.

6. Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

7. Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

8. Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell *163*, 594-606.

9. Wojtowicz, W.M., Wu, W., Andre, I., Qian, B., Baker, D., and Zipursky, S.L. (2007). A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains. Cell *130*, 1134-1145.

10. Nicoludis, J.M., Green, A.G., Walujkar, S., May, E.J., Sotomayor, M., Marks, D.S., and Gaudet, R. (2019). Interaction specificity of clustered protocadherins inferred from sequence covariation and structural analysis. Proc Natl Acad Sci U S A *116*, 17825-17830.

11. Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science *347*, 673-677.

12. Itzkovitz, S., Tlusty, T., and Alon, U. (2006). Coding limits on the number of transcription factors. BMC Genomics *7*, 239.

13. Friedlander, T., Prizak, R., Guet, C.C., Barton, N.H., and Tkacik, G. (2016). Intrinsic limits to gene regulation by global crosstalk. Nat Commun *7*, 12307.

14. Baker, C.R., Hanson-Smith, V., and Johnson, A.D. (2013). Following gene duplication, paralog interference constrains transcriptional circuit evolution. Science *342*, 104-108.

15. Bentovim, L., Harden, T.T., and DePace, A.H. (2017). Transcriptional precision and accuracy in development: from measurements to models and mechanisms. Development *144*, 3855-3866.

16. Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676-680.

17. Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science *317*, 364-369.

18. Chen, Z., Boyken, S.E., Jia, M., Busch, F., Flores-Solis, D., Bick, M.J., Lu, P., VanAernum, Z.L., Sahasrabuddhe, A., Langan, R.A., et al. (2019). Programmable design of orthogonal protein heterodimers. Nature *565*, 106-111.

19. Dang, L.T., Miao, Y., Ha, A., Yuki, K., Park, K., Janda, C.Y., Jude, K.M., Mohan, K., Ha, N., Vallon, M., et al. (2019). Receptor subtype discrimination using extensive shape complementary designed interfaces. Nat Struct Mol Biol *26*, 407-414.

20. Netzer, R., Listov, D., Lipsh, R., Dym, O., Albeck, S., Knop, O., Kleanthous, C., and Fleishman, S.J. (2018). Ultrahigh specificity in a network of computationally designed protein-interaction pairs. Nat Commun *9*, 5286.

21. McClune, C.J., Alvarez-Buylla, A., Voigt, C.A., and Laub, M.T. (2019). Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. Nature *574*, 702-706.

22. Sockolosky, J.T., Trotta, E., Parisi, G., Picton, L., Su, L.L., Le, A.C., Chhabra, A., Silveria, S.L., George, B.M., King, I.C., et al. (2018). Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. Science *359*, 1037-1042.

23. Diss, G., and Lehner, B. (2018). The genetic landscape of a physical interaction. Elife *7*.

24. Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems. Annu Rev Microbiol *66*, 325-347.

25. Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell *150*, 222-232.

26. Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M.M., and Correia, B.E. (2019). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods.

27. Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat Methods *15*, 816-822.

28. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat Methods *16*, 1315-1322.

29. Gao, X.J., Chong, L.S., Kim, M.S., and Elowitz, M.B. (2018). Programmable protein circuits in living cells. Science *361*, 1252-1258.

30. Langan, R.A., Boyken, S.E., Ng, A.H., Samson, J.A., Dods, G., Westbrook, A.M., Nguyen, T.H., Lajoie, M.J., Chen, Z., Berger, S., et al. (2019). De novo design of bioactive protein switches. Nature *572*, 205-210.

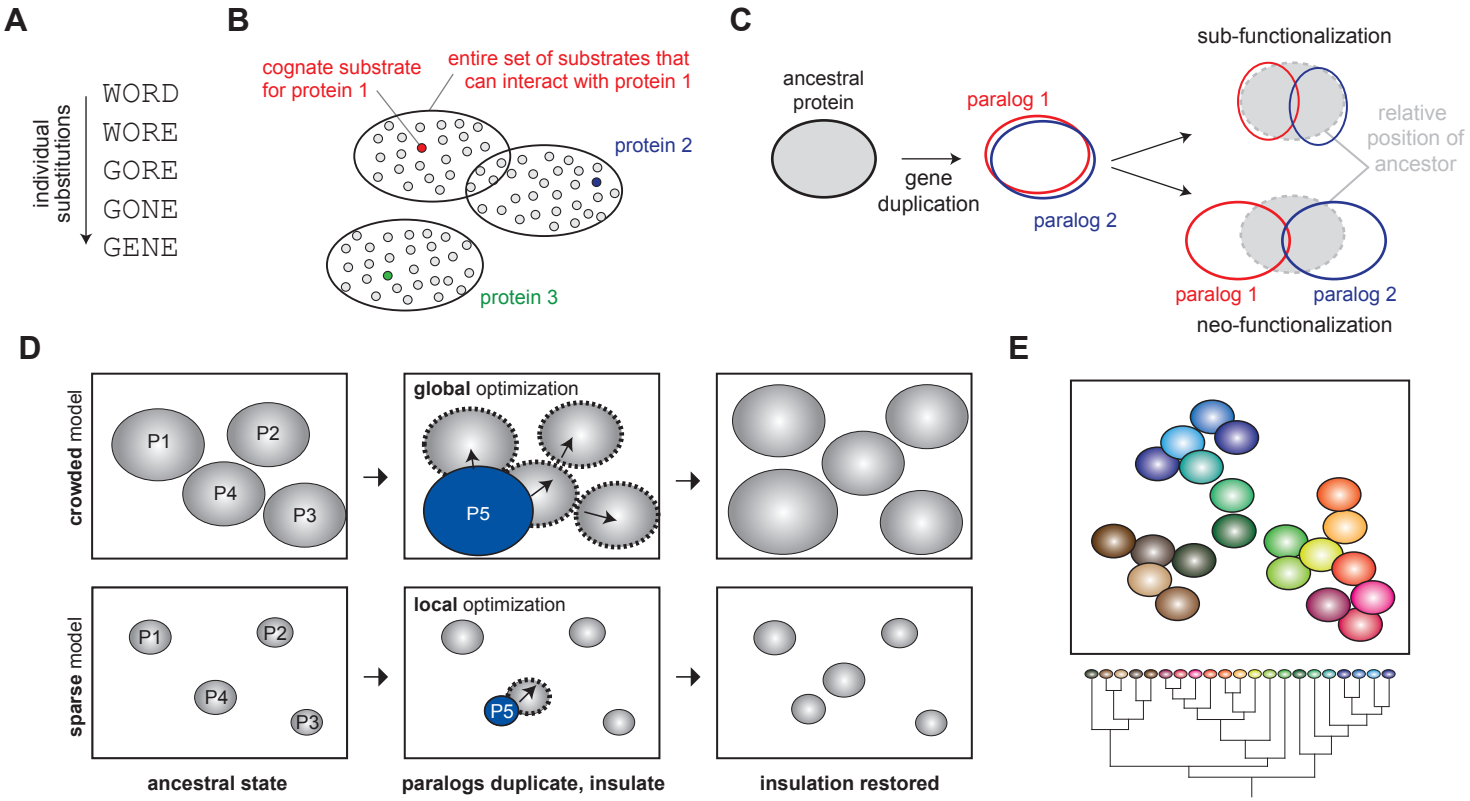

Figure 1. (A) The example given by J. M. Smith in 1970 to illustrate the idea of a sequence space and mutational 'trajectories'. See text for details. (B) Schematic of sequence space relevant to three paralogous proteins. Each oval represents the entire set of partner proteins that can interact with a given paralog (protein 1, 2, or 3), with the cognate partner colored. These ovals, or niches, in sequence space may overlap (as with proteins 1 and 2) or not (as with proteins 1 and 3), but are positioned such that the cognate partner of each is contained only within the niche of one paralog, reflecting the insulation of the three systems. (C) Post-duplication, two paralogs can become insulated through sub-functionalization (top) or neo-functionalization (bottom). (D) Two extreme models for the distribution of paralogs in sequence space. Each oval represents the set of substrates that a given paralogous kinase can interact with. These ovals, or niches, may be crowded (top row) or sparsely distributed (bottom row), requiring global (top) or only local (bottom) optimization of specificity when a new paralog is introduced. (E) If sequence space is sparse, the expansion of a paralogous protein family by duplication results in a series of local optimizations to drive insulation post-duplication. Thus, the organization of paralogs reflects their evolutionary history.