Linguistic networks uncover grammatical constraints of protein sentences comprised of domain-based words.

Adrian A. Shimpi[1,2], Kristen M. Naegle[1,2,*]

1 Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, 22903
2 Department of Genome Sciences, University of Virginia, Charlottesville, VA, 22903

*** Correspondence: kmn4mj@virginia.edu**

Abstract

Evolution has developed a set of principles that determine feasible domain combinations analogous to grammar within natural languages. Treating domains as words and proteins as sentences, made up of words, we apply a linguistic approach to represent the human proteome as an n-gram network. Combining this with network theory and application, we explore the functional language and rules of the human proteome. Additionally, we explored subnetwork languages by focusing on reversible post-translational modifications (PTMs) systems that follow a reader-writer-eraser paradigm. We find that PTM systems appear to sample grammar rules near the onset of the system expansion, but then convergently evolve towards similar grammar rules, which stabilize during the post-metazoan switch. For example, reader and writer domains are typically tightly connected through shared n-grams, but eraser domains are almost always loosely or completely disconnected from readers and writers. Additionally, after grammar fixation, domains with verb-like properties, such as writers and erasers, never appear – consistent with the idea of natural grammar that leads to clarity and limits futile enzymatic cycles. Then, given how some cancer fusion genes represent the possibility for the emergence of novel language, we investigate how cancer fusion genes alter the human proteome n-gram network. We find most cancer fusion genes follow existing grammar rules. Collectively, these results suggest that n-gram based analysis of proteomes is a complement to the more direct protein-protein interaction networks. N-grams can capture abstract functional connections in a more fully described manner, limited only by the definition of domains within the proteome and not by the combinatorial challenge of capturing all protein interaction connections.

Introduction

Domains are modular units of structure and function that enable protein complex formation or translate biochemical information between signaling effectors[1]. About half of the human proteome is comprised of multidomain proteins, where domains help define overall protein functionality through their independent contributions. The combination of domains within a protein, or domain architecture, arises primarily through the shuffling of preexisting domains, rather than the emergence of new domains[2–4]. Evolutionary jumps, encoded within the domain architectures of protein families, can predict the acquisition of new protein functions[3, 5]. Changes in domain architectures most commonly occur from the gain or loss of domains from the terminal ends of preexisting proteins[6]. Interestingly, only a small fraction of possible domain combinations are observed, which cannot be described by the random shuffling of domains during genetic recombination events[3, 4, 7], but can be attributed to few domains having multiple domain partners[8, 9]. These observations have motivated the representation of proteins as vectors comprised of their domains for evaluating the evolution of protein families and the complete proteome[3, 5, 10]. Further, these representations have been used to predict the subcellular localization and gene ontology terms for individual proteins[11, 12] demonstrating the breadth of information encoded within overall protein domain architectures.

Insights from broad surveys of both protein domain architectures and amino acid sequences suggest that proteins operate with sets of rules akin to grammar within natural languages[10, 13, 14]. However, the development of protein language models often focus on the amino acid sequence to predict the structure, function, and evolution of unresolved proteins or guide novel protein design[13, 15–17]. Despite the breadth of information encoded within domain architectures few linguistic approaches have used domains as the fundamental unit of the protein language. Applications that utilize domains as "words" within a protein have shown linguistics can recover protein functionality and evolution across the tree of life independent of known protein interaction networks or signaling pathways[5, 10]. However, these past methods rarely consider the word ordering and word repeats, which likely reflect the evolutionary pressures related to the non-random shuffling of domains and the observation of the predominant modification of domain architectures at the terminal ends[3, 4, 6, 7]. The sequential order can have important consequences on protein functions like modifying the avidity or specificity of catalytic domains like the tyrosine kinase domain[18, 19]. Like natural languages, there are likely constraints on the relative location and combination of certain word types that make up human proteins. For example, catalytic domains might represent verbs and the binding domains they appear with are adverbs, or possibly as the nouns or subject of the sentence that help define the action the verb will have. Certain types of adverbs or the combination of too many verbs without modifying nouns may be prohibited from a language that ultimately needs to be interpreted by the biochemical networks of the cell, while also maintaining sentence clarity (akin to minimizing energy usage).

N-gram analysis is a linguistic approach which maintains both the composition and sequential order of words within natural languages and can be adapted to protein domain architectures. By treating individual domains as words and the complete domain architecture as a sentence, n-grams with n domains can be extracted from either single or multiple domain architectures. A 2-gram model has shown that pairwise domain combinations can recapitulate the evolution of proteome complexity[10]. The smaller fraction of multidomain proteins in prokaryotes than eukaryotes[20] has limited n-gram analysis to mostly 2-gram models for comparative genomic studies. However, certain 2-3 domain combinations, or supra-domains, are rearranged together as a unit across protein families[21]. If an n-gram model does not extract n-grams longer than these supra-domains, the diversity of feasible domain combinations in a proteome may not be sufficiently recovered.

Here, we evaluate n-gram models of various lengths to describe the human proteome. We integrate these models with network analysis techniques to identify protein domains and multidomain n-grams that act as hubs to connect obligate domain families, enabling a simplification of the overall network. An advantage of representing the human proteome as a language-based network is that all proteins have defined domains, unlike networks that rely on protein-protein interactions that have yet to be fully annotated[22, 23]. However, the connections within the network reflect more about the related functional connections amongst key words in the human proteome than it does direct protein-protein interactions. Here, we develop and test n-gram networks at the level of the entire proteome, characterizing the entropic information needed to recover most of the proteome. Next, we applied n-gram analysis to measure the emergent properties of the specific words and languages within reader-write-eraser systems that coordinate post-translational modifications (PTMs). Surprisingly, we find that despite vastly different evolutionary timescales, most reversible PTM-systems convergently evolved to have tight connections between the readers and writers and very loosely, if at all connected, erasers. Looking across evolution, it appears that how PTM systems sampled different word ordering gives a possible measure of the relative time since the appearance of the system and the time at which the language becomes fixed in terms of its grammar rules and composition. We then ask if n-gram based network analysis can lead to novel insights to how cancer gene fusions alter, or not, the functional connections within the proteome. Interestingly, we find that predominantly somatic fusion genes arise with the same characteristics that appear to guide the overall evolution of most multidomain protein architectures[3, 4]. Hence, we find that n-gram linguistic analysis of proteomes is highly useful at making entire proteome-level insights about the functional connections between proteins and systems within cells, along with being useful for subnetwork analysis, such as evaluated in the PTM system.

Results

Generating domain n-gram networks to describe the domain architecture landscape

In order to describe the entire human (or other species) proteome as an n-gram network, we relied on the InterPro database, using our recently developed python-based package CoDIAC[24] to annotate the unique proteome domain and domain architecture (sequential ordering of domains in a protein). The InterPro database[25] consolidates annotations from databases such as Pfam[26], SMART[27], CATH-Gene3D[28] and others into one central database to define domain families. Importantly, InterPro is closely integrated with the UniProt database that serves as a central, comprehensive repository of protein sequences and functions. The result of our access is a controlled vocabulary based on the InterPro IDs of the architecture and the order of domains in proteins within the proteome of interest. For example, the SRC and ABL kinase families have the domain architecture SH3-SH2-Kinase. We then define all possible n-grams within each protein, maintaining the continuous sequence of domains from the N- to C-terminal. For the SRC and ABL family kinases, this includes three 1-grams (SH3, SH2, Kinase), two 2-grams (SH3|SH2, SH2|Kinase), and one 3-gram (SH3|SH2|Kinase).

We applied this n-gram extraction to the entire human proteome, and then broadly surveyed domain architectures to identify highly prevalent domains and n-grams that span multiple protein families. About 95% of the proteome have domain architectures containing up to 10 domains (Fig. 1B, S1A). The two longest domain architectures were for the protein Titin (TTN, UniProt ID: Q8WZ42) at 303 domains and Obscurin with 66 domains (OBSCN, UniProt ID: Q5VST9) (Fig S1A). Given the roughly 5-fold difference between the two largest domain architectures and TTN being an outlier in our protein set, we extracted n-grams with a length up to 66 domains. In total, 44,425 n-grams were extracted from >18000 proteins. The majority of n-grams were found in only one protein (Fig. 1C). Eight out of the top ten most reoccurring domain n-grams were repeats of the Zinc Finger C2H2 type (Znf-C2H2) domain (Fig. S1B). Beyond the highly repetitive Znf-C2H2 domain containing n-grams, the protein kinase domain and the seven transmembrane region of rhodopsin-like G-protein coupled receptors (GPCR_Rhodpsn_7TM) were the only additional 1-grams within the top 10 overall n-grams (Fig. S1B). Given the dominance in overall n-grams due to repetition, we identified additional n-grams that contain domain repeats. We retrieved n-grams with 2 or 3 domain repeats and found n-grams consisting entirely of the EGF-like, Cadherin-like, Fibronectin type 3 (FN3), Ig-subtype 2 (Ig_sub2), or the RNA recognition motif (RRM) domains, which occured in more than 100 proteins each (Fig. S1C). Only one heterotypic n-gram (an n-gram that contains different domains) existed within the top 10, which contained both the Znf-C2H2 domain and the KRAB (Krueppel-associated box) domain (Fig. S1C). The KRAB domain has been noted to only occur in proteins with Znf-C2H2 domains and acts as a transcriptional repressor[29, 30]. Interestingly, n-grams with repetitive domains occurred within roughly 100-200 proteins except for n-grams containing the Znf-C2H2 domain that occur in about 700 proteins. Further, the KRAB and Znf-C2H2 domain n-gram was also the only multidomain n-gram that was returned when retrieving either 1-grams or heterotypic n-grams (Fig S1D). Interestingly, Znf-C2H2 containing proteins are one of the largest classes of transcription factors and Znf-C2H2 domain mediates DNA interactions by recognizing 3 or more bases to create a diverse range of recognition motifs[31, 32]. However, while the Znf-C2H2 domain was the smallest of the domains identified in the top n-grams (Fig. S1E), other small domains like the homeobox domain (HD) and Znf-RING domains were not found in large protein families with highly repetitive copies of each domain. Collectively, these results establish the diversity of domain n-grams that exist across the proteome along with helping identify obligate grammar structures (such as the KRAB|Znf-C2H2|Znf-C2H2 3-gram). However, these results highlight that certain protein families, such as the Znf-C2H2 containing proteins, dominate n-gram counting metrics and may obscure the importance of other critical domain families.

Given that n-grams are structured units of language, we can consider n-grams that share common words to have a connection between the individual n-grams, which represents a functional property that determines feasible domain locations and combinations. Thus, we assembled the n-grams of the complete proteome as a network where nodes represent individual n-grams (e.g. SH2, SH3|SH2, or

KRAB|Znf-C2H2), and edges designate parent-child relationships where shorter n-grams are found within the longer n-grams (Fig. 1D). For example, if we consider the SH3|SH2 2-gram found within the SRC and ABL kinase family architecture, it will be connected directly to the SH2 and SH3 1-grams, and the SH3|SH2|Kinase 3-gram. We extended this to the 44,245 n-grams we extracted from the proteome to construct a complete n-gram network (Fig. 1D). We then used gross topographical network features, like the number of connected components, to identify n-gram families that share a common set of words. The n-gram network of the proteome contained 1345 connected components and 700 isolates, which represent domains only found on their own across proteins (Fig. S2A). Most non-isolate connected components represented 5 or less proteins, while the largest connected component represented 9937 proteins (Fig. 1E). Given that only a small number of domains have a diverse number of domain partners[8], and whole protein families may be represented within individual connected components of the n-gram network, we determined how integrating a node collapsing step within the domain n-gram network construction alters the overall network topology. By collapse groups of n-grams that fully represent the same set of proteins we can reduce the redundancy of information encoded by each node, and further identify n-grams that represent distinct grammar structures (Fig. 1F). The collapsed network (Fig. 1G) increased the number of isolates to 1106 nodes with most representing n-grams of length 5 or less, but one isolate represents an n-gram family containing 14 domains (Fig. S2). Altogether, this suggests that the collapsed n-gram network maintains connections that represent non-redundant grammar structures, which can then be used to explore the rules within the human proteome.

Next, we asked if we can use node-specific metrics to identify critical n-grams of the network. For this, we calculate: 1) the degree centrality – to identify the domains or domain combinations with a diverse set of domain partners, and 2) the betweenness centrality – to identify domain n-grams that connect n-gram families by laying on the shortest path between individual nodes. Using the largest connected component, which represents the most proteins within the proteome, we showcase how these centrality measurements can identify important domain n-grams within the subnetwork. For each n-gram, the degree and betweenness centrality was calculated and shows that a small fraction of n-grams have high values (>0.01) for each metric (Fig. 1H). Exploring the relationship between each centrality measurement (Fig. 1I), we can identify which n-grams act as hubs to connect other n-gram families within the network, and thus represent words with a diverse set of grammar contexts or functions. N-grams with high degree and betweenness centrality values, such as the protein kinase, EGF-like, and Znf-C2H2 domains, have a diverse set of domain partners that demonstrate their flexibility in generating word structures that maintain clear functionality. However, n-grams with only high degree centrality values can represent domain combinations that have several partners but only connect a distinct n-gram family (e.g. highly repetitive domain n-grams that only link to other repetitive n-grams). If an n-gram has a high betweenness centrality but low degree centrality, like the GPCR_Rhodpsn_7TM domain, then it likely acts as a connection between a small fraction of other n-grams to the rest of the network. Even if the individual n-gram is highly prevalent across the proteome, it is highly constrained in what are considered feasible domain partners. This category of n-grams can represent grammar structures that rarely require modifiers to define their action. Altogether, our results demonstrate how an n-gram network can be used to explore the rules of the protein language, and can find abstract connections which maintain protein functionality.

Characterizing the information content and network topology of different n-gram length models

Since prior studies used 2-gram models to study domain architectures[3, 8, 10], we set out to evaluate how different n-gram lengths alter the grammar structures of the protein language, which are captured within the topography of the n-gram networks. Importantly, we had found that more than 100 n-gram families comprised of 5 to 14 domains can be collapsed as isolates in the n-gram network (Fig. 1, S2), which suggests 2-gram models may overlook the contributions of supradomains in defining the connections that determine protein functionality[21]. We extracted n-grams ranging from 2-grams up to 15-grams, and constructed n-gram networks. Thus, the 3-gram model contained 1-,2-, and 3-grams, while

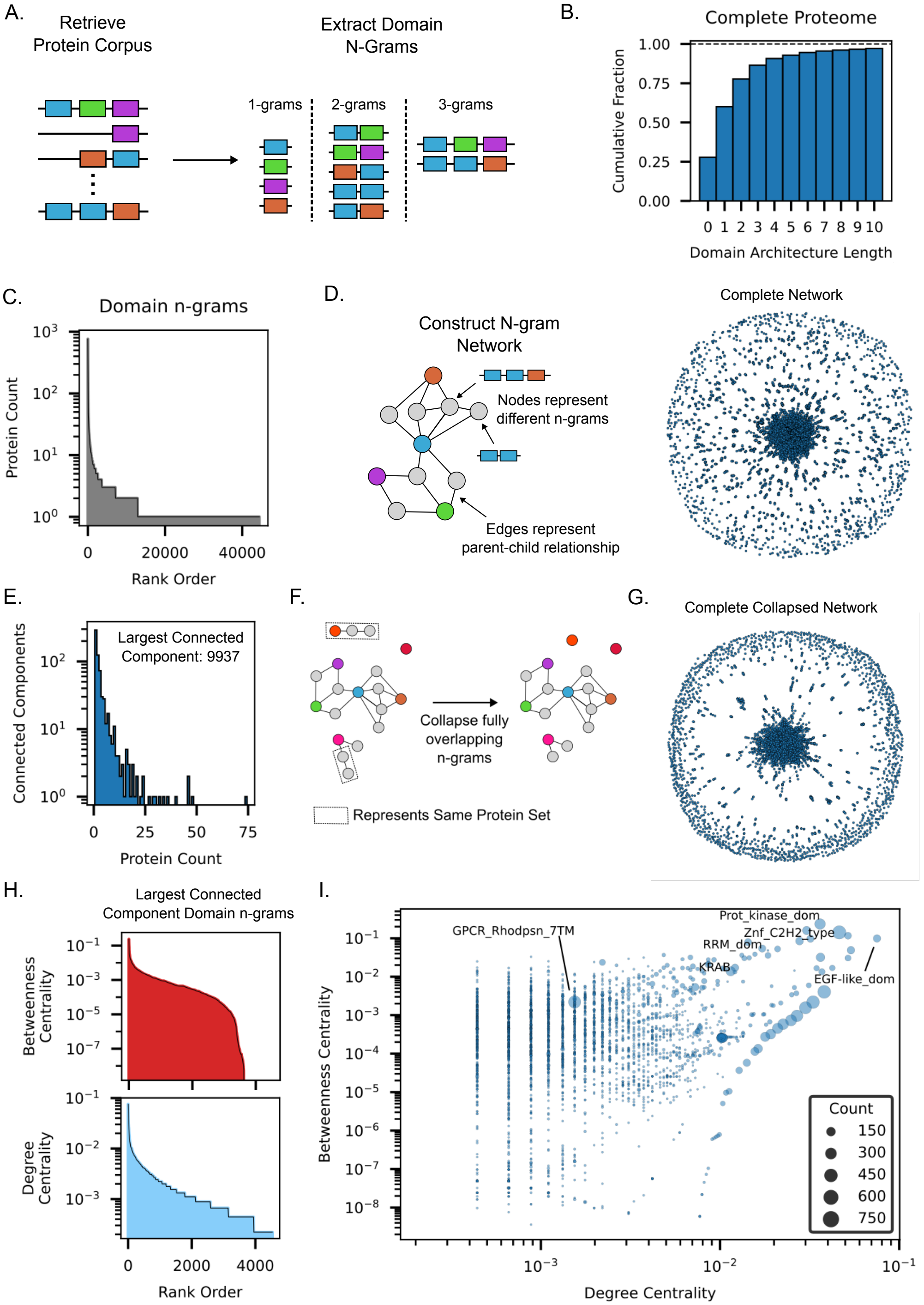

Figure 1. Human proteome n-gram network characterization. A) Overview of extracting n-grams of length n from a collection of protein domain architectures. B) The cumulative distribution of domain architecture lengths across all proteins in the human proteome. C) The number of proteins for each 44,425 domain n-grams. D) The schematic (left) of how the n-gram network was constructed with individual n-grams as nodes within the network and edges representing parent-child relationships where a shorter n-gram is found in a longer n-grams. The actual n-gram network of the complete proteome (right). E) The number of proteins represented for each connected component except the largest connected component which has 9937 proteins. F) Schematic of collapsing n-grams that which fully represent the same set of proteins. G) The collapsed n-gram network of the human proteome. H) Network centrality measurements for each node in the largest connected component of the collapsed n-gram network in rank order. I) The relationship of each centrality measurement with each other and marker size representing the protein count of each n-gram.

the 5-gram model contained all those n-grams plus 4- and 5-grams. We measured the entropy of each network model to understand broad changes through emergence or modification of grammatical rules as longer n-grams are included. We compared these entropy values to the natural distribution of domains represented by 1-grams alone, to calculate the relative information gain for each n-gram model, and relative to the full network model. A large information gain of 2 bits occurs with the 2-gram model, however gains became more modest as longer n-grams were included in each model's corpus. The 5-, 10-, and 15-gram models had information gains of 3.77, 4.56, and 4.85 bits respectively, which correspond to 71%, 85%, and 91% of the information captured in the full, collapsed n-gram model (Fig. 2A). Using the n-gram networks for the 2-, 5-, 10-, and 15-gram models (Fig 2B), we find most isolates from the original n-gram model were retained across each model. However, the 2-gram and 5-gram models also generated 771 and 28 additional isolates respectively (Fig 2C), which suggests that the n-gram families that represent the same set of proteins are being spread across multiple nodes. We next analyzed the changes in non-isolate connected components, and found the 2-gram model only recapitulated half of the existing components in their entirety, and split 16 components into two or more additional components. Meanwhile, the 5-gram split 2 connected components and the 10- and 15-gram models only truncated the connected components by removing nodes associated with longer n-grams (Fig 2D). The 15-gram model only truncated the largest connected component. We calculated the relative entropy of each model to the complete n-gram model to determine the divergence of the n-gram probability distribution, and thus the information content, within each individual model. Like the network changes, the 2-gram model exhibit the highest relative entropy (i.e. largest difference), while the 5-, 10-, and 15-gram models had relative entropy values less than 0.5 bits. Collectively, our results suggest that a 2-gram model is insufficient to accurately recapitulate the diversity of domain n-grams while representing the same set of proteins across different 2-grams. However, n-gram models that include up to 15 domains within an individual n-gram can recapture most of the diversity, but will lose information related to longer n-grams found in roughly 5% of the proteome. However, we observed relatively minimal gains in information content and minimial changes in network topology between 10- and 15-gram models. Thus, for representing the human proteome, we selected a 10-gram model, which appears to be a nice tradeoff between maximizing the information encoded within protein domain architectures and complexity.

Domain modules within reversible PTM systems infrequently share n-grams.

Given that we have established a representation of domain architectures for the entire proteome, we wanted to explore the insights that can be generated from n-gram networks constructed for individual signaling subnetworks, like the phosphorylation system. Phosphorylation is found across many biological processes[33], and is suggested to have developed because the biochemical properties of phosphate groups allow it to be rapidly and readily reversible[34]. This diversity of biological functions can also help explain why the kinase domain is one of the critical n-grams within the complete proteome n-gram network. Signal transduction pathways mediated by phosphorylation are tightly controlled by a three module system which operates under a reader-writer-eraser paradigm. For example, the pTyr machinery consists of the Tyr kinase (writer), Tyr phosphatase (PTP, eraser), SH2 and PTB (readers) domains. Kinases families fall under two broad families: the pTyr and pSer/Thr. However, the pTyr system is considered to be evolutionarily newer having evolved near the origins of metazoan species[35, 36]. Unfortunately, the transient nature of phosphorylation and the relatively low affinity of reader domains have made protein-protein interaction network definition especially challenging. To evaluate if linguistic approaches can identify unique characteristics that differ between the two systems, we generated n-gram networks on the components of each phosphorylation system separately (domains and classification in Table S1). Specifically, we generated two n-gram networks for each system, one which contains the complete n-gram corpus, and one with n-grams that only contain at least one of the domains of the PTM system (PTM System Domain Focused). By analyzing the PTM System Domain Focused networks, we can study broad differences in the grammar of system, and how it constrains the combinations of different word types represented by each domain to modulate the overall system. The pTyr system generated a complete connected graph (i.e. a single connected component), while the pSer/Thr system had multiple connected components with n-grams containing the phosphatase (eraser), 14-3-3 (reader),

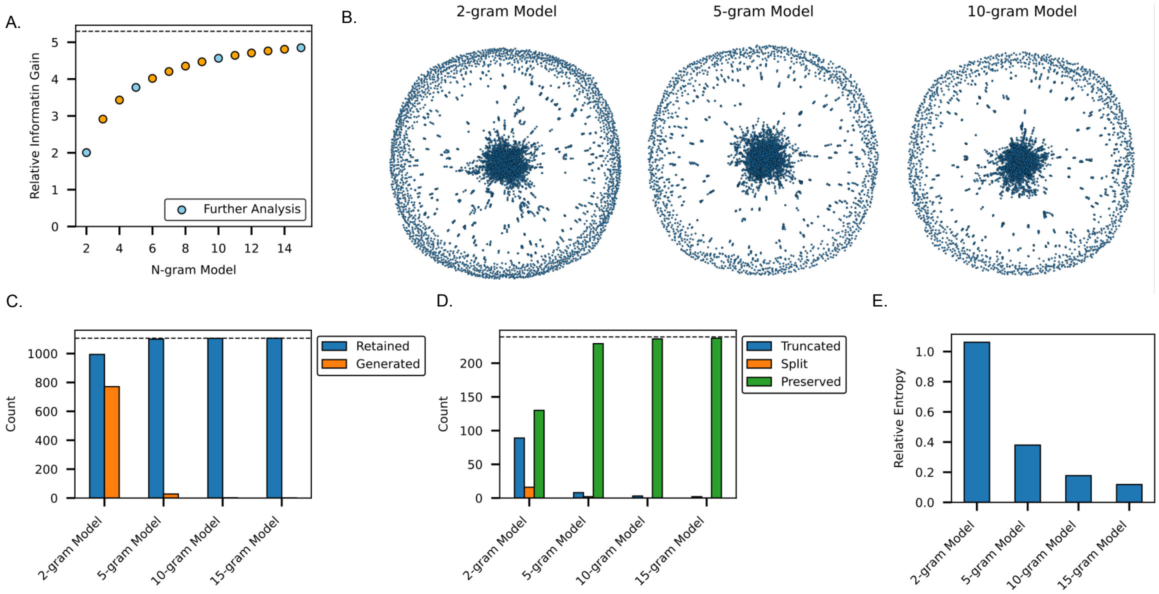

Figure 2. Comparison of different n-gram length models. A) Information gain of different n-gram models relative to a unigram model. B) The n-gram networks of the specified n-gram models. C) The number of isolates generated or retained in each individual n-gram model. D) The number of connected components that were either truncated, split, or fully preserved. E) The relative entropy of each n-gram model to the complete proteome model.

or MH2 (reader) domains as individual, disconnected nodes from the rest of pSer/Thr machinery (Fig. 3A). Comparing the complete versus PTM System Domain Focused n-gram networks of the pTyr system highlighted that only the erasers PTPN6 and PTPN11 with the domain architecture SH2|SH2|PTP use other pTyr regulatory domains. Meanwhile, the SH2 and Tyr Kinase domains share multiple n-grams supporting findings that SH2 domains modulate kinase processivity[19, 37]. Within the PTM system domain focused network of the pSer/Thr system only the FHA domain was directly connected to Ser/Thr-kinase domain subnetwork (Fig. 3A). Collectively, these results suggest a common set of rules between the two systems, which includes that eraser domains rarely if at all use the other system components to modify their function.

To determine if the network topologies of the pTyr or pSer/Thr systems were characteristic of other reversible PTM systems that share the reader-writer-eraser paradigm, we generated networks for the acetylation, methylation, and ubiquitin systems. From these networks, we found they more closely resembled the pSer/Thr system with multiple connected components (Fig. 3B, S3A). However, the methylation system had n-grams with JmjC eraser domains connected to the rest of the network within the PTM domain focused network, and the DOT writer domains were isolated within both networks (Fig. S3A). Notably, the PTM System Domain Focused n-gram networks showed each PTM system the eraser domains are rarely found within the n-grams that contain other system components. To determine if the PTM System Domain Focused networks were sufficient to recapture the information within the complete n-gram networks, we calculated the relative entropy between the two models, and found that the domain focused networks encode similar n-gram distributions across each system (Fig. S3B). Collectively, these results suggest these PTM systems have developed a common set of grammatical rules that determine the feasible domain combinations to maintain biochemical functions. Interestingly, these rules suggest that eraser domains do not require additional domains of the system to modify their activity. Thus reader domains that act as "adverbs" to modify activity of the writer domains, rarely are used to modulate erasers. Additionally, these networks suggest a constraint that the catalytic domains - writers and erasers - that act as "verbs" must be on separate proteins that likely prevents inefficient and futile processing of the PTMs.

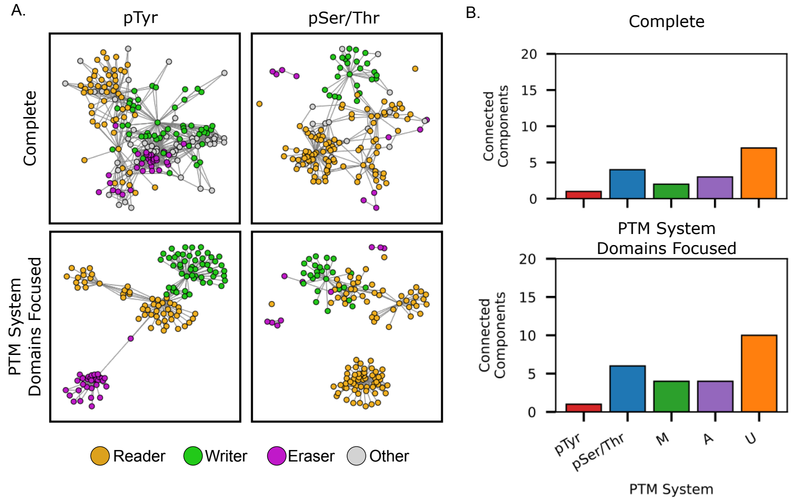

Figure 3. Characterizing the n-gram networks or reversible PTM systems. A) N-gram networks for the phosphotyrosine (pTyr) or phosphoserine/threonine (pSer/Thr) machinery. The top row contains all possible domain n-grams to represent a complete network representation, while the bottom row has n-grams that must contain individual components of the machinery. With the PTM System Domain Focused networks any node which represents both a reader and writer/eraser domain will be colored based on writer/eraser colors. B) The number of connected components for both the phosphorylation systems and additional reversible PTM systems which operate under a reader-writer-eraser paradigm. M: methylation, A: Acetylation, Ub: Ubiquination

Characterizing the evolution of the phosphorylation domain modules.

Since the pTyr n-gram networks was the only PTM system that could be represented as a complete graphs and it is considered one of the most recently evolved PTM systems[35, 36], we wanted to determine if the n-gram network topology had evolved over time and could reflect the evolutionary age of the system. We retrieved both the pTyr and pSer/Thr systems from 20 species starting from *Saccharomyces cerevisiae*, which contains a single proto-SH2 domain and three PTP domains[35] (Fig. S4). Assembling the n-gram networks for each species, we observe a rapid expansion of both the n-grams and edges between n-grams for the pTyr system but not the pSer/Thr system (Fig. 4A, S7, S8). The emergence of metazoans led to the stabilization of the pTyr n-gram network, and is reflected in the number of connected components and the relative distributions of individual domains (Fig. 4A,B, S4). For the pSer/Thr system network topology, the separated Ser/Thr phosphatase subnetwork was established early and few changes in the distribution of individual domains (Fig. 4A,B, S5). These observations were further supported by using relative entropy to compare the n-gram distributions for each species to a network generated from all species (Fig. 4C). The relative entropy for the pSer/Thr networks were lower to begin with in most pre-metazoan species, which reflects the increased similarity index of the n-grams found within the pSer/Thr network for these species (Fig. S6). Interestingly, these results suggest a potential convergence of n-grams in both systems during the evolution of vertebrates within our queried species. However, the individual species networks show an interesting evolution of the PTM system, which is not readily observed when using the number of individual domains to study the system evolution alone. The pTyr network of the pre-metazoan species *Capsaspora owczarzaki* and *Monosiga brevicollis* had several n-grams that connect the PTP domain to the rest of the network that are lost within metazoans. However, once processes advantageous to metazoans evolved, these n-grams, with the exception of the SH2|SH2|PTP architecture, were lost. Interestingly, these involved connections between both catalytic domains, but the only metazoan species where a connection remained was in *Nematostella vectensis*, which is one of earliest metazoans included in our analysis. This suggests during this evolutionary period representing the transition to metazoans, species were sampling several configurations of the network during the rapid pTyr system expansion. However, the pTyr system converged to the similar set of grammatical rules as the other PTM systems, where eraser domains are loosely connected to the rest of the network, and domains with opposing verb actions are not found within the same protein.

Using n-gram networks to study gene fusion domain architectures.

While advantageous domain architectures have been selected for throughout evolution, selective pressures during the development and progression of cancer present new opportunities for the exploration of new grammatical structures. In particular, the development of cancer fusion genes represents one avenue for generating novel architectures by forming chimeric proteins from genes with disparate functions. For example, the FGFR3-TACC3 fusion gene combines the kinase domain from FGFR3, a receptor Tyr kinase, with the TACC domain (a coiled coil domain) from TACC3, a protein involved in the organization of microtubules during mitosis to form a constitutively active FGFR variant[38]. Relative to the complete n-gram network, gene fusions can generate domain architecture that relate to the gross topology of the n-gram network in three different fashions: 1) the fusion n-gram already exists within the network, and leads to no change, as seen with STK11-TYK2 fusion gene, which retains only the two kinase domains from TYK2 (referred to as no change), 2) a novel domain architecture is generated, which uses domains that share a common domain or n-gram partner to shorten the network path and reinforce an existing connected component, as seen with the BCR-ABL1 fusion gene (referred to as reinforcement), or 3) the novel domain architecture connects domains that do not share any common domain partners and thus only creates articulation points to bridge connected components like the FGFR3-TACC3 fusion (Fig 5A) (referred to as connected components). To investigate the nature and extent of how cancer gene fusions alter n-gram networks, we retrieved gene fusions identified in patients from 20 study cohorts within the Cancer Genome Atlas (TCGA) from the ChimerDB[39]. We limited fusions to those mapped to the current human genome build (hg38) and result in in-frame fusions. Predicted domain architectures were generated by mapping genomic breakpoint information to the protein coding sequence, and identifying the domains being donated from each parent gene (Fig. S9). Across 20 cancer types, at least 40% of all unique fusion gene domain architectures do not create a novel domain architecture with the exceptions of Acute Myeloid Leukemia (LAML) and Testicular Germ Cell Tumors (TGCT). However, for both of these cancers, fewer than 50 fusion genes were identified in their respective studies and resulted in less than 20 unique domain architectures (Fig. 5B). Meanwhile, most of the novel domain architectures reinforced existing domain connections (20-30%), and only 10% of all fusions bridge connected components within the n-gram network (Fig. 5B). The low degree of novelly connected components across pancancer gene fusions, suggests that fusion sampling and evolutionary pressures during tumor progression rarely expand the proteome to generate proteins that span disparate functions. However, given the novelty of those fusions that do create new connections, we evaluated whether these fusions represent highly advantageous grammatical structures, which encourage a widespread development across multiple cancers. For each fusion gene, we retrieved their impact on the n-gram network, and then determined whether they were a recurrent fusion across multiple cancer types or within a single cancer. For example, the FGFR3-TACC3 fusion is highly recurrent across multiple cancers[38] and represents a fusion that bridge connected components within the n-gram network. Meanwhile, the TMPRSS2-ERG fusion is highly recurrent within prostate cancer[40], and does not alter the n-gram network. Comparing the n-gram network effects by novel domain architectures generated by these and singleton gene fusions, which have only been identified in one patient, we found irrespective of their recurrent status each group of gene fusions primarily reinforce existing domain combinations (Fig. 5C). Altogether, these results suggest that the domain architectures of gene fusions still adhere to the same grammatical rules that were established for domain architectures within the complete proteome.

Having established that gene fusions do not readily expand the proteome, we next sought to determine if the few fusions that do connect disparate n-gram families impact patient survival. We stratified patients based on the network impact of their the predicted protein domain architecture of their gene fusion, and if a patient had multiple gene fusions. A pan-cancer, univariate analysis suggested that only patients with gene fusions that reinforce existing domain connections or had multiple fusions were predicted to have worse overall survival (Fig. S10). However, for individual cancer cohorts, the

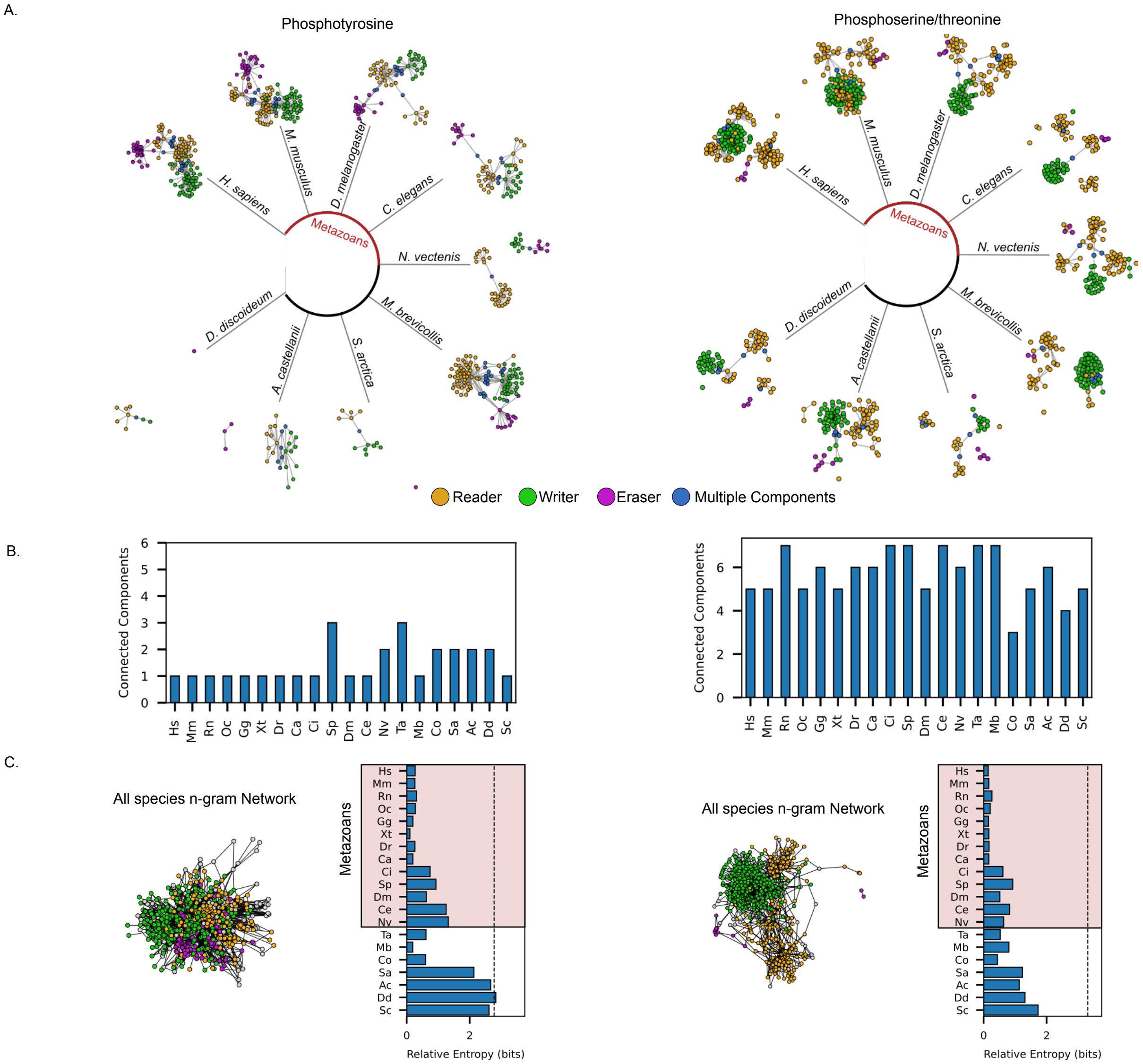

Figure 4. The evolution of the protein domain architectures of the pTyr and pSer/Thr systems. A) Complete n-gram networks of the phosphotyrosine (left) or phosphoserine/threonine (right) systems for species ranging from *Dictyostelium discoideum* to humans. B) The number of connected components for each species phosphorylation system. C) The complete n-gram network containing all n-grams across all queried species and the relative entropy for each species compared to the all species network. Hs: *Homo sapiens*, Ms: *Mus musculus*, Rn: *Rattus norvegicus*, Oc: *Oryctolagus cuniculus*, GG: *Gallus gallus*, Xt: *Xenopus tropicalis*, Dr: *Danio rerio*, Ca: *Carassius auratus*, Sp: *Strongylocentrotus purpuratus*, Dm: *Drosophila melanogaster*, Ce: *Caenorhabditis elegans*, Nv: *Nematostella vectensis*, Ta: *Trichoplax adhaerens*, Mb: *Monosiga brevicollis*, Co: *Capsaspora owczarzaki*, Sa: *Sphaeroforma arctica*, Ac: *Acanthamoeba castellanii*, Dd: *Dictyostelium discoideum*, Sc: *Saccharomyces cerevisiae*

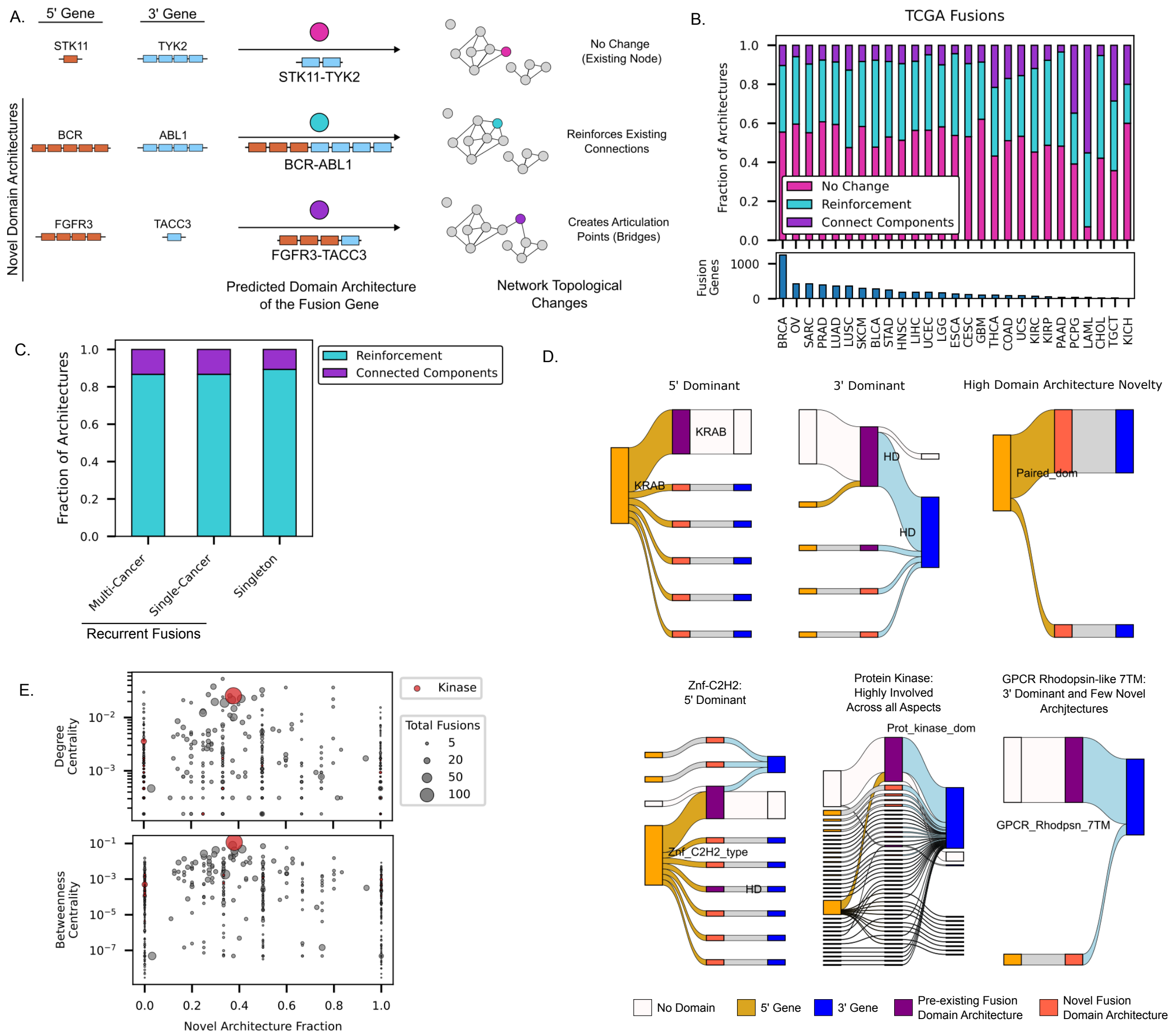

Figure 5. Characterizing predicted protein domain architectures of cancer gene fusions. A) Schematic of the possible changes to the n-gram network caused by individual fusion genes with example fusion genes are predicted domain architectures. For novel domain architectures these can either reinforce existing connected components or reduce the number of connected components. B) The fractional distribution of fusion gene impacts on the n-gram network (top) and the total number of fusion genes (bottom) retrieved for each TCGA study cohort. C) The distribution of architecture n-gram network impacts for recurrent or singleton fusions within either multiple or single cancer. D) Representative Sankey diagrams of select protein domains that exhibit 5'/3' donation propensity and/or generation of novel domain architectures, and diagrams for the top 3 most prevalent domains across the human proteome. E) Network centrality measurements relative to the fraction of novel domain architectures individual domain n-grams generate. N-grams which contain the protein kinase domain are highlighted in red.

presence of gene fusions generally did not predict worse patient survival, except for seven cancers (Fig. S11). Further, stratification by network impacts of gene fusions, and excluding patients with multiple fusions, showed most cancers did not have significant changes in patient survival. However, three cancer cohorts had significantly worse patient survival based on fusion gene stratification, but each cohort had less than 25 patients (5-10% of the patient population) with a single fusion gene limiting the conclusions that can be made on changes in survival probabilities by the n-gram network categories (Fig. S12). Together, these results corroborate that gene fusions are relatively rare events compared to somatic point mutations[41–43]. Further, our results suggest that while gene fusions can identify therapeutic candidates[40, 44], they may not be bona fide biomarkers to predict patient survival, and can even represent passenger mutations[45].

Given that our results have established that n-gram networks can highlight broad descriptions of the rules guiding domain architecture development, we constructed an n-gram network on the gene fusions and analyzed the domain architectures of the parent genes. Similar to the natural proteome, most parent genes have domain architectures with two or fewer domains (Fig. S13A). Extracting the individual n-grams donated by the parent genes and the final fusion architecture, we generated a new n-gram network that tracked whether an n-gram was from a parent gene or the final fusion. Thus, if a kinase domain was donated from the 3' fusion gene (e.g. with TYK2) it would be a separate node in the network than a kinase domain that represents the fusion gene (e.g. the complete STK11-TYK2 fusion). From this network, a total of 106 fusion gene domain architectures utilized domain architectures only found from their parent genes to form unique complete domain fusion families (Fig. S13B). The large connected component within the network suggests several domain architectures are highly recurrent across either parent genes or the final fusion and span multiple fusion families. Importantly, while most parent genes have domains, many of the fusion genes did not involve any of the domains from at least one parent gene (Fig. S13B-D). Without a domain from one of the parent genes, a novel domain architecture was not produced, which would explain the large fraction of fusion genes that led to no change in the n-gram networks across the TCGA cohorts (Fig. 5B). Next, we identified domain n-grams that either 1) generate a large fraction of novel domain architecture or 2) display a propensity to be donated by either the 5' or 3' parent genes. We find domains such as the KRAB domain and Paired domain are frequently donated by the 5' parent, while the homeobox domain (HD) is donated by the 3' gene. However, most fusions involving the KRAB or HD domains frequently have no domains donated from the other partner gene, resulting in a fusion with only the KRAB or HD domain. If the fusion gene partner for the KRAB containing proteins does donate a domain, it results in a novel domain architecture. Meanwhile, some domains like the Paired domain only result in novel domain architectures (Fig. 5D). When analyzing fusions containing the Znf-C2H2 or protein kinase domains given their importance in the natural proteome, we find each had at least 50% of fusion result in novel domain architectures. However, when we analyze fusions with the GPCR-Rhodpsn-7TM domain, which is found in >600 proteins in the proteome, but does not appear to have a critical role in the n-gram network architecture (Fig. 1I), we find it only within seven fusion genes, and it rarely results in novel domain architectures. Altogether, these results suggest that individual domains have an intrinsic property that determines their role within generating gene fusions, which can be uncovered with these linguistic networks.

Individual domains can vary in the number of multiple domain partners[4, 7, 8] they combine with, and our networks can identify this through centrality measurements. To determine if domains involved in gene fusions were the promiscuous domains, we calculated the Gini-Simpson Diversity Index for these domains within the natural proteome. For the diversity index, we calculated it for both the diversity of domains that precede an individual n-gram and the domains that follow the n-gram. This allows us to determine whether an n-gram has a propensity to have a diverse set of partners before or after it, and relate it to whether the n-gram is donated primarily by the 5' (reflecting diversity indices for n-grams following it), or the 3' (reflecting the preceding n-gram diversity index). For all n-grams within fusions genes, the distribution of the diversity indices were skewed slightly higher than the rest of the proteome, but this was not a defining characteristic of the top overall fusion n-grams (Fig. S14). However, for the KRAB and Znf-C2H2 domains, their propensity to generate novel domain architectures may be related to having multiple existing domain partners – as reflected in a high degree centrality – but being comprised of one dominate n-gram, lowers their diversity index (Fig. 1I, S14B). Alternatively, the GPCR-Rhodpsn-7TM domain suggests its propensity to not form novel domain architectures may be due to having few other n-grams it connects, and having single n-gram dominate the collection of domains in combines with and result in a diversity index close to 0 (Fig. 1I, S14B). Meanwhile, for the protein kinase and HD domains, these trends with the diversity index could not explain their properties within fusion gene domain architectures. Rather, their centrality measurements within the complete n-gram network could help explain their involvement in both preexisting and novel domain architectures. However, when plotting the novel domain architecture fraction against the centrality metrics for all n-grams within fusion genes, we found no correlation (Fig. 5E). Altogether, our results further suggest that the fusion gene domain architectures follow the principles regulating domain architecture development. Further, the collection of properties including the diversity of domain partners and the involvement in connecting obligate proteins can contextualize the grammatical rules for individual n-grams.

Kinase fusion have been frequently identified across most cancers[44] with the tyrosine kinase domain specifically being overrepresented within gene fusions[42, 46, 47]. We wanted to understand whether kinase fusions identified in our analysis differed if a kinase domain was related to the pTyr or pSer/Thr system. We found similar numbers of both the pSer/Thr and pTyr kinase domains across all fusions and both domain families generated the same fraction of novel domain architectures (Fig. S15A,C). Constructing an n-gram network from these gene fusions only 20 n-grams were common to both kinase fusion types and the most highly connected nodes within the network were either specific to pSer/Thr kinases (AGC-kinase, C-terminal domain) or common to both (FN3, Ig subtype 2, or the PH domains) (Fig. S15B-D). Collectively, these results suggest that kinase domains generally have intrinsic properties that enable neofunctionalization through diverse domain combinations. This result further supports our suggestion that the high centrality measurements of the kinase domain reflect its involvement across biological functions by connecting distinct n-gram families.

Discussion

Here, we applied n-gram analysis with network approaches to characterize the protein domain and multidomain landscape of the human proteome. Assembling the domain n-grams into networks agrees with past findings, that a small fraction of domains combine with a diverse set of domain partners[3, 7, 8] (Fig. 1, S1), but also highlights that about 300 multidomain architectures represent single proteins or protein families (Fig. S2). About 95% of the human proteome contains up to 10 domains which allows for a 10-gram model to sufficiently recover the distribution of domain n-grams and recreate a complete n-gram network that 2-gram networks cannot (Fig. 2). To understand the insights n-gram networks can provide for individual signaling subsystems, we further investigated the different phosphorylation systems. We found that the domains making up the pTyr system were unique in generating a complete connected graph and that this property likely evolved during the origins of metazoan species (Fig. 3, 4). Interestingly, the n-gram network topologies highlights how selective pressures from evolution can generate specific n-grams to bridge different PTM system components such as the SH2|SH2|PTP architecture, but represent a small fraction of the complete system. Further, this evolutionary analysis suggests that each PTM system converges toward a common set of grammatical rules, such as eraser domains being loosely connected to the rest of system, that reflects the evolutionary age of the PTM system. Since cancer progression can represent a potentially active selection process that can create novel domain architectures to expand the grammatical rules of the proteome, we analyzed the gene fusions. We found few fusions connect domains with obligate functions that do share common domain partners (Fig. 5), but certain domains such as the protein kinase domain are frequently found within fusion genes (Fig. 5F). However, these fusions cannot easily predict patient survival outcomes. Studying gene fusions highlight that the principles that determine feasible domain combinations from evolution remain during cancer progression. Collectively, the results highlight the uses of domain architectures to study molecular functions beyond predicting the evolution and functionality of protein families. Further, these results highlight that our n-gram network analysis can uncover rules akin to grammar that determine feasible domain combinations, which can complement existing protein-protein interaction networks by abstracting the functional connections within signaling subnetworks.

We applied the n-gram network analysis to cancer fusion genes to understand the frequency and types of novel domain architectures being generated. From this analysis we found that most predicted fusion gene domain architectures result in preexisting n-grams agree and that novel domain architectures primarily utilizes domains with common domain partners (Fig. 5), but rarity of gene fusions compared to other somatic mutations[41–43] makes findings on changes in patient survival inconclusive (Figs. S10, S11). Kinase fusions though represent one of the largest classes of gene fusions and have been widely identified across a variety of cancers[42, 44], which our full pancancer n-gram network analysis corroborated (Fig. 5F). One of the other major classes of frequent gene fusions are transcription factors fusions, which when involved in gene fusions are suggested to exert dominant negative effects[46]. In our analysis, only Pointed|ETS transcription factor domain architecture was one of the most common domain n-grams in our dataset, but rarely generated novel domain architectures (Fig. S13C) nor impacted n-gram network topology. This reflects the prevalence of ETS fusions especially through the TMPRSS2-ERG fusion within prostate cancer[39, 48], but has limited correlations to clinical outcomes[40, 49]. However, this in combination with our results on patient survival (Fig. S10, S11) emphasize the complexity of interpreting the presence of gene fusion in patient prognosis. Few gene fusions are considered to be putative drivers of disease[50], and results have suggested in some instances gene fusions are passenger aberrations[45]. However, recent studies have identified chimeric mRNA species predicted to generate fusion genes within disease-free tissues[51, 52], but the contributions of these gene fusions to cancer development or prognosis remain unclear.

Interestingly, by incorporating evolutionary analysis of the n-gram networks for the pTyr and pSer/Thr systems (Fig. 4), we uncovered a set of grammatical rules that each reversible PTM system appears to converge towards. During the rapid expansion of the pTyr system, which established pTyr residues as novel, orthogonal signaling currency[36, 53], species are sampling several configurations of the rules that determine domain ordering. However, evolution still promotes a convergence of rules which keep domains that act as "verbs" separate from one another, and loosely connect the eraser domains if at all with the rest of the system. The pTyr system is not the only signaling system that recently evolved. The KRAB domain, which was also identified across our analyses to exert some influence in the n-gram network, recently evolved during the transition towards vertebrates[54] to counter the expansion of transposable elements within mammalian genomes[55]. While the function of many KRAB containing Zinc-finger proteins (KRAB-ZFPs) have not been widely characterized, their role in embryonic development have been widely appreciated due to KRAB-ZFPs recruitment of KAP1 to modulate chromatin states[56]. This rapid evolution, similar to the pTyr system, emphasize that domains and their combinations not only encode the evolutionary jumps of protein families[5] but reflect changes in the molecular ecosystems available to cells. However, our n-gram networks can uncover the guiding grammatical rules of individual signaling subnetworks, and could be applied to further study and understand KRAB-ZFPs and the wider transposable element regulatory system.

Our analysis has generated a computational framework for describing domain architectures using both linguistics and network approaches. However, our analysis still is limited in describing the complete domain landscape present within cells. We retrieved and analyzed only the canonical protein isoforms, which can omit proteoformes of individual genes that arise due to alternative splicing that impact protein-protein and domain-domain interactions[57]. However, alternative splicing infrequently impacts domain architectures but when reported involves repetitive domains such as the Znf-C2H2 and Ig-like domains[58]. Additionally, the continued improvement of protein structure algorithms like AlphaFold[16] have led to an expansion of the predicted structural folds without known functions but still represent the evolution of protein families[28, 59]. Altogether, these suggest the proteome continues to evolve through various mechanisms. However, the computational framework we have described here is flexible towards describing and characterizing the expanded proteome, which can be further supplemented by molecular interaction or function annotations.

Materials and Methods

Retrieving UniProt IDs for the human proteome and post-translational modification systems for additional species. For the human proteome, UniProtKB IDs were retrieved using pybiomart to map between Ensembl gene IDs and reviewed UniProtKB IDs. For individual post-translational modification (PTM) systems, InterPro IDs for the different domains that make up the PTM system were retrieved using the IDs listed in Table S1 using the InterPro module of CoDIAC[24]. For analyzing the phosphorylation systems in all species except humans, mice (*Mus musculus*) and rats (*Rattus norvegicus*) both reviewed and unreviewed UniProt records were fetched. To ensure unreviewed UniProtKB records are not mapping to same gene within individual species, UniProt entries on the evidence level, cross references to Ensembl, the NCBI Gene, and RefSeq databases, protein length, and gene symbol were fetched to determine which record represented identified genes and full protein coding sequences. Reviewed records were given top priorty followed by records with cross references to other databases. Records which mapped to the same cross reference IDs were then compared by protein existence levels and finally amino acid length sequence. For UniProtKB records which required comparing amino acid length sequences, the record with the longest length was retained. The complete list of species and strain names used during fetching is provided in Table S2.

Fetching InterPro protein domain architectures and generating domain n-grams. To fetch protein domain architectures from InterPro, we utilized the UniProt module from our recently developed python package CoDIAC[24] by inputting fetched UniProtKB IDs into the UniProt module from CoDIAC. The resulting reference file contains all InterPro and UniProt domain architectures and reference sequences for each queried protein and was used for downstream analysis. All results were fetched using the 2024 September 4th build of the InterPro and UniProt databases. The retrieved protein domain architectures were then separated into n-grams of the length of interest for each n-gram model. For building the PTM System Domain Focused n-gram models, only n-grams that contained domains of interest were fetched from the complete protein domain architecture.

Measuring n-gram model information gain and relative entropy. N-gram models rely on the Markov assumption where the probability of a specific n-gram depends on the conditional probability of the next domain, d_n, given the preceding sequence of domains, d_{n-N+1}. This can can be estimated for n-gram using the maximum likelihood estimate (MLE):

$$p_{MLE}(d_n|d_{n-N+1:n-1}) = \frac{C(d_{n+N-1:n-1}, d_n)}{C(d_{n+N-1:n-1})}$$

Where N represents the maximum length of n-grams being evaluated (i.e. N=2 for bigrams or N=5 for 5-grams). The counts of a specific n-gram $(d_{n+N-1:n-1}, d_n)$ is represented by $C(d_{n+N-1:n-1}, d_n)$. These probabilities are then used to calculate the entropy of an n-gram model $(H_n(x))$ using Shannon's entropy:

$$H_n(x) = -\sum p(x)\log_2 p(x)$$

For individual domains the probabilities are given by the relative frequencies for each domain within the corpus of domain architectures. For longer n-grams, the entropy represents the sum of weighted probabilities which can be estimated by:

$$\begin{aligned} H_n(x) &= -\frac{1}{N_{ng}}\sum C(d_{n+N-1:n-1})\log_2\frac{C(d_{n+N-1:n-1}, d_n)}{C(d_{n+N-1:n-1})} \\ &= -\frac{1}{N_{ng}}\sum C(d_{n+N-1:n-1})\log_2 p_{MLE}(d_n|d_{n+N-1:n-1}) \end{aligned}$$

Where N_{ng} represents the total number of n-grams with length N. The entropy of an n-gram model is then used to determine the relative information gain $I(x)$ from the unigram (only single domain frequency distributions) model ($H_1(x)$):

$$I(x) = H_n(x) - H_1(x)$$

To compare how the distributions of n-grams change using different n-gram models we use the relative entropy also known as the Kullback-Leibler divergence defined as:

$$D_{KL}(P||Q) = \sum P(X) \log_2 \frac{P(X)}{Q(X)}$$

Where $P(X)$ and $Q(X)$ represent the probability distributions within both n-gram models and $Q(X)$ is the baseline model that contains all n-grams within $P(X)$.

Calculating network centrality measurements. For nodes within the largest connected component, the betweenness and degree centrality measurements were calculated as implemented in the networkx python package. Degree centrality is defined as the total fraction of all nodes connected to node v. The betweenness centrality measurement of node v is the sum of fractions of pairwise shortest paths that pass through node v, which is defined as:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Where V is the set of nodes, $\sigma(s,t)$ represents the number of shortest paths between nodes s and t, while $\sigma(s,t|v)$ represents the number of shortest paths that go through node v.

Predicting domain architectures for gene fusions. Genomic breakpoints for gene fusions were retrieved from the ChimerDB[39] for the TCGA cohort. Fusions that were mapped to the current human genome build (hg38), predicted to be in-frame, and were designated as found within the ChimerSeq+ dataset representing high confidence fusions were selected for further analysis (5579 total fusions). Genomic breakpoints were then retrieved for each parent gene and mapped to exon and the protein coding sequence positions. The base pair position was then translated to an amino acid position and used to determine which domains were donated towards the fusion gene (Fig. S9). Domains which were truncated by the breakpoint location were not included in the final predicted domain architecture.

Calculating Gini-Simpson Diversity Index for domain n-grams. For n-grams of with a length of 10 or less, for each domain n-gram of interest d_n all n+1 n-grams containing the n-gram were retrieved. The set of domain n-grams were then split into n-grams where d_n started or ended the n-gram. For the n-grams starting with d_n were used to calculate the diversity index for following n-grams, while those ended were used for the diversity index of preceding n-grams. The diversity index was calculated using the Gini-Simpson Diversity Index as defined for small datasets:

$$D = 1 - l = 1 - \frac{\sum_{i=1}^{R} n_i(n_i - 1)}{N(N-1)}$$

Where R is the collection of n+1 domain n-grams, N is the total count of the n-grams in the set, and n_i is the count for each individual domain n-gram.

Species N-gram Similarity Index For comparing the n-grams of the pTyr and pSer/Thr systems between individual species, the Jaccard similarity index, J, was calculated using:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B represent the sets of domain n-grams found within individual species.

TCGA Patient Survival Analysis. Clinical attributes for the TCGA study cohorts were retrieved from cBioPortal[60]. For univariate patient survival analysis, patients were stratified on if a tumor harbored one or multiple of the fusion genes within our analyzed fusion gene dataset. Patients with single fusions were then further stratified based on the n-gram network impacts caused by the fusion gene (Fig. 5A). Statistically significant changes in survival probabilities were determined by using pairwise log-rank tests as implemented in the lifelines python package.

Acknowledgments

We would like to acknowledge the constructive and critical assessment of the presented work from the entire Naegle lab. The authors thank Bruce Mayer for helpful discussions about the phosphotyrosine system of *Monosiga brevicollis*. This work was supported by funding from the National Institute of General Medical Sciences (5R35GM138127).

References

1. Mayer BJ (2015) The discovery of modular binding domains: building blocks of cell signalling. *Nat Rev Mol Cell Biol* 16(11):691–8.

2. Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. *Trends Biochem Sci* 33(9):444–51.

3. Ekman D, Björklund AK, Elofsson A (2007) Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol* 372(5):1337–48.

4. Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol* 8(4):319–30.

5. Jin J, et al. (2009) Eukaryotic protein domains as functional units of cellular evolution. *Science Signaling* 2(98).

6. Buljan M, Bateman A (2009) The evolution of protein domain families. *Biochem Soc Trans* 37(Pt 4):751–5.

7. Apic G, Huber W, Teichmann SA (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomics* 4(2/3):67–78.

8. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. *Genome Research* 18(3):449–461.

9. Basu MK, Poliakov E, Rogozin IB (2008) Domain mobility in proteins: functional and evolutionary implications. *Briefings in Bioinformatics* 10(3):205–216.

10. Yu L, et al. (2019) Grammar of protein domain architectures. *Proceedings of the National Academy of Sciences* 116(9):3636–3645.

11. Hayete B, Bienkowska JR (2005) Gotrees: predicting go associations from protein domain composition using decision trees. *Pac Symp Biocomput* pp. 127–38.

12. Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem* 277(48):45765–9.

13. Ofer D, Brandes N, Linial M (2021) The language of proteins: Nlp, machine learning & protein sequences. *Comput Struct Biotechnol J* 19:1750–1758.

14. Bepler T, Berger B (2021) Learning the protein language: Evolution, structure, and function. *Cell Systems* 12(6):654–669.e3.

15. Lin Z, et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379(6637):1123–1130.

16. Jumper J, et al. (2021) Highly accurate protein structure prediction with alphafold. *Nature* 596(7873):583–589.

17. Scaiewicz A, Levitt M (2015) The language of the protein universe. *Curr Opin Genet Dev* 35:50–6.

18. Errington WJ, Bruncsics B, Sarkar CA (2019) Mechanisms of noncanonical binding dynamics in multivalent protein-protein interactions. *Proc Natl Acad Sci U S A* 116(51):25659–25667.

19. Mayer BJ, Hirai H, Sakai R (1995) Evidence that sh2 domains promote processive phosphorylation by protein-tyrosine kinases. *Curr Biol* 5(3):296–305.

20. Wang M, Kurland CG, Caetano-Anollés G (2011) Reductive evolution of proteomes and protein structures. *Proc Natl Acad Sci U S A* 108(29):11954–8.

21. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol* 336(3):809–23.

22. Luck K, et al. (2020) A reference map of the human binary protein interactome. *Nature* 580(7803):402–408.

23. Peng X, Wang J, Peng W, Wu FX, Pan Y (2017) Protein-protein interactions: detection, reliability assessment and applications. *Brief Bioinform* 18(5):798–819.

24. Kandoor A, et al. (2024) Codiac: A comprehensive approach for interaction analysis reveals novel insights into sh2 domain function and regulation. *bioRxiv* p. 2024.07.18.604100.

25. Paysan-Lafosse T, et al. (2023) Interpro in 2022. *Nucleic Acids Res* 51(D1):D418–D427.

26. Mistry J, et al. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res* 49(D1):D412–D419.

27. Letunic I, Khedkar S, Bork P (2021) Smart: recent updates, new developments and status in 2020. *Nucleic Acids Res* 49(D1):D458–D460.

28. Waman VP, et al. (2024) Cath 2024: Cath-alphaflow doubles the number of structures in cath and reveals nearly 200 new folds. *J Mol Biol* 436(17):168551.

29. Urrutia R (2003) Krab-containing zinc-finger repressor proteins. *Genome Biol* 4(10):231.

30. Rosspopoff O, Trono D (2023) Take a walk on the krab side. *Trends in Genetics* 39(11):844–857.

31. Najafabadi HS, et al. (2015) C2h2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* 33(5):555–62.

32. Schmitges FW, et al. (2016) Multiparameter functional diversity of human c2h2 zinc finger proteins. *Genome Res* 26(12):1742–1752.

33. Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ (2019) Illuminating the dark phosphoproteome. *Science Signaling* 12(565).

34. Hunter T (2012) Why nature chose phosphate to modify proteins. *Philos Trans R Soc Lond B Biol Sci* 367(1602):2513–6.

35. Pincus D, Letunic I, Bork P, Lim WA (2008) Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc Natl Acad Sci U S A* 105(28):9680–4.

36. Lim WA, Pawson T (2010) Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142(5):661–7.

37. Machida K, et al. (2007) High-throughput phosphotyrosine profiling using sh2 domains. *Mol Cell* 26(6):899–915.

38. Costa R, et al. (2016) Fgfr3-tacc3 fusion in solid tumors: mini review. *Oncotarget* 7(34):55924–55938.

39. Jang YE, et al. (2020) Chimerdb 4.0: an updated and expanded database of fusion genes. *Nucleic Acids Res* 48(D1):D817–D824.

40. Pettersson A, et al. (2012) The tmprss2:erg rearrangement, erg expression, and prostate cancer outcomes: a cohort study and meta-analysis. *Cancer Epidemiol Biomarkers Prev* 21(9):1497–509.

41. Mitelman F, Johansson B, Mertens F (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat Genet* 36(4):331–4.

42. Latysheva NS, Babu MM (2016) Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Res* 44(10):4487–503.

43. Forbes SA, et al. (2017) Cosmic: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 45(D1):D777–D783.

44. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C (2014) The landscape of kinase fusions in cancer. *Nat Commun* 5:4846.

45. Kalyana-Sundaram S, et al. (2012) Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. *Neoplasia* 14(8):702–8.

46. Frenkel-Morgenstern M, Valencia A (2012) Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics* 28(12):i67–74.

47. Ortiz de Mendíbil I, Vizmanos JL, Novo FJ (2009) Signatures of selection in fusion transcripts resulting from chromosomal translocations in human cancer. *PLoS One* 4(3):e4805.

48. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM (2008) Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* 8(7):497–511.

49. Esgueva R, et al. (2010) Prevalence of tmprss2-erg and slc45a3-erg gene fusions in a large prostatectomy cohort. *Mod Pathol* 23(4):539–46.

50. Mertens F, Johansson B, Fioretos T, Mitelman F (2015) The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* 15(6):371–81.

51. Singh S, et al. (2020) The landscape of chimeric rnas in non-diseased tissues and cells. *Nucleic Acids Res* 48(4):1764–1778.

52. Babiceanu M, et al. (2016) Recurrent chimeric fusion rnas in non-cancer tissues and cells. *Nucleic Acids Res* 44(6):2859–72.

53. Liu BA, et al. (2011) The sh2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes. *Sci Signal* 4(202):ra83.

54. Imbeault M, Helleboid PY, Trono D (2017) Krab zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543(7646):550–554.

55. Rosspopoff O, Trono D (2024) Take a walk on the krab side: (trends in genetics, 39:11 p:844-857, 2023). *Trends Genet* 40(2):203–205.

56. Senft AD, Macfarlan TS (2021) Transposable elements shape the evolution of mammalian development. *Nat Rev Genet* 22(11):691–711.

57. Yang X, et al. (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164(4):805–17.

58. Light S, Elofsson A (2013) The impact of splicing on protein domain architecture. *Curr Opin Struct Biol* 23(3):451–8.

59. Schaeffer RD, et al. (2023) Classification of domains in predicted structures of the human proteome. *Proc Natl Acad Sci U S A* 120(12):e2214069120.

60. Cerami E, et al. (2012) The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2(5):401–4.