TRADITIONAL AND REPRESENTATION LEARNING APPROACHES FOR PROTEIN SEQUENCE ANALYSIS

by

WAYLAND YEUNG

(Under the Direction of Natarajan Kannan)

ABSTRACT

The analysis of protein sequence information is an important part of bioinformatics, used for high-throughput predictions of protein structure, function, and evolution. While traditional analytical methods utilize sequence alignments, recent advances in representation learning facilitate alternative, alignment-independent strategies. In this work, I develop and apply both alignment-based and alignment-independent approaches to analyze the protein kinase superfamily, a biomedically-relevant and highly conserved class of signaling enzymes. Using a large curated sequence alignment, I characterized sequence variations of the αC-β4 loop across diverse protein kinase enzymes and identified the region as a major kinase regulatory hotspot. Using a more focused alignment, I characterized the functional evolution of tyrosine kinases families across diverse holozoan taxa and proposed a new representative phylogeny. Finally, I infer the evolutionary relationships which connect the protein kinases superfamily to structurally divergent lipid and small-molecule kinases using an alignment-independent approach, facilitated by sequence embeddings learned from Transformer protein language models. My work provides new insights on the functional evolution of the protein kinase superfamily using a combination of traditional and novel approaches inspired by unsupervised analytical techniques from

representation learning. The broad applicability of my sequence embedding-based framework is further demonstrated in pilot analyses of phosphatase enzymes as well as the radical S-adenosyl-L-methionine (SAM) superfamily.

INDEX WORDS: protein kinase, evolution, sequence analysis, representation learning, deep learning, artificial intelligence

TRADITIONAL AND REPRESENTATION LEARNING APPROACHES FOR PROTEIN SEQUENCE ANALYSIS

by

WAYLAND YEUNG

B.S., University of Georgia, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

TRADITIONAL AND REPRESENTATION LEARNING APPROACHES FOR PROTEIN SEQUENCE ANALYSIS

by

WAYLAND YEUNG

Major Professor:	Natarajan Kannan
Committee:	Sheng Li
	Eileen Kennedy
	Robert Woods

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2022

to that Keurig machine in the IOB office

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Natarajan Kannan, for his guidance and support during my Ph.D. study as well as everyone who has served on my committee: Dr. Robert Woods, Dr. Eileen Kennedy, Dr. Sheng Li, Dr. Casey Bergman, and Dr. Eva Strauch. I also want to thank the members of my lab, past and present, who I have worked with over the years: Dr. Annie Kwon, Dr. Rahil Taujale, Dr. Liang-Chin Huang, Dr. Zheng Ruan, Dr. George Bendzunas, Dr. Carlos Sanz, Dr. Samiksha Katiyar, Zhongliang Zhou, Claire Bunn, Nathan Gravel, Safal Shrestha, Aarya Venkat, Nolan Ross-Kempinnen, Brady O'Boyle, Niral Thaker, Grace Watterson, and Steven Scott. I also want to thank the IOB staff: April King Mosley and Sandra Getz.

TABLE OF CONTENTS

Page

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

1.1 - Motivation

Protein macromolecules are biological machines that carry out the diverse biochemical processes that enable life. As biopolymers built from amino acid monomers, a distinct protein can be represented by a sequence of letters where each letter denotes a distinct amino acid residue. The analysis of protein sequence information is a fundamental aspect of biology, which provides a means of studying unique biological processes and how they evolved. Technological advances of the post-genomic era have brought about a rapid increase in sequence information, publicly available across many online databases. In fact, a recent update from the UniProt protein sequence database reported over 189,000,000 sequence records across ~290,000 proteomes as of 2020 [1]. Many methods have been developed to harness this wealth of information towards inferring protein structure, function, and evolutionary history.

Traditional methods for protein sequence analysis are largely facilitated by sequence alignments. Invented in 1970, Needleman and Wunsch developed the first algorithm for pairwise sequence alignment [2]. Multiple sequence alignments were later developed as an extension of this framework in order to facilitate the comparison of more than two sequences [3]. Following their conception, multiple sequence alignments have remained widely used as a highly effective framework for protein sequence comparisons. In more recent times, deep learning methods have begun playing a more significant role in sequence analysis due to the rising accessibility of high-performance computing. Building upon the established framework, many deep learning models also utilize a sequence alignment-based framework to represent protein sequence information. Providing the means for an alternative strategy, recent advances in unsupervised representation learning have yielded methods for meaningfully translating protein sequences from strings of discrete characters into matrices of continuous floating-point numbers called embedding vectors [4,5]. The translation of discrete protein sequence information into a continuous numeric space opens up exciting new possibilities for data analysis.

Here, I sought to develop an array-based framework for modern alignment-based protein sequence analysis within the scientific Python ecosystem [6]. Utilizing sequence embeddings, I also developed a framework for alignment-independent protein sequence analysis. I apply these tools towards the study of the protein kinase superfamily, a biomedically relevant set of enzymes that control cellular signaling. By using a combination of conventional and non-conventional means, I shed light upon the complex evolutionary history of the protein kinase superfamily. By identifying unique features which coincide with the emergence of new protein kinase families, I gain insight into family-specific functions. Finally, I demonstrate the broader applicability of my embedding-based approach towards the study of other protein superfamilies.

1.2 - Background

1.2.1 - Traditional methods for protein sequence analysis

The majority of methods for protein sequence analysis utilize multiple sequence alignments, which arrange related sequences such that homologous regions are organized into columns. This is accomplished by inserting gaps throughout each sequence such that similar or identical residues appear in the same columns. Given a pair of aligned sequences, mismatches represent point mutations while gaps represent insertions or deletions. To account for (non)synonymous substitutions, alignment algorithms typically penalize mismatches using a substitution matrix which describes the likelihood of one residue mutating into another. Extending these comparisons to multiple sequences, highly conserved positions typically indicate structure or functional importance [7]. Multiple sequence alignments can also be used to model sequence evolution through a variety of phylogenetic methods [8]. In fact, many biological insights can be obtained from multiple sequence alignments alone by investigating the interdependence between sequence, structure, function, and evolution.

The goal of a sequence alignment algorithm is to define equivalent positions shared across a dataset of sequences. Due to the combinatorial space of all possible alignments, this is a computationally difficult problem that requires heuristic algorithms for approximate solutions. Programs such as MAFFT [9], MUSCLE [10], and ClustalO [11] can generate multiple sequence alignments from small datasets of unaligned protein sequences, scaling poorly to larger sequence datasets. Conversely, profile-based methods such as HMMER [12] and MAPGAPS [13] utilize different heuristics that are better suited for aligning large sequence datasets. By aligning a small, representative sample of a large sequence dataset, the resulting alignment can be converted into a

profile which is used as a heuristic for aligning all other sequences in the dataset. Given a database of profiles corresponding to diverse protein superfamilies or families, profile-based methods can also be used to classify sequences based on how well they align to each individual profile. The quality of multiple sequence alignments can also be improved by incorporating additional structural information, implemented by programs such as DALI [14] and mTM-align [15]. Because protein structure evolves at a significantly slower rate than protein sequence, structural homology is highly informative of equivalent sequence regions [16]. This heuristic is particularly useful for aligning highly divergent sequences.

A common method for visually analyzing a multiple sequence alignment is using sequence logos which show the conservation of each aligned column, typically quantified by statistical entropy [17]. A meaningful alignment of any protein family will reveal that some columns are more conserved than others. Highly conserved columns are predictive of functionally important residues which may be involved with catalysis or substrate binding [7]. Furthermore, columns that are uniquely conserved by a subset of sequences are predictive of residues that are important for family-specific functions [18–20]. Overall, the individual characterization of each column represents first-order statistics.

Second-order statistics describe coevolutionary relationships between alignment columns — measuring the degree to which variations in any two columns are correlated within the same sequence. Strongly coevolving pairs are highly predictive of structural contacts because interacting residues are under evolutionary pressure to remain mutually compatible. Although this can be computationally expensive to determine, second-order statistics are powerful in that they infer protein structure from sequence information alone. There are many methods for quantifying second-order statistics [21–23]; however, the most widely used method is direct

coupling analysis which uniquely disentangles direct correlations from indirect correlations [24]. For instance, if A is directly correlated with B and B is directly correlated with C, then A would be indirectly correlated with C. Additional statistics have also been developed to correct for random noise and sampling bias resulting from common ancestry, which offers further improvements to the method [25]. Taking advantage of recent advances in high-performance computing, preliminary studies have shown that the algorithm for direct coupling analysis can be further expanded to model third-order statistics [26].

Multiple sequence alignments can also be used to infer phylogenetic trees which describe the evolutionary relationships between sequences. Statistical methods for determining phylogenies such as maximum likelihood [27] and Bayesian inference [28] attempt to identify the most likely explanation for how the currently observed (present day) sequences may have emerged from a common ancestor. This inference depends on a variety of parameters such as the underlying multiple sequence alignment, which defines homologous positions in each sequence, the evolutionary rates, which define how quickly each column mutates, and the substitution model, which defines the probability of all possible point mutations. Resampling methods such as bootstrap are typically used to measure branch support values which indicate how frequently each clade was observed across replicate trees [29]. By inferring the progression of evolutionary history, phylogenetic trees provide a meaningful organization for related protein sequences — closely-related sequences tend to share more structure-functional similarities relative to distantly-related sequences.

Although the majority of sequence analysis methods depend on multiple sequence alignments, several alignment-independent methods for sequence analysis have also been developed [30]. One of the most common strategies utilizes word-based methods, which divide

sequences into subsequences of a constant length. Implemented by programs such as CD-HIT [31], word-based methods represent protein sequences by the frequency in which each word appears in the protein sequence. These methods are robust to domain shuffling and are generally more computationally efficient compared to alignment-based methods. However, options for conducting in-depth alignment-independent analyses are limited compared to the relative abundance of alignment-based methods due to the technical challenges in developing them.

1.2.2 - Applying artificial neural networks in protein sequence analysis

Artificial neural networks are a predictive model inspired by the communication system of neurons within the brain consisting of layers of interconnected nodes. Within an artificial neural network, nodes are linked by directed connections which are associated with weights that determine how strongly each source node influences each target node. Each node is also associated with an activation function which is used to calculate a value that gets outputted to connected nodes. Nonlinear activation functions allow artificial neural networks to model complex relationships [32]. To apply this basic framework for predictive modeling, input values are assigned to input nodes which are connected to a network of intermediate nodes (also called hidden states) and/or output nodes located downstream. These input values are used to calculate output values (as well as all other intermediate nodes) through forward calculations that propagate through each layer of nodes in the network. Given a set of input values, output values are influenced by the connection weights, which can be fitted to a training dataset using the backpropagation algorithm [33,34]. Using these basic principles, artificial neural networks have been applied towards the prediction of various biological properties [35] such as protein structure [36,37], cellular localization [38,39], secondary structure [40,41], and mutation impact [42,43].

In order for an artificial neural network to parse protein sequence information, the amino acid sequence must be converted into a numeric representation [44]. One-hot encoding is the most straightforward method for deriving a numeric representation for discrete data by using a binary variable to represent each discrete character. In order to represent a protein sequence, each residue would need to be represented as an individual one-hot encoding. However, a major disadvantage is that one-hot encodings are relatively uninformative because they do not reflect similarities between amino acids [45]. To account for these similarities, amino acids can also be encoded based on physical-biochemical features such as mass, charge, and hydrophobicity [46]. Another popular strategy is to encode amino acids based on their corresponding column in a substitution matrix [44]. Although these methods encode amino acid similarities, they do not account for sequence context — each of the 20 amino acids is always represented using the same 20 unique vectors.

There are various strategies for encoding protein sequences context. Instead of individually encoding each residue, ProtVec encodes local sequence context using learned encoding for three-residue chunks within the sequence [47], similar to the aforementioned word-based methods for alignment-free protein sequence analysis. Context-naive encodings can also be trained to include contextual information by masked language modeling [48]. This technique is commonly associated with Transformer models [49], which can produce highly nuanced protein sequences encodings (also referred in literature as representations or embedding vectors) that account for the full sequence context [4,5]. These embeddings encode a wide range of biological properties and are very useful as machine learning features for predicting a diverse range of target variables.

Many artificial neural networks utilize aligned protein sequences as input because it pre-establishes a common frame of reference for sequence positions with a fixed length. Input nodes corresponding to a given position can assume that they will always receive an equivalent site for every sequence input. Operating without a common frame of reference, modeling unaligned sequences is a major challenge because it requires the artificial neural network to parse sequences of variable length and identify their features without external guidance. This would require an architecture capable of recognizing important sequence regions irrespective of their position in an unaligned sequence. Offering one potential strategy, convolutional neural networks [50] with pooling operations are capable of establishing positional invariance and are commonly used for applications in object detection [51]. Performance for detecting long-range interactions is highly dependent on the convolution size [52,53]. Offering another potential strategy, recurrent neural networks with long or short-term memory storage [54] can read sequences one position at a time and are capable of remembering important features such as the presence of sequence motifs. However, recurrent networks are prone to forgetting information over time [55]. Sharing a common disadvantage, both architectures struggle to model long-range interactions or long sequences in general due to the requirement of deeper networks which are prone to gradient vanishing and require more computational resources [56,57].

Utilizing a different strategy, Transformer models implement an attention mechanism that provides direct connections between all positions of an input — irrespective of distance — while positional information is maintained through positional encodings [49]. Furthermore, Transformers have proven to be highly effective for modeling variable-length inputs and long-range interactions [4,5,48,58]. Consequently, Transformers and attention-based architectures have recently become a popular platform for bioinformatics applications [59].

1.3 - Key challenges and unresolved questions

Since their original conception in the late 1900s, multiple sequence alignments have been established as a core framework in protein sequence analysis. Consequently, the majority of protein sequence research operates on a decades-old paradigm where data analysis pipelines are typically performed by chaining together a series of single-function programs in the UNIX terminal. This paradigm presents a number of common issues such as the requirement of managing program-specific file formats, the accumulation of excessive intermediate files, and difficulties in reproducibility.

Alignment-based methods are very effective in modeling conserved sequence regions. However, they struggle to characterize fast-evolving or divergent regions which are difficult to align. Aligning distantly-related proteins is a major challenge due to difficulties in reliably detecting homologous regions across divergent sequences. Previous work has defined ~25% sequence identity as the “twilight zone” where true homologs sharing similar structural folds become indistinguishable from random unrelated sequences [60]. Finally, most alignment-based analyses work under the assumption that the underlying alignment is correct — statistics calculated from an unreliable multiple sequence alignment are likely to yield meaningless or misleading results.

While multiple sequence alignments are highly useful for protein sequence analysis, there is a scarcity of modern computational frameworks for doing so. Furthermore, the development of orthogonal methods is also important as a means of external validation. However, there is also a scarcity of orthogonal methods for alignment-independent sequence analysis due to the technical challenges of developing these methods.

1.4 - Major research questions addressed

To address the aforementioned challenges, I have built an array-based framework for alignment-based sequence analysis which facilitates a programmatic interface with the larger scientific Python ecosystem, a centralized platform used by a diverse range of quantitative fields [6,61]. These tools were applied in large-scale comparative sequence analyses of the protein kinases. First, I characterize sequence variations in the αC-β4 loop region as it relates to protein kinase structure, function, evolution, and disease [62]. Next, I characterize the functional evolution of diverse tyrosine kinase families across holozoan organisms — animals and their closest single-celled relatives. Extending beyond the current human-centric evolutionary model [63], I inferred a new phylogeny of the tyrosine kinome, representative across all holozoan organisms [64]. Finally, I developed a methodological framework for alignment-independent sequence analysis using embedding vectors generated from Transformer protein language models. I applied these methods towards inferring a rooted phylogeny of the protein kinase superfamily which has been a major challenge for alignment-based methods due to the difficulties in reliably aligning the protein kinase superfamily to an evolutionary outgroup of small molecule and lipid kinases. To demonstrate the broader applicability of my embedding-based approach, I also apply these methods towards studying phosphatase enzymes and the radical S-adenosyl-L-methionine (SAM) enzyme superfamily.

1.4.1 - Emerging roles of the αC-β4 loop in protein kinase structure, function, evolution, and disease

The αC-β4 loop refers to a flexible region of the protein kinase domain located between the αC helix and β4 strand. Previous research has shown that the αC-β4 loop of the eukaryotic

protein kinase superfamily is distinct from distantly related eukaryotic-like and atypical kinases [65]. Furthermore, variations in the αC-β4 loop seem to modulate αC-helix dynamics which play a major role in regulating kinase activity. In this study, I described sequence variations of the αC-β4 loop region across diverse eukaryotic protein kinases. The majority of eukaryotic protein kinases conserve an HxN motif at the αC-β4 loop, however this motif is absent among CK1 kinases and divergent among AGC kinases which instead conserve an HPF motif at the equivalent position. While the αC-β4 loop is typically eight residues long, extended loops can be observed across the protein kinome where the insert typically results in an additional helical segment. Finally, I map cancer mutations and post-translational modifications and identify the αC-β4 loop as a hotspot for cancer mutations, especially in the tyrosine kinase group. Using a wide variety of comparative analyses, this study identifies the αC-β4 loop as a central hub for many important protein kinase regulatory mechanisms.

1.4.2 - Evolution of functional diversity in the holozoan tyrosine kinome

Tyrosine kinases are a major group of protein kinases, the expansion of which has been strongly associated with the evolution of metazoan multicellularity [66]. In this study, I propose a new representative phylogeny of the tyrosine kinome using sequences from diverse holozoan organisms. The holozoan tyrosine kinase can be divided into two major clades one mainly consisting of cytoplasmic tyrosine kinases and the other mainly consisting of receptor tyrosine kinases. Nearly all members of the latter clade conserve a fast-evolving insertion region between the αD and αE-helices. The analysis also identifies three major subgroups of tyrosine kinases which conserve subgroup-specific sequence motifs which reflect subgroup-specific kinase regulatory mechanisms. Based on the taxonomic conservation of tyrosine kinase families, I speculate as to how evolutionary innovations of anciently conserved tyrosine kinases could have

led to the emergence of multicellularity. Furthermore, I note that the emergence of new biological phenotypes often coincided with the emergence of functionally-related tyrosine kinase families. Further solidifying the connection between tyrosine kinases and metazoan evolution, I end by noting an interesting trend that more recent tyrosine kinases families tend to be overrepresented in cancer mutations.

1.4.3 - Alignment-free evolutionary analysis and sequence classification using Transformer protein language models

Generated from Transformer protein language models [49], protein sequence embeddings are numerical matrix representations of amino acid sequences [4,5] and are typically used as input features for supervised machine learning applications [67]. In this study, I propose various unsupervised methods for directly analyzing protein sequence embeddings, without the requirement of labeled data. Demonstrating unique technical advantages, I find that embedding vectors are highly amenable for applications in alignment-free sequence analysis. I develop methods for embedding-based evolutionary inference, sequence classification, and fast/slow-evolving sites identification, then demonstrate the broad applicability of these methods across three diverse protein sequence datasets. First, I inferred an embedding-based phylogeny of the protein kinase superfamily, rooted to an outgroup of structure-functionally divergent small molecule and lipid kinases. The tree indicated that the Casein Kinase 1 (CK1) is the most ancestral protein kinase group. Next, I conducted an embedding-based hierarchical clustering of phosphatase enzymes which span ten different structural folds. Clustering results yielded a biologically-meaningful organization which reflected convergent motifs shared between divergent structural folds. Finally, I inferred a representative phylogenetic tree of the radical SAM superfamily which has been challenging to align due to the high degree of structural

variations between families. Overall, I propose embedding-based sequence analysis as a new class of techniques for alignment-free sequence analysis — particularly useful for studying divergent protein superfamilies.

Bibliography

[1] The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research. 49 (2021) D480–D489. https://doi.org/10.1093/nar/gkaa1100.

[2] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J Mol Biol. 48 (1970) 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.

[3] D.F. Feng, R.F. Doolittle, Progressive sequence alignment as a prerequisite to correct phylogenetic trees, J Mol Evol. 25 (1987) 351–360. https://doi.org/10.1007/BF02603120.

[4] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, PNAS. 118 (2021). https://doi.org/10.1073/pnas.2016239118.

[5] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing, IEEE Transactions on Pattern Analysis and Machine Intelligence. (2021) 1–1. https://doi.org/10.1109/TPAMI.2021.3095381.

[6] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods. 17 (2020) 261–272. https://doi.org/10.1038/s41592-019-0686-2.

[7] J.A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics. 23 (2007) 1875–1882. https://doi.org/10.1093/bioinformatics/btm270.

[8] Z. Yang, B. Rannala, Molecular phylogenetics: principles and practice, Nat Rev Genet. 13 (2012) 303–314. https://doi.org/10.1038/nrg3186.

[9] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research. 30 (2002) 3059–3066. https://doi.org/10.1093/nar/gkf436.

[10] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics. 5 (2004) 113. https://doi.org/10.1186/1471-2105-5-113.

[11] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol Syst Biol. 7 (2011) 539. https://doi.org/10.1038/msb.2011.75.

[12] S.R. Eddy, Accelerated Profile HMM Searches, PLOS Computational Biology. 7 (2011) e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

[13] A.F. Neuwald, Rapid detection, classification and accurate alignment of up to a million or more related protein sequences, Bioinformatics. 25 (2009) 1869–1875. https://doi.org/10.1093/bioinformatics/btp342.

[14] L. Holm, DALI and the persistence of protein shape, Protein Science. 29 (2020) 128–140. https://doi.org/10.1002/pro.3749.

[15] R. Dong, Z. Peng, Y. Zhang, J. Yang, mTM-align: an algorithm for fast and accurate multiple protein structure alignment, Bioinformatics. 34 (2018) 1719–1725. https://doi.org/10.1093/bioinformatics/btx828.

[16] K. Illergård, D.H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence—A study of structural response in protein cores, Proteins: Structure, Function, and Bioinformatics. 77 (2009) 499–508. https://doi.org/10.1002/prot.22458.

[17] T.D. Schneider, R.M. Stephens, Sequence logos: a new way to display consensus sequences, Nucleic Acids Research. 18 (1990) 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

[18] A.F. Neuwald, A Bayesian Sampler for Optimization of Protein Domain Hierarchies, Journal of Computational Biology. 21 (2014) 269–286. https://doi.org/10.1089/cmb.2013.0099.

[19] A.F. Neuwald, The CHAIN program: forging evolutionary links to underlying mechanisms, Trends in Biochemical Sciences. 32 (2007) 487–493. https://doi.org/10.1016/j.tibs.2007.08.009.

[20] A.F. Neuwald, Surveying the Manifold Divergence of an Entire Protein Class for Statistical Clues to Underlying Biochemical Mechanisms, Statistical Applications in Genetics and Molecular Biology. 10 (2011). https://doi.org/10.2202/1544-6115.1666.

[21] S.M. Larson, A.A. Di Nardo, A.R. Davidson, Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions11Edited by F. E. Cohen, Journal of Molecular Biology. 303 (2000) 433–446. https://doi.org/10.1006/jmbi.2000.4146.

[22] S.W. Lockless, R. Ranganathan, Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families, Science. 286 (1999) 295–299. https://doi.org/10.1126/science.286.5438.295.

[23] D.K.Y. Chiu, T. Kolodziejczak, Inferring consensus structure from nucleic acid sequences, Bioinformatics. 7 (1991) 347–352. https://doi.org/10.1093/bioinformatics/7.3.347.

[24] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E. 87 (2013) 012707. https://doi.org/10.1103/PhysRevE.87.012707.

[25] S.D. Dunn, L.M. Wahl, G.B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, Bioinformatics. 24 (2008) 333–340. https://doi.org/10.1093/bioinformatics/btm604.

[26] M. Schmidt, K. Hamacher, hoDCA: higher order direct-coupling analysis, BMC Bioinformatics. 19 (2018) 546. https://doi.org/10.1186/s12859-018-2583-6.

[27] J. Felsenstein, Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters, Systematic Biology. 22 (1973) 240–249. https://doi.org/10.1093/sysbio/22.3.240.

[28] J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics. 17 (2001) 754–755. https://doi.org/10.1093/bioinformatics/17.8.754.

[29] B. Efron, E. Halloran, S. Holmes, Bootstrap confidence levels for phylogenetic trees, PNAS. 93 (1996) 13429–13429. https://doi.org/10.1073/pnas.93.23.13429.

[30] A. Zielezinski, S. Vinga, J. Almeida, W.M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biology. 18 (2017) 186. https://doi.org/10.1186/s13059-017-1319-7.

[31] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics. 22 (2006) 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

[32] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control Signal Systems. 2 (1989) 303–314. https://doi.org/10.1007/BF02551274.

[33] H.J. Kelley, Gradient Theory of Optimal Flight Paths, ARS Journal. 30 (1960) 947–954. https://doi.org/10.2514/8.5282.

[34] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature. 323 (1986) 533–536. https://doi.org/10.1038/323533a0.

[35] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Briefings in Bioinformatics. 18 (2017) 851–869. https://doi.org/10.1093/bib/bbw068.

[36] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back,

S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[37] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, J. Wang, Q. Cong, L.N. Kinch, R.D. Schaeffer, C. Millán, H. Park, C. Adams, C.R. Glassman, A. DeGiovanni, J.H. Pereira, A.V. Rodrigues, A.A. van Dijk, A.C. Ebrecht, D.J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M.K. Rathinaswamy, U. Dalwadi, C.K. Yip, J.E. Burke, K.C. Garcia, N.V. Grishin, P.D. Adams, R.J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network, Science. 373 (2021) 871–876. https://doi.org/10.1126/science.abj8754.

[38] Y.-D. Cai, X.-J. Liu, K.-C. Chou, Artificial Neural Network Model for Predicting Membrane Protein Types, Journal of Biomolecular Structure and Dynamics. 18 (2001) 607–610. https://doi.org/10.1080/07391102.2001.10506692.

[39] F. Teufel, J.J. Almagro Armenteros, A.R. Johansen, M.H. Gíslason, S.I. Pihl, K.D. Tsirigos, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 6.0 predicts all five types of signal peptides using protein language models, Nat Biotechnol. (2022) 1–3. https://doi.org/10.1038/s41587-021-01156-3.

[40] J.-M. Chandonia, M. Karplus, Neural networks for secondary structure and structural class predictions, Protein Science. 4 (1995) 275–285. https://doi.org/10.1002/pro.5560040214.

[41] S. Wang, J. Peng, J. Ma, J. Xu, Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields, Sci Rep. 6 (2016) 18962. https://doi.org/10.1038/srep18962.

[42] L. Sundaram, H. Gao, S.R. Padigepati, J.F. McRae, Y. Li, J.A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, K.K.-H. Farh, Predicting the clinical impact of human mutation with deep neural networks, Nat Genet. 50 (2018) 1161–1170. https://doi.org/10.1038/s41588-018-0167-z.

[43] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, (2021) 2021.07.09.450648. https://doi.org/10.1101/2021.07.09.450648.

[44] V.I. Jurtz, A.R. Johansen, M. Nielsen, J.J. Almagro Armenteros, H. Nielsen, C.K. Sønderby, O. Winther, S.K. Sønderby, An introduction to deep learning on biological sequence data: examples and solutions, Bioinformatics. 33 (2017) 3685–3690. https://doi.org/10.1093/bioinformatics/btx531.

[45] K. Lin, A.C.W. May, W.R. Taylor, Amino Acid Encoding Schemes from Protein Structure Alignments: Multi-dimensional Vectors to Describe Residue Types, Journal of Theoretical Biology. 216 (2002) 361–365. https://doi.org/10.1006/jtbi.2001.2512.

[46] S. Kawashima, M. Kanehisa, AAindex: Amino Acid index database, Nucleic Acids Research. 28 (2000) 374. https://doi.org/10.1093/nar/28.1.374.

[47] E. Asgari, M.R.K. Mofrad, Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics, PLOS ONE. 10 (2015) e0141287. https://doi.org/10.1371/journal.pone.0141287.

[48] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019: pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 6000–6010.

[50] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition?, in: 2009 IEEE 12th International Conference on Computer Vision, 2009: pp. 2146–2153. https://doi.org/10.1109/ICCV.2009.5459469.

[51] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM. 60 (2017) 84–90. https://doi.org/10.1145/3065386.

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

[53] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE. 86 (1998) 2278–2324. https://doi.org/10.1109/5.726791.

[54] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[55] N.Y. Masse, G.D. Grant, D.J. Freedman, Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization, Proc Natl Acad Sci U S A. 115 (2018) E10467–E10475. https://doi.org/10.1073/pnas.1803839115.

[56] J.F. Kolen, S.C. Kremer, Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies, in: A Field Guide to Dynamical Recurrent Networks, IEEE, 2001: pp. 237–243. https://doi.org/10.1109/9780470544037.ch14.

[57] G.B. Goh, N.O. Hodas, A. Vishnu, Deep learning for computational chemistry, J Comput Chem. 38 (2017) 1291–1307. https://doi.org/10.1002/jcc.24764.

[58] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, A. Rives, Transformer protein language models are unsupervised structure learners, (2020) 2020.12.15.422761. https://doi.org/10.1101/2020.12.15.422761.

[59] H. Li, S. Tian, Y. Li, Q. Fang, R. Tan, Y. Pan, C. Huang, Y. Xu, X. Gao, Modern deep learning in bioinformatics, Journal of Molecular Cell Biology. 12 (2020) 823–827. https://doi.org/10.1093/jmcb/mjaa030.

[60] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins., The EMBO Journal. 5 (1986) 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.

[61] S. van der Walt, S.C. Colbert, G. Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science Engineering. 13 (2011) 22–30. https://doi.org/10.1109/MCSE.2011.37.

[62] W. Yeung, Z. Ruan, N. Kannan, Emerging roles of the αC-β4 loop in protein kinase structure, function, evolution, and disease, IUBMB Life. 72 (2020) 1189–1202. https://doi.org/10.1002/iub.2253.

[63] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The Protein Kinase Complement of the Human Genome, Science. 298 (2002) 1912–1934. https://doi.org/10.1126/science.1075762.

[64] W. Yeung, A. Kwon, R. Taujale, C. Bunn, A. Venkat, N. Kannan, Evolution of Functional Diversity in the Holozoan Tyrosine Kinome, Molecular Biology and Evolution. 38 (2021) 5625–5639. https://doi.org/10.1093/molbev/msab272.

[65] N. Kannan, A.F. Neuwald, Did Protein Kinase Regulatory Mechanisms Evolve Through Elaboration of a Simple Structural Component?, Journal of Molecular Biology. 351 (2005) 956–972. https://doi.org/10.1016/j.jmb.2005.06.057.

[66] N. King, M.J. Westbrook, S.L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K.J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J.B. Lyons, A. Morris, S. Nichols, D.J. Richter, A. Salamov, J.G.I. Sequencing, P. Bork, W.A. Lim, G. Manning, W.T. Miller, W. McGinnis, H. Shapiro, R. Tjian, I.V. Grigoriev, D. Rokhsar, The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans, Nature. 451 (2008) 783–788. https://doi.org/10.1038/nature06617.

[67] C. Dallago, K. Schütze, M. Heinzinger, T. Olenyi, M. Littmann, A.X. Lu, K.K. Yang, S. Min, S. Yoon, J.T. Morton, B. Rost, Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets, Current Protocols. 1 (2021) e113. https://doi.org/10.1002/cpz1.113.

Chapter 2

EMERING ROLES OF THE αC-β4 LOOP IN PROTEIN KINASE STRUCTURE, FUNCTION, EVOLUTION, AND DISEASE

W. Yeung, Z. Ruan, N. Kannan, Emerging roles of the αC-β4 loop in protein kinase structure, function, evolution, and disease, IUBMB Life. 72 (2020) 1189–1202.

Reprinted here with permission of publisher.

Abstract

The faithful propagation of cellular signals in most organisms relies on the coordinated functions of a large family of protein kinases that share a conserved catalytic domain. The catalytic domain is a dynamic scaffold that undergoes large conformational changes upon activation. Most of these conformational changes, such as movement of the regulatory αC-helix from an "out" to "in" conformation, hinge on a conserved, but understudied, loop termed the αC-β4 loop, which mediates conserved interactions to tether flexible structural elements to the kinase core. We previously showed that the αC-β4 loop is a unique feature of eukaryotic protein kinases. Here, we review the emerging roles of this loop in kinase structure, function, regulation, and diseases. Through a kinome-wide analysis, we define the boundaries of the loop for the first time and show that sequence and structural variation in the loop correlate with conformational and regulatory variation. Many recurrent disease mutations map to the αC-β4 loop and contribute to drug resistance and abnormal kinase activation by relieving key auto-inhibitory interactions associated with αC-helix and inter-lobe movement. The αC-β4 loop is a hotspot for post-translational modifications, protein-protein interaction, and Hsp90 mediated folding. Our kinome-wide analysis provides insights for hypothesis-driven characterization of understudied kinases and the development of allosteric protein kinase inhibitors.

2.1 Introduction

Protein kinases are one of the largest gene families in the human genome and regulate virtually all cellular processes. Dysregulation of protein kinase activity can lead to a variety of disease phenotypes such as cancer [1], diabetes [2], neurodegeneration [3], and cardiovascular disease [4]. Consequently, there is a need to understand the diverse regulatory mechanisms of protein kinases as a foundation for developing protein kinase inhibitors. To this end, comparative studies on protein kinase sequence and structure have provided important insights into protein kinase activation, regulation, evolution, and inhibition [5–7].

Drug discovery efforts on protein kinases have traditionally focused on the conserved catalytic domain, which adopts a bi-lobal fold. The N-terminal ATP binding lobe consists of 5 strands and a helix, while the larger C-terminal substrate binding lobe is primarily composed of helices. Extensive structural studies on the catalytic domain and comparisons of active and inactive conformations have highlighted the role of key flexible elements in kinase conformational regulation. The activation segment [8,9] and αC-helix [10,11] are two such flexible elements that undergo dramatic conformational changes upon activation of most protein kinases [12]. Another critical example of a flexible element is the dynamic assembly of the regulatory spine (RS) [11,13], a spatially connected network of hydrophobic interactions spanning the ATP and substrate binding lobes. RS assembly is correlated with kinase activation and conformational strain in the catalytic loop [14].

At the advent of the post-genomic era, the newfound wealth of sequencing information allowed large-scale comparisons across diverse protein kinases. In particular, quantitative comparisons of the evolutionary constraints acting on eukaryotic and distantly related eukaryotic

like kinases in prokaryotes revealed that conformational flexibility and allosteric regulation evolved progressively in protein kinases through addition of key flexible elements such as the activation loop [5]. In addition to the activation loop and the substrate binding lobe, sequence motifs in the αC-β4 loop were also identified as unique to eukaryotic protein kinases [5]. Structurally, the αC-β4 loop immediately follows the αC helix and connects to the β8 strand, which immediately precedes the activation loop DFG motif. The αC-β4 loop also serves as a hinge point for inter-lobe movement.

The αC-β4 loop resides at the intersection of many essential regulatory mechanisms for protein kinase function. Disease-related mutations in this region are capable of altering kinase activity and drug response [15–17]. While our knowledge of the kinase activation loop is quite extensive, relatively little is known about the role of the αC-β4 loop in kinase function. In this review, we aim to provide a comprehensive review on the αC-β4 loop region and centralize the knowledge to facilitate comparisons across the protein kinome.

2.2 Results and Discussion

2.2.1 Conservation and variation in the αC-β4 loop of protein kinases

Structure and sequence conservation of αC-β4 loop

The αC-β4 loop is located on the N-lobe of the kinase domain and connects the αC-helix to the β4 strand **(Figure 2.1A)**. To provide an unbiased overview on the kinase αC-β4 loop, we mined the Protein Data Bank (PDB) for kinase structures solved by X-ray crystallography (4900 structures, 8122 chains). We generated a non-redundant dataset of kinase structures by only including kinase chains with unique Uniprot IDs. During this filtering procedure, priority was

given to structures with high resolution and fully resolved αC-β4 loops. The final filtered dataset contained 426 kinase chains and was used for all subsequent structural analyses.

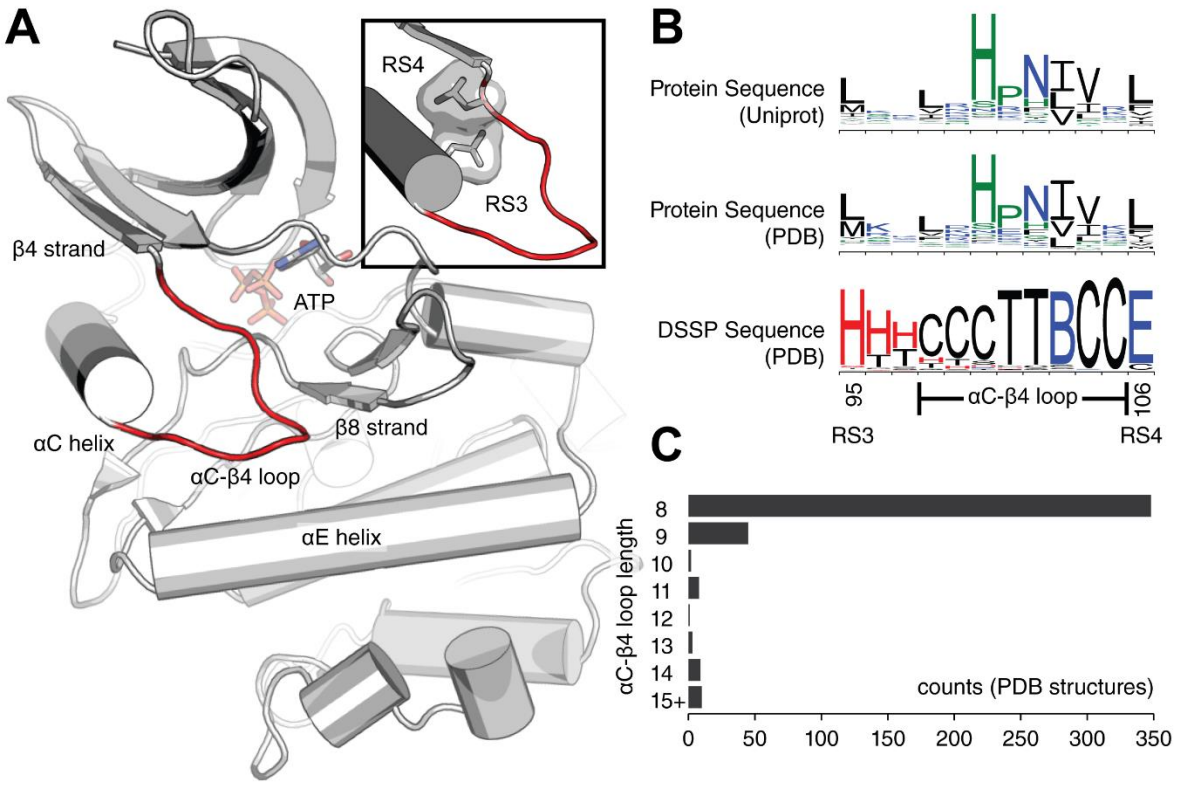


Figure 2.1: Definition of the αC-β4 loop
(A) The αC-β4 loop (red) of protein kinase A (PDB ID: 1ATP) [18]. Structural regions near the αC-β4 loop are labeled for reference. At the top-right corner, a close-up shows the αC-β4 loop flanked by the RS3 and RS4 residues. (B) Sequence logo plots spanning from RS3 to RS4 are shown. This span of residues was used on all logo plots throughout the review. Sequence logo plots include amino acid sequences from Uniprot proteomes (top), amino acid sequences from PDB (middle), and secondary structure sequences from PDB (bottom). Secondary structure sequences were defined by DSSP where red = helix, blue = strand, black = coil. DSSP classifications: G = 3_{10} helix, H = α-helix, I = π-helix, B = isolated β-bridge, E = extended β-strand, T = hydrogen bonded turn, S = non-hydrogen bonded bend, C = coil. (C) Histogram showing the distribution of αC-β4 loop lengths calculated from unique protein kinase structures in the PDB. The 15+ category includes lengths greater than or equal to 15.

To analyze the sequence conservation of the αC-β4 loop, we identified and aligned 600,734 protein kinase sequences from Uniprot proteomes [19]. We also converted our filtered dataset of 426 kinase chains into amino acid sequences. Amino acid sequence logos from Uniprot **(Figure 2.1B, top)** and PDB **(Figure 2.1B, middle)** are similar, suggesting that our filtered dataset of kinase structures is representative.

To define the boundaries of the αC-β4 loop, we assigned a secondary structure sequence to each kinase structure using the Define Secondary Structure of Proteins (DSSP) algorithm [20] **(Figure 2.1B)**. Based on secondary structure propensity, we defined the αC-β4 as an 8-residue segment starting from RS3+3 (PKA position 98) and ending at RS4-1 (PKA position 105). While the αC-β4 loop is typically 8 residues in length **(Figure 2.1C)**, we note exceptions in multiple kinase structures **(Table 2.1)**.

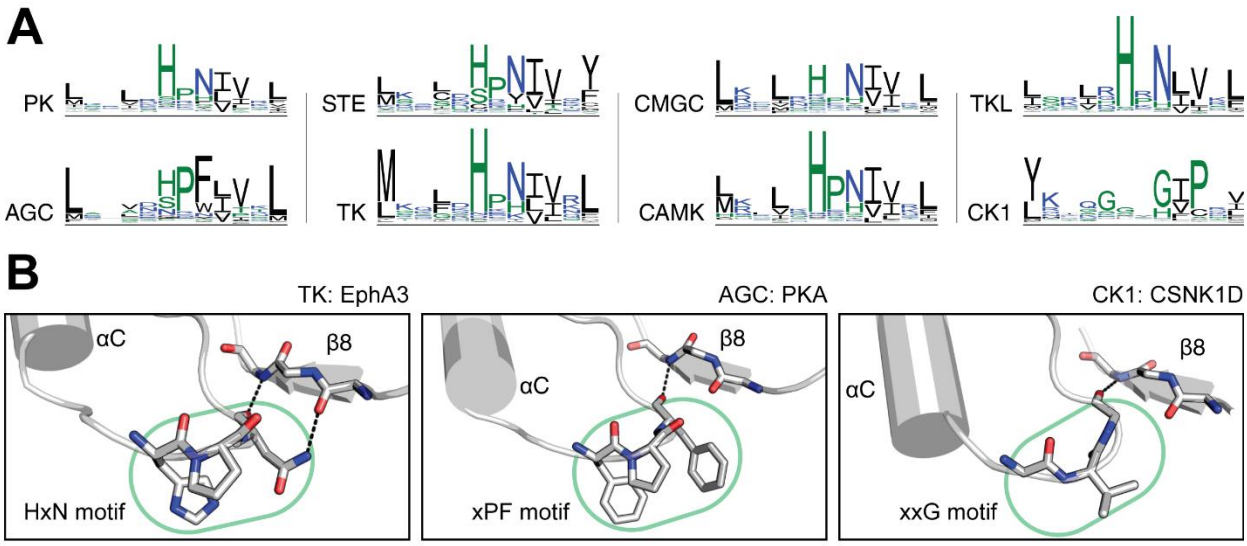


Figure 2.2: Sequence conservation of the αC-β4 loop
(A) Sequence logo plots spanning from RS3 to RS4 are shown for different protein kinase groups. (B) The canonical HxN motif is shown in tyrosine kinase EphA3 (PDB ID: 3dzq) (left). The AGC-specific variant is shown in PKA (PDB ID: 1atp) [18] (middle).

The HxN motif

The αC-β4 loop and the associated HxN motif is uniquely conserved in eukaryotic protein kinases including pseudokinases [21], but is absent/divergent in distantly related atypical protein kinases and eukaryotic-like small molecule kinases [5]. To investigate sequence variations in the αC-β4 loop, we analyzed sequence conservation within each kinase group using the aforementioned Uniprot dataset of 600,734 protein kinase sequences **(Figure 2.2A)**. This analysis revealed a consensus sequence for the αC-β4 loop: Φ-X-H-X-N-Φ-Φ-X **(Figure 2.2A, top-left)**, where Φ represents a hydrophobic residue and X represents any amino acid (wildcard). The HxN motif facilitates a β-turn connecting the αC helix and β4 strand **(Figure 2.2B, left)**. The HxN-Asn hydrogen bonds the backbone of the β8 strand via an isolated β-bridge and the carboxamide side chain **(Figure 2.3A)**. These interactions tether the αC-β4 loop to the hinge region of the protein kinase domain. The HxN wildcard residue is usually a proline and shows varying levels of conservation across different kinase groups.

The CK1-specific xxG motif is shown in CSNK1D (PDB ID: 4twc) [22] (right). Residue numbers (not shown) are provided: 679-681 for the HxN motif in EphA3, 100-102 for xPF motif in PKA, and 62-64 for the xxG motif in CSNK1D. Side chains are not shown for the β8 strand.

AGC and CK1 kinases display group-specific variations within the HxN motif to accommodate unique regulatory functions **(Figure 2.2A)**. The AGC-specific xPF motif **(Figure 2.2B, middle)** facilitates *cis*-interactions with the C-terminal tail and is hypothesized to modulate ATP binding and inter-lobe movement [23,24]. However, the CK1-specific xxG motif **(Figure 2.2B, right)** is not well understood. Similar to the HxN-Asn, the xPF-Phe and xxF-Gly form isolated β-bridges with the β8 strand **(Figure 2.2B)**.

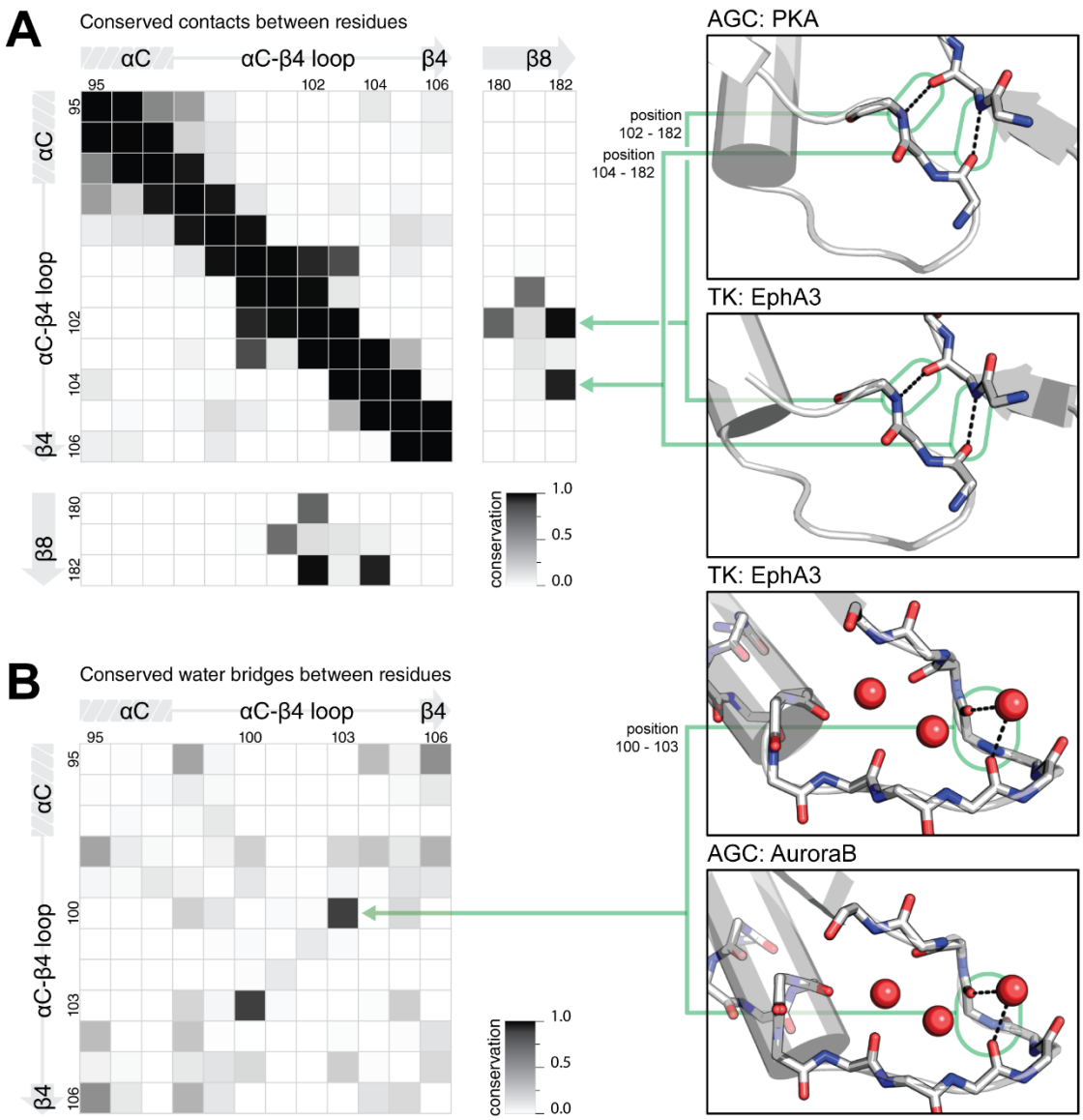


Figure 2.3: Conserved interactions within the αC-β4 loop
Conservation was defined by the fraction of structures containing the interaction. All heatmaps use PKA numbering for residue positions. (A) Conserved contacts found in kinase structures are shown in an all-vs-all comparison of residues in the αC-β4 loop and β8 strand. Two examples of the two most highly conserved long-range contacts in the αC-β4 loop and β8 strand are shown in PKA (PDB ID: 1atp) [18] and EphA3 (PDB ID: 3dzq). (B) Conserved water bridges found in kinase structures are shown in an all-vs-all comparison of residues in the αC-β4 loop. Two examples of a highly conserved water bridge is shown in EphA3 (PDB ID: 3dzq) and AuroraB (PDB ID: 2vrx) [25].

Conserved interactions involving the αC-β4 loop

The αC-β4 loop mediates many conserved structural interactions [23,24]. We independently quantified these interactions using the aforementioned dataset of 426 representative kinase chains with unique Uniprot IDs. To identify conserved contacts, we performed an all-vs-all residue comparison for residue pairs within 3.2 Å (heavy atom distance) **(Figure 2.3A, left)**. This cutoff was chosen to be greater than hydrogen bond distance and less than the van der Waals contact distance [26]. This should account for uncertainty in electron density mapping while excluding hydrophobic packing interactions. Residue pairs within this cutoff are expected to either be covalently linked or hydrogen bonded.

Our analysis identified conserved hydrogen bonds involving the αC-β4 loop **(Figure 2.3A, left)**. When considering interactions between all possible residue pairs within the αC-β4 loop, we identified a single conserved β-turn between position 100 (HxN-His) and 103 (HxN+1) **(Figure 2.3A, left)**. By extending our analysis to the entire kinase domain, we further identified conserved contacts between the αC-β4 loop and the β8 strand. On the β8 strand, 182 forms two highly conserved isolated β-bridges with 102 and 104 on the αC-β4 loop **(Figure 2.3A, right)**. The isolated β-bridge with 102 (HxN-Asn) was described in the previous section **(Figure 2.2B)**. These isolated β-bridge are observed in CK1 and AGC kinases, despite the absence of the HxN motif. Furthermore, many kinases with extended αC-β4 loops also maintain an isolated β-bridge with the β8 strand. A lesser conserved contact is detected between 102 and 180. This is only conserved amongst HxN containing kinases and reflects hydrogen bonds allowed by the carboxamide side chain of the HxN-Asn **(Figure 2.2B, left)**.

Our analysis also identifies conserved water bridges in the αC-β4 loop **(Figure 2.3B, left)**. Water bridges were defined as two residues residing within 3.2 Å (heavy atom distance) of a shared water molecule. To identify conserved water bridges, we perform an all-vs-all residue comparison for residue pairs within the αC-β4 residues. Crystal structures that lack water densities were excluded. We identified a single highly conserved water bridge connecting position 100 (HxN-His) and 103 (HxN+1) **(Figure 2.3B, right)**. This water bridge helps stabilize the conserved β-turn [27]. While our analysis did not cover water-water interactions, we note a conserved network of water molecules in some high-resolution structures.

The αC-β4 loop contains three conserved hydrophobic positions (HxN-2, HxN+1, HxN+2) which are buried within the kinase core, forming hydrophobic packing interactions with the RS. Furthermore, the HxN+2 residue takes part in a hydrophobic ensemble critical for RS assembly and thus catalytic activation [28]. In addition, recent studies suggest that conservative substitution of these hydrophobic residues can modify the shape of the active site cleft [29].

Extended αC-β4 loop conformations

Although the length of the αC-β4 is typically conserved across the protein kinase superfamily, we observed extended conformations in multiple kinase crystal structures. Extended αC-β4 loops are most commonly found in the CK1 and CMGC groups **(Figure 2.4A)** usually in the form of a short helical insert **(Figure 2.4B)**. In many cases, extended αC-β4 loops seem to be linked to constitutive enzyme activity [30–33]. **Table 2.1** shows a list of kinases containing an extended αC-β4 loop.

PDB chain	αC-β4 length	Group	Name
4jr7_A [30]	47	CMGC	*sc*CK2α (*Saccharomyces cerevisiae*)
5oat_A [34]	23	Other	PINK1 (*Tribolium castaneum*)
6bru_A	21	CK1	VRK1 (*Homo sapiens*)
2v62_A [33]	21	CK1	VRK2 (*Homo sapiens*)
2jii_A [33]	21	CK1	VRK3 (*Homo sapiens*)
1q8y_A [35]	19	CMGC	SKY1 (*Saccharomyces cerevisiae*)
4qtc_A [36]	17	Other	HASPIN (*Homo sapiens*)
5my8_A [37]	16	CMGC	SRPK1 (*Homo sapiens*)
2x7g_A	16	CMGC	SRPK2 (*Homo sapiens*)
5yk0_A [38]	15	pknB	Rv3197 (*Mycobacterium tuberculosis*)
6s14_A	14	CMGC	DYRK1A (*Homo sapiens*)
6fyv_A [39]	14	CMGC	CLK4 (*Homo sapiens*)
6fyr_A [39]	14	CMGC	CLK3 (*Homo sapiens*)
6fyl_A [39]	14	CMGC	CLK2 (*Homo sapiens*)
6ft8_A [40]	14	CMGC	CLK1 (*Homo sapiens*)
5y86_A [41]	14	CMGC	DYRK3 (*Homo sapiens*)
5lxc_A [42]	14	CMGC	DYRK2 (*Homo sapiens*)
4iir_A [43]	14	CMGC	PRPF4B (*Homo sapiens*)
3llt_A	13	CMGC	PF3D7_1445400 (*Plasmodium falciparum*)
6p5s_A [44]	13	CMGC	HIPK2 (*Homo sapiens*)
4nt4_A [45]	13	CK1	Gilgamesh (*Drosophila melanogaster*)
5wtk_A [46]	12	Other	cas13a (*Leptotrichia shahii*)
4x7q_B [47]	12	CAMK	PIM2 (*Homo sapiens*)

Table 2.1. List of kinases containing an extended αC-β4 loop
Examples were retrieved from the aforementioned dataset of 426 representative kinase chains which was filtered by unique Uniprot IDs with priority given to high resolution structures and fully resolved αC-β4 loops.

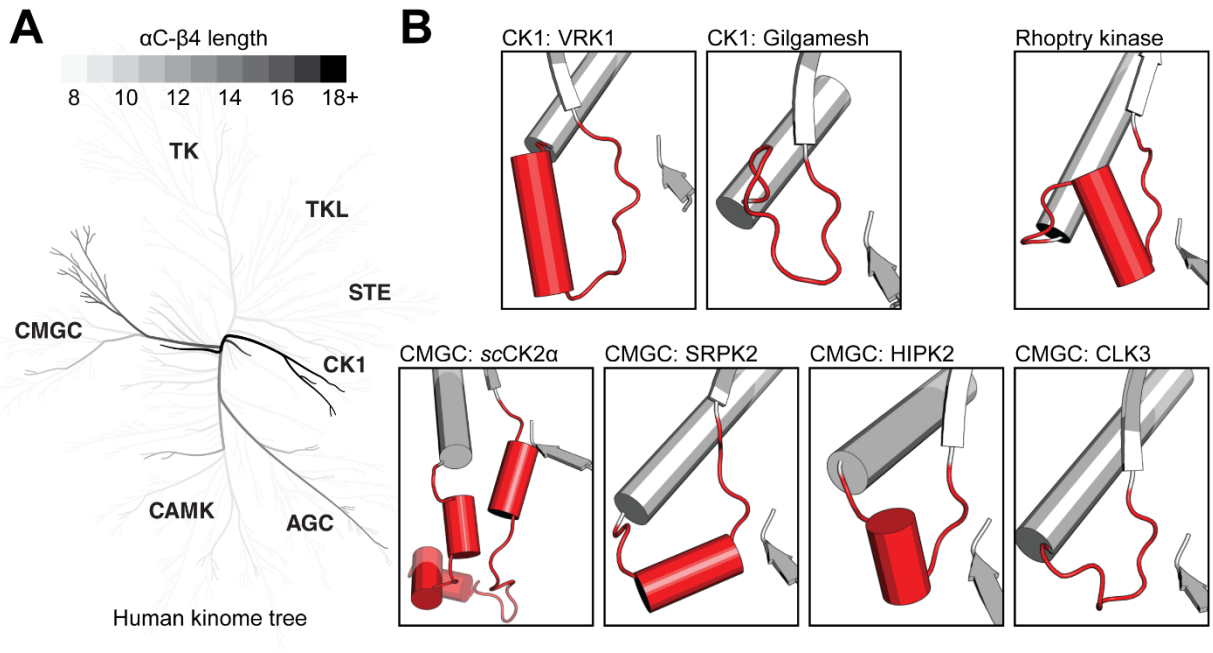


Figure 2.4: Extended conformations of the αC-β4 loop in the protein kinome
(A) A phylogenetic tree shows αC-β4 loop lengths of human protein kinases [48,49]. Branches containing longer αC-β4 loops are colored darker. The 18+ color category includes lengths greater than or equal to 18. (B) Structural examples of extended αC-β4 loops are shown in red. The β8 strand is shown for reference. Protein names are provided alongside its kinase group. Locations for αC-β4 loops are provided: 93-135 in scCKA1 (PDB ID: 4jr7) [30], 147-160 in SRPK2 (PDB ID: 2x7g), 253-262 in HIPK2 (PDB ID: 6p5s) [44], 212-222 in CLK3 (PDB ID: 6fyr) [39], 92-112 in VRK1 (PDB ID: 6bru), 95-106 in Gilgamesh (PDB ID: 4nt4) [45], and 322-340 in Rhoptry kinase (PDB ID: 3byv) [50].

In the CMGC group, *Saccharomyces cerevisiae* CK2α (*sc*CK2α) contains the longest resolved αC-β4 loop (47 residues) [30]. *sc*CK2α is a homologue of human CK2α1, a member of one of the most phylogenetically ancient CMGC kinases families. We note that the human homologue only has a 9 residue αC-β4 loop. Experimentally, *sc*CK2α has broad substrate specificity and is constitutively active [30]. A crystal structure of scCK2α reveals that the extended αC-β4 loop is tethered to the surface of the kinase C-lobe and interacts with both the N and C-terminal tails flanking the kinase domain. Deletion of the elongated segment negatively

impacts ATP binding and results in a 6-fold increase of K_m [51]. CMGC kinases SRPK2 [31] and HIPK2 [44] contain a short helical insertion in their extended αC-β4 loop **(Figure 2.4B)**. SRPK1 has been shown to maintain constitutive activity *in vitro* despite extensive mutation at the activation loop [31]. This short helical segment is not found in structures of CLK [52] and DYRK1A. However, DYRK1A maintains a similar resilience against inactivation [53,54].

In the CK1 group, the VRK family also contains an elongated αC-β4 loop with a helical insert [33]. The αC helix is tightly linked with the αE helix of the kinase domain through aromatic packing interactions, presumably rigidifying the αC helix into an active conformation [33,55]. Consequently, VRK1 and VRK2 are constitutively active, while VRK3 is a pseudokinase lacking ATP binding and phosphoryl-transfer activity [33]. A homologue of human CK1-γ, Gilgamesh kinase from *Drosophila melanogaster* also carries an extended αC-β4 loop [45]. This insertion is not observed in the sequence of the human homologue, and its function is yet to be determined.

Rhoptry kinases also contain an elongated αC-β4 loop with a helical insert [50]. The rhoptry kinases are specific to the protozoan parasite *Toxoplasma gondii* and have also been shown to be important virulence factors secreted by coccidian parasites [56]. Comparative sequence analyses identified the helical insert to be one of the most distinguishing features of rhoptry kinases [56] **(Figure 2.3)**. However, the biological role of the conserved helical insert is not well understood.

2.2.2 Disease variants in the αC-β4 loop

Many mutations in protein kinases play a direct role in cancer progression. To explore cancer-related variants in the αC-β4 loop, we retrieved missense mutations deposited in the

Catalogue of Somatic Mutations in Cancer (COSMIC) v90 [57] **(Figure 2.5A-B)**. To evenly sample all protein kinases, we only considered mutations from genome-wide screens. Furthermore, we removed redundancy caused by alternative transcripts by only including mutations with a unique tumor sample and genomic location. The resulting dataset contains a diverse combination of cancer driver mutations, passenger mutations, and drug resistance mutations.

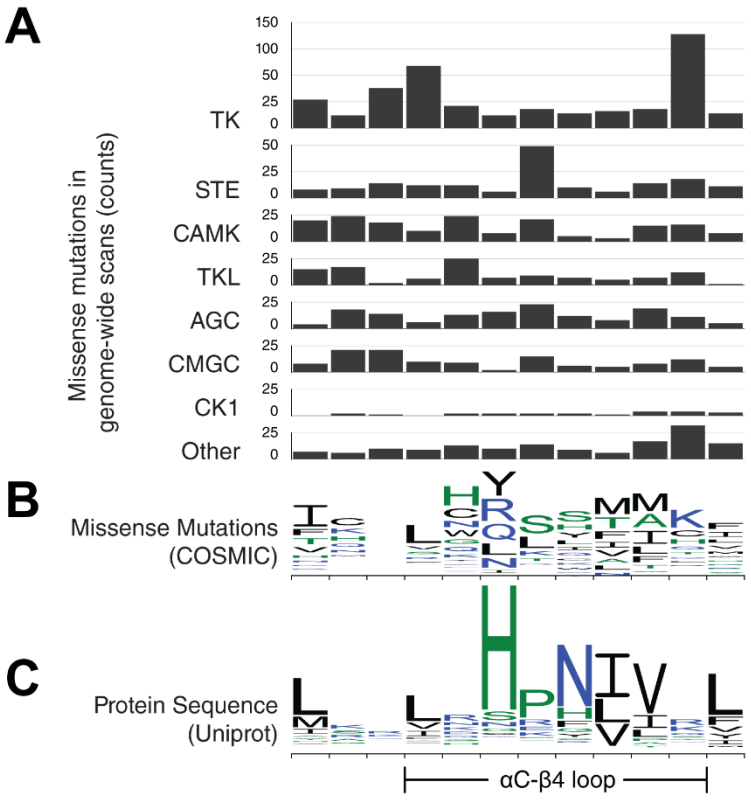


Figure 2.5: Missense mutations in the αC-β4 loop
For easy comparison, residue position on the x-axis is kept consistent throughout all graphs. (A) Bar graphs showing the number of missense mutations in the αC-β4 loop from 7 different protein kinase groups. The y-axis scale is consistent across all bar graphs to allow cross comparison (B) The missense mutations are shown using a sequence logo plot. Similar to a sequence logo, each column shows the relative frequency of all substitutions occurring at that position. (C) A sequence logo for wildtype αC-β4 sequences from Uniprot is provided as reference.

The molecular brake is a mutational hotspot in the αC-β4 loop

The most prevalent αC-β4 loop mutations occur at the HxN+3 position in tyrosine kinases **(Figure 2.5A)**. The HxN+3 position has been known to take part in the "molecular brake", a regulatory mechanism conserved in receptor tyrosine kinases (RTK) [58]. The molecular brake is a hydrogen bonding network mediated by three polar residues located at the kinase hinge region, including HxN+3. This regulatory mechanism was first identified in FGFR2 and extended to include several other RTKs through sequence comparison. In FGFR2, the molecular brake triad (N549, E565, and K641) locks the kinase in an inactive conformation. Mutations at the HxN+3, FGFR2$^{N549H/T}$, disengage the brake and activate the kinase [58].

Within the COSMIC genome-wide screens dataset, the majority of HxN+3 missense mutations substitute the RTK-conserved arginine **(Figure 2.2A)** for lysine **(Figure 2.5B)**. For example, cancer related HxN+3 mutations have been found in three FGFR family members, including FGFR1^{N546K}, FGFR2^{N549H}, FGFR3$^{N540K/S}$. These mutations have all been experimentally determined to be gain-of-function [58–62]. In other RTKs, PDGFRA$^{N659K/S}$ relieves the molecular brake, triggers constitutive STAT5 phosphorylation, and results in growth factor-independent cell proliferation [63,64]. Similarly, FLT3^{N676K} (HxN+3 position) increases autophosphorylation and downstream AKT/MAPK phosphorylation [65,66]. EGFR^{R776H} also increases autophosphorylation and preferentially adopts the acceptor position in the EGFR asymmetric dimer [67]. These examples suggest that HxN+3 position mutations are a common mechanism for tyrosine kinase activation in cancer cells.

Molecular brake mutations alter drug response

Comparative sequence studies have hypothesized that the αC-β4 loop is coupled with protein kinase activation by regulating inter-lobe movement and αC dynamics [32,67]. Supporting this notion, biophysical and biochemical studies suggest that the αC-β4 loop maintains auto-inhibitory interactions to prevent inadvertent kinase activation [32,58,67]. For example, nuclear magnetic resonance (NMR) and hydrogen-deuterium exchange mass spectrometry (HDX-MS) experiments on FGFR1 suggest that the molecular brake mechanism is coupled to activation loop conformation and active-inactive transition [5,23].

By taking advantage of this coupling, it is possible that molecular brake mutations in the αC-β4 loop (HxN+3) may confer drug resistance by altering the conformational equilibrium of a kinase, as opposed to directly altering the active site cleft [68]. The molecular brake stabilizes the auto-inhibited conformation of the kinase. HxN+3 mutations typically disrupt this inhibitory interaction and push the equilibrium toward the active conformation. Mutations that favor the active conformation (activating mutations) are generally resistant against Type II inhibitors, which target the inactive conformation [69]. At the HxN+3 position, examples of activating mutations that resist Type II inhibitors include KIT^{N655K} against imatinib and sunitinib [70,71] and FLT3-ITD^{N676K} ($FLT3^{N676K}$ with internal tandem duplication) resistance against quizartinib (AC220) [65,66,72,73]. Conversely, activating mutations can also result in sensitivity towards Type I inhibitors, which target the active conformation. For instance, $FLT3^{N676K}$ is sensitive to Type I inhibitor crenolanib [65,74].

Gatekeeper-proximal mutations in the αC-β4 loop associated with drug resistance

In the αC-β4 loop, the HxN+1 and HxN+2 positions are spatially proximal to the gatekeeper position: a well-studied hotspot for secondary drug resistance mutations [75,76]. Mutations at HxN+1 and HxN+2 positions have been associated with drug resistance in several tyrosine kinases. Both HxN+1 and HxN+2 take part in hydrophobic packing interactions that help form the kinase active site cleft in the N-lobe. Drawing parallels to resistance mutations at the gatekeeper position, mutations in HxN+1 and HxN+2 may alter the shape and packing of the active site cleft, which could sterically block or disfavor drug binding [69].

Abl^{L298V} (HxN+1) and Abl^{V299L} (HxN+2) have been shown to confer secondary drug resistance in leukemia patients [15,17,77]. Computational docking and molecular dynamics studies have predicted that these secondary mutations raise the free energy barrier of drug binding [29]. In another RTK, c-Kit^{V654A} (HxN+2) has been documented in imatinib-resistant gastrointestinal stromal tumors [78]. Although c-Kit^{V654A} does not result in constitutive kinase activity by itself, it can occur in conjunction with co-occurring mutations such as c-Kit^{V560G}, resulting in elevated kinase activity and factor-independent growth [78].

To the best of our knowledge, experimental characterization of HxN+1 and HxN+2 mutants are currently limited to the tyrosine kinases. However, HxN+2 seems to be a mutational hotspot across all 7 eukaryotic protein kinase groups **(Figure 2.5A)**. The position is highly conserved as a valine in all major protein kinase groups except CK1 **(Figure 2.2A)**. Although examples are currently limited to the tyrosine kinases, HxN+1 and HxN+2 mutations may be capable of conferring drug resistance in all protein kinases by modifying the shape/biochemical environment of the active site cleft.

Insertion mutations in the αC-β4 loop

Drug resistance mutations have been reported for EGFR and HER2 at exon 20 [79–81]. Exon 20 overlaps the αC-β4 loop and is a hotspot for insertion mutations. Historically, EGFR exon 20 insertion mutations are associated with resistance to first and second generation TK inhibitors. However, recent studies demonstrate differential responses to irreversible covalent inhibitors [16]. These differential responses depend on the sequence and location of the insertion. Further detailed characterization of these αC-β4 insertion mutations is crucial for understanding drug resistant mechanisms and, ultimately, for the development of effective protein kinase inhibitors [16,32].

Cis domain interactions affected by αC-β4 loop mutations

Many disease mutations target cis-interactions and interfere with normal kinase regulation. In TGF β-receptor I family, the regulatory GS domain interacts with a family-conserved arginine at the HxN-1 position [82–84]. Phosphorylation of the GS domain results in a conformational change that activates the kinase [84]. This conserved arginine interacts with the GS domain and shields it from phosphorylation. Disease mutations such as ACVR1$^{R258G/S}$ destabilize this interaction and result in constitutive kinase activity [82]. Constitutive activation of ACVR1 leads to fibrodysplasia ossificans progressiva (FOP), a rare disorder in extraskeletal bone formation [85]. In this example, the αC-β4 loop is capable of controlling kinase activity through interactions with regulatory domains.

Many kinases are regulated by long disordered regions flanking the kinase domain, also referred to as “tails”. EGFR is negatively regulated by its C-terminal tail, which makes electrostatic interactions with the αC-β4 loop and the hinge region of the kinase domain [86].

The autoinhibitory interaction at the αC-β4 loop is compromised by oncogenic mutations at the HxN+3 position, $EGFR^{R776H/C}$ [87]. Molecular dynamics and cell-based assays suggest that $EGFR^{R776H}$ weakens the inhibitory interaction and results in constitutive autophosphorylation [67]. Equivalent mutations are observed in HER2 and HER4, suggesting that disrupting this inhibitory mechanism is a common strategy for cancer cells to activate members of the EGFR family. In addition to the EGFR family, we note more examples of C-terminal tail interactions mediated by the αC-β4 loop in MAPK family [88,89] and IGF-1R [90].

Mutations that alter cis-interactions can also confer drug resistance. $MEK1^{P124L/S}$ was discovered to be resistant against selumetinib (AZD6244) in a random mutagenesis study [91]. $MEK1^{P124}$ is the HxN wildcard residue and packs against the MEK1 A-helix, a negative regulatory element located N-terminal to the kinase domain. Being a highly specific inhibitor, it is possible that selumetinib targets the inactive conformation which is disfavored in the absence of the A-helix interaction.

Another example can be found in ALK, an RTK. ALK is inhibited by its N-terminal juxtamembrane (JM) segment which makes hydrophobic contacts with the αC-helix and αC-β4 loop [92]. Phosphorylation of JM tyrosines results in ALK activation by disengaging the JM segment. ALK^{F1174L} is a recurring oncogenic mutation that disrupts pi-stacking interactions between the αC-β4 loop and the JM domain. Furthermore, ALK^{F1174L} results in constitutive kinase activity and confers resistance to crizotinib [93].

Protein-protein interactions affected by αC-β4 loop mutations

Mutations in the αC-β4 loop can also interfere with protein-protein interaction interfaces. In ERK2, the αC-β4 loop takes part in the D-recruitment site which helps the kinase bind to

effector proteins [94,95]. A patient derived mutation on the HxN wildcard residue, ERK2^{E81K}, activates the kinase [95]. Furthermore, ERK2^{E81K} may disrupt negative regulation by DUSP6 phosphatase [95].

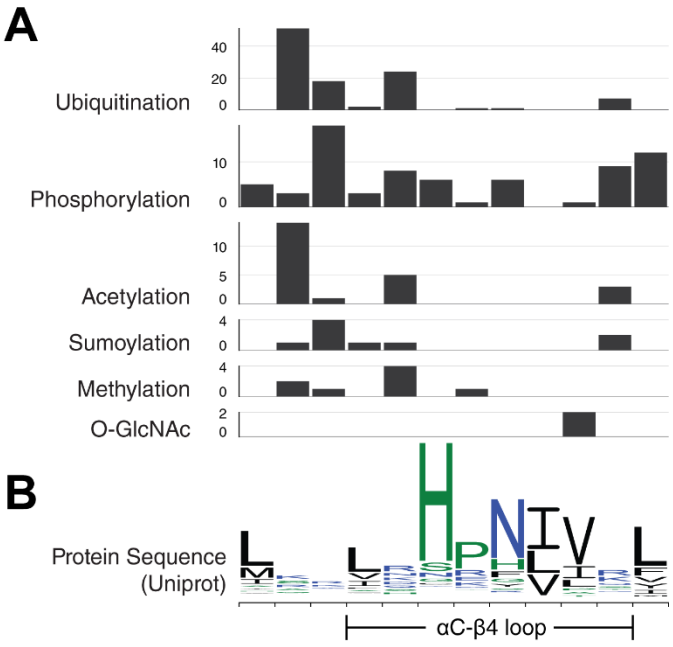


Figure 2.6: Post-translational modifications in the αC-β4 loop
For easy comparison, residue position on the x axis is kept consistent throughout all graphs. (A) Bar graphs showing the number of PTMs found at each αC-β4 position separated by PTM. Please note that the y-axis scale is not consistent across bar graphs. (B) A sequence logo for wildtype αC-β4 sequences from Uniprot is provided as reference.

2.2.3 Post-translational modifications in the αC-β4 loop

The catalytic function of many protein kinases is regulated by post-translational modifications (PTMs). For example, phosphorylation of the activation loop segment is required for the activation of many kinases [9]. The αC-β4 loop is also targeted by a variety of PTMs. To explore the landscape of PTMs within the αC-β4 loop, we retrieved a variety of mammalian PTMs from the PhosphoSitePlus database [96] **(Figure 2.6A)**. The majority of PTM sites were

identified by mass spectrometry and filtered by a statistical cutoff for assignment ($p < 0.05$). Within the database, available PTM assignments included ubiquitination, phosphorylation, acetylation, sumoylation, methylation, O-GlcNAc, and O-GalNAc. O-GalNAc was the only PTM without assignments to the αC-β4 loop.

Phosphorylation, one of the most abundant PTMs, plays a major role in modulating conformation and protein-protein interfaces. In the MST family, the HxN motif is replaced by a phosphorylatable SPx motif at the equivalent position. JNK phosphorylates $MST1^{S82}$ which is the SPx-Ser position. Phosphorylation of $MST1^{S82}$ enhances MST1 activity and promotes apoptosis [97]. The SPx wildcard residue is sometimes phosphorylatable. For instance, c-Abl phosphorylates $MST2^{Y81}$ at the SPx wildcard position. Phosphorylation of $MST2^{Y81}$ prevents MST2 from interacting with Raf-1 and promotes MST2 homodimerization [98].

O-linked β-N-acetylglucosamine (O-GlcNAc) is another important PTM that varies in response to many factors such as extracellular stress, cell cycle, and development [99]. Western blot and mass spectrometry (MS) assignments have revealed many O-GlcNAcylation sites on PKC family in rats [100]. WIthin the PKC family, O-GlcNAcylation sites were assigned to the αC-β4 loop of rat PKCA and PKCB. Some of these glycosylation sites intersect with phosphorylation sites suggesting that these modifications may modulate each other [101]. Although not in the αC-β4 loop, examples of kinase regulations via cross-talk between glycosylation and phosphorylation have been described in CaMKIV [101].

S-nitrosylation is also an important cysteine PTM that provides a mechanism for redox-based regulation [102]. In human InsR kinase, a modifiable cysteine replaces HxN-His. A study

on cultured skeletal muscle cells demonstrates that S-nitrosylation of $InsR^{C1083}$ results in the inhibition of kinase activity [103].

2.2.4 The αC-β4 loop mediates protein-protein interactions.

The αC-β4 loop is involved in kinase dimer interfaces

Many protein kinases are regulated by the formation of dimeric complexes. RAF kinases are activated by a side-to-side dimer interface involving the αC-β4 loop [104,105]. A mutation at the HxN-1 position, $BRAF^{R509H}$, impairs dimer formation and results in the kinase-dead phenotype [106]. Pseudokinase KSR is also capable of dimerizing with BRAF. Consequently, the equivalent mutation KSR^{R665H} also results in the loss of BRAF activity [107].

The αC-β4 loop can also take part in a symmetric back-to-back dimer interface. This conformation exposes the active site cleft and is usually associated with catalytically active kinase. One of the first examples was discovered in PKR, where the active back-to-back dimer was solved in two different crystallographic environments [108]. The IRE1 back-to-back dimer is also associated with an active kinase and high RNase activity [109]. Similarly, the Nek7-Nek9 back-to-back heterodimer is associated with rapid autophosphorylation of Nek7. Autophosphorylation assays showed that $Nek7^{N90R}$ (HxN-1) resulted in reduced kinase activity [110]. The proposed mechanism for PknB activation suggests that a back-to-back active dimer is induced by ligand binding to the PknB extracellular sensor domain [111,112]. PknE has also been crystallized in the back-to-back conformation [113].

The αC-β4 loop plays an important role in Hsp90-mediated kinase folding

One of the most important roles of the αC-β4 loop is the recognition of molecular chaperone Hsp90 and co-chaperone cdc37. Hsp90 promotes proper folding in many proteins including 60% of the human kinome [114]. This discovery started from an observation that human EGFR neither requires nor associates with Hsp90, a stark contrast to its paralog, HER2 [115,116]. However, mutation of the HER2 αC-β4 loop to the EGFR sequence abolished Hsp90 association in HER2 [115,117]. Furthermore, Fer^{Y616}, an αC-β4 loop residue, is essential for Hsp90 association and kinase activity [118]. Cryo-EM experiments have shown that co-chaperone cdc37 mimics the conformation of the αC-β4 loop and uses the HxN motif to form hinge interactions with the client kinase [119]. Although the full mechanism for Hsp90-recognition remains a mystery, results have shown that the HxN motif plays a role in cdc37-mediated folding for more than half of the human kinome [114].

2.2.5 Concluding remarks and predictions on understudied dark kinases

In this article, we have highlighted the αC-β4 loop as a central hub for many essential regulatory mechanisms for protein kinase function. To investigate the diverse functions of this region, we have compiled a list of disease-related mutations and PTMs that localize to the αC-β4 loop. We provide many examples of disease-related mutations linked to aberrant signaling and drug resistance. Experimental characterization shows that these mutations can alter both conserved and family-specific regulatory mechanisms. We believe that the αC-β4 loop is a conserved, yet understudied hotspot for regulatory interactions within the eukaryotic protein kinome.

Our kinome-wide analysis provides a useful resource for investigating understudied kinases. Recently, the NIH common fund program initiated a large-scale effort to identify and characterize new druggable proteins within the human genome. To guide research efforts, this initiative has maintained a list of understudied kinases (last updated on June 2019), collectively referred to as the dark kinome.

Interestingly, we find that nearly all kinases with extended αC-β4 loop segments have been classified as dark kinases **(Figure 2.4A)**. In the CMGC group, most members of the DYRK, HIPK, CLK, and SRPK families are classified as dark kinases. As opposed to the typical 8-residue loop, these families form a large clade whose members have a conserved ~14-residue αC-β4 loop. Members such as HIPK2, DYRK1A, and SRPK2 are some of the few characterized members of this clade. Using existing knowledge, we noticed clade-specific trends such as constitutive activity and helical inserts. These observations can guide hypothesis-driven research in the clade's understudied members. This approach can also be applied to the VRK family of the CK1 group. This family of dark kinases has a single well-characterized member, VRK1, and contains a conserved 21-residue αC-β4 loop. From a drug discovery perspective, these extended loop conformations may also provide a targetable interface for high-specificity protein kinase inhibitors.

The αC-β4 loop remains a regulatory hotspot within the complex web of interactions that modulate protein kinase activity. As such, mutations within this region can trigger a variety of human diseases. An in-depth understanding of the structure, function, and evolution of the αC-β4 loop will provide new insights on kinase regulation and enhance the discovery of novel protein kinase inhibitors.

Bibliography

[1] S. Gross, R. Rahal, N. Stransky, C. Lengauer, K.P. Hoeflich, Targeting cancer with kinase inhibitors, J Clin Invest. 125 (2015) 1780–1789. https://doi.org/10.1172/JCI76094.

[2] A. Fountas, L.-N. Diamantopoulos, A. Tsatsoulis, Tyrosine Kinase Inhibitors and Diabetes: A Novel Treatment Paradigm?, Trends in Endocrinology & Metabolism. 26 (2015) 643–656. https://doi.org/10.1016/j.tem.2015.09.003.

[3] J.-Q. Li, L. Tan, J.-T. Yu, The role of the LRRK2 gene in Parkinsonism, Molecular Neurodegeneration. 9 (2014) 47. https://doi.org/10.1186/1750-1326-9-47.

[4] R. Kumar, V.P. Singh, K.M. Baker, Kinase inhibitors for cardiovascular disease, Journal of Molecular and Cellular Cardiology. 42 (2007) 1–11. https://doi.org/10.1016/j.yjmcc.2006.09.005.

[5] N. Kannan, A.F. Neuwald, Did Protein Kinase Regulatory Mechanisms Evolve Through Elaboration of a Simple Structural Component?, Journal of Molecular Biology. 351 (2005) 956–972. https://doi.org/10.1016/j.jmb.2005.06.057.

[6] E.D. Scheeff, P.E. Bourne, Structural Evolution of the Protein Kinase–Like Superfamily, PLOS Computational Biology. 1 (2005) e49. https://doi.org/10.1371/journal.pcbi.0010049.

[7] J.A. Ubersax, J.E. Ferrell Jr, Mechanisms of specificity in protein phosphorylation, Nat Rev Mol Cell Biol. 8 (2007) 530–541. https://doi.org/10.1038/nrm2203.

[8] L.N. Johnson, M.E.M. Noble, D.J. Owen, Active and Inactive Protein Kinases: Structural Basis for Regulation, Cell. 85 (1996) 149–158. https://doi.org/10.1016/S0092-8674(00)81092-2.

[9] B. Nolen, S. Taylor, G. Ghosh, Regulation of protein kinases; controlling activity through activation segment conformation, Mol. Cell. 15 (2004) 661–675. https://doi.org/10.1016/j.molcel.2004.08.024.

[10] L. Palmieri, G. Rastelli, αC helix displacement as a general approach for allosteric modulation of protein kinases, Drug Discovery Today. 18 (2013) 407–414. https://doi.org/10.1016/j.drudis.2012.11.009.

[11] S.S. Taylor, A.S. Shaw, N. Kannan, A.P. Kornev, Integration of signaling in the kinome: Architecture and regulation of the αC Helix, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1854 (2015) 1567–1574. https://doi.org/10.1016/j.bbapap.2015.04.007.

[12] M. Huse, J. Kuriyan, The conformational plasticity of protein kinases, Cell. 109 (2002) 275–282.

[13] S.S. Taylor, A.P. Kornev, Protein kinases: evolution of dynamic regulatory proteins, Trends in Biochemical Sciences. 36 (2011) 65–77. https://doi.org/10.1016/j.tibs.2010.09.006.

[14] K. Oruganty, N.S. Talathi, Z.A. Wood, N. Kannan, Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases, PNAS. 110 (2013) 924–929. https://doi.org/10.1073/pnas.1207104110.

[15] D. Jones, S.S. Chen, E. Jabbour, M.B. Rios, H. Kantarjian, J. Cortes, Uncommon BCR-ABL kinase domain mutations in kinase inhibitor–resistant chronic myelogenous leukemia and Ph+ acute lymphoblastic leukemia show high rates of regression, suggesting weak selective effects, Blood. 115 (2010) 5428–5429. https://doi.org/10.1182/blood-2009-11-252155.

[16] T. Kosaka, J. Tanizaki, R.M. Paranal, H. Endoh, C. Lydon, M. Capelletti, C.E. Repellin, J. Choi, A. Ogino, A. Calles, D. Ercan, A.J. Redig, M. Bahcall, G.R. Oxnard, M.J. Eck, P.A. Jänne, Response Heterogeneity of EGFR and HER2 Exon 20 Insertions to Covalent EGFR and HER2 Inhibitors, Cancer Research. 77 (2017) 2712–2721. https://doi.org/10.1158/0008-5472.CAN-16-3404.

[17] F.E. Nicolini, S. Corm, Q.-H. Lê, N. Sorel, S. Hayette, D. Bories, T. Leguay, L. Roy, S. Giraudier, M. Tulliez, T. Facon, F.-X. Mahon, J.-M. Cayuela, P. Rousselot, M. Michallet, C. Preudhomme, F. Guilhot, C. Roche-Lestienne, Mutation status and clinical outcome of 89 imatinib mesylate-resistant chronic myelogenous leukemia patients: a retrospective analysis from the French intergroup of CML (Fi(ϕ)-LMC GROUP), Leukemia. 20 (2006) 1061–1066. https://doi.org/10.1038/sj.leu.2404236.

[18] J. Zheng, E.A. Trafny, D.R. Knighton, N. Xuong, S.S. Taylor, L.F. Ten Eyck, J.M. Sowadski, 2.2 Å refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor, Acta Crystallographica Section D. 49 (1993) 362–365. https://doi.org/10.1107/S0907444993000423.

[19] A.F. Neuwald, Rapid detection, classification and accurate alignment of up to a million or more related protein sequences, Bioinformatics. 25 (2009) 1869–1875. https://doi.org/10.1093/bioinformatics/btp342.

[20] W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers. 22 (1983) 2577–2637. https://doi.org/10.1002/bip.360221211.

[21] A. Kwon, S. Scott, R. Taujale, W. Yeung, K.J. Kochut, P.A. Eyers, N. Kannan, Tracing the origin and evolution of pseudokinases across the tree of life, Sci. Signal. 12 (2019). https://doi.org/10.1126/scisignal.aav3810.

[22] J. Bischof, J. Leban, M. Zaja, A. Grothey, B. Radunsky, O. Othersen, S. Strobl, D. Vitt, U. Knippschild, 2-Benzamido-N-(1H-benzo[d]imidazol-2-yl)thiazole-4-carboxamide derivatives as potent inhibitors of CK1δ/ε, Amino Acids. 43 (2012) 1577–1591. https://doi.org/10.1007/s00726-012-1234-x.

[23] N. Kannan, N. Haste, S.S. Taylor, A.F. Neuwald, The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module, Proc Natl Acad Sci U S A. 104 (2007) 1272–1277. https://doi.org/10.1073/pnas.0610251104.

[24] N. Kannan, A.F. Neuwald, S.S. Taylor, Analogous regulatory sites within the αC-β4 loop regions of ZAP-70 tyrosine kinase and AGC kinases, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1784 (2008) 27–32. https://doi.org/10.1016/j.bbapap.2007.09.007.

[25] F. Girdler, F. Sessa, S. Patercoli, F. Villa, A. Musacchio, S. Taylor, Molecular Basis of Drug Resistance in Aurora Kinases, Chemistry & Biology. 15 (2008) 552–562. https://doi.org/10.1016/j.chembiol.2008.04.013.

[26] A. Bondi, van der Waals Volumes and Radii, J. Phys. Chem. 68 (1964) 441–451. https://doi.org/10.1021/j100785a001.

[27] N. Thanki, Y. Umrania, J.M. Thornton, J.M. Goodfellow, Analysis of protein main-chain solvation as a function of secondary structure, Journal of Molecular Biology. 221 (1991) 669–691. https://doi.org/10.1016/0022-2836(91)80080-E.

[28] H.S. Meharena, P. Chang, M.M. Keshwani, K. Oruganty, A.K. Nene, N. Kannan, S.S. Taylor, A.P. Kornev, Deciphering the Structural Basis of Eukaryotic Protein Kinase Regulation, PLOS Biology. 11 (2013) e1001680. https://doi.org/10.1371/journal.pbio.1001680.

[29] D.L. Gibbons, S. Pricl, P. Posocco, E. Laurini, M. Fermeglia, H. Sun, M. Talpaz, N. Donato, A. Quintás-Cardama, Molecular dynamics reveal BCR-ABL1 polymutants as a unique mechanism of resistance to PAN-BCR-ABL1 kinase inhibitor therapy, Proc Natl Acad Sci U S A. 111 (2014) 3550–3555. https://doi.org/10.1073/pnas.1321173111.

[30] H. Liu, H. Wang, M. Teng, X. Li, The multiple nucleotide–divalent cation binding modes of Saccharomyces cerevisiae CK2α indicate a possible co-substrate hydrolysis product (ADP/GDP) release pathway, Acta Cryst D. 70 (2014) 501–513. https://doi.org/10.1107/S1399004713027879.

[31] J.C.K. Ngo, J. Gullingsrud, K. Giang, M.J. Yeh, X.-D. Fu, J.A. Adams, J.A. McCammon, G. Ghosh, SR Protein Kinase 1 Is Resilient to Inactivation, Structure. 15 (2007) 123–133. https://doi.org/10.1016/j.str.2006.11.011.

[32] Z. Ruan, N. Kannan, Altered conformational landscape and dimerization dependency underpins the activation of EGFR by αC-β4 loop insertion mutations, Proc Natl Acad Sci U S A. 115 (2018) E8162–E8171. https://doi.org/10.1073/pnas.1803152115.

[33] E.D. Scheeff, J. Eswaran, G. Bunkoczi, S. Knapp, G. Manning, Structure of the Pseudokinase VRK3 Reveals a Degraded Catalytic Site, a Highly Conserved Kinase Fold, and a Putative Regulatory Binding Site, Structure. 17 (2009) 128–138. https://doi.org/10.1016/j.str.2008.10.018.

[34] A. Kumar, J. Tamjar, A.D. Waddell, H.I. Woodroof, O.G. Raimi, A.M. Shaw, M. Peggie, M.M. Muqit, D.M. van Aalten, Structure of PINK1 and mechanisms of Parkinson's disease-associated mutations, ELife. 6 (2017) e29985. https://doi.org/10.7554/eLife.29985.

[35] B. Nolen, J. Ngo, S. Chakrabarti, D. Vu, J.A. Adams, G. Ghosh, Nucleotide-Induced Conformational Changes in the Saccharomyces cerevisiae SR Protein Kinase, Sky1p, Revealed by X-ray Crystallography, Biochemistry. 42 (2003) 9575–9585. https://doi.org/10.1021/bi0344331.

[36] A. Chaikuad, E. M C Tacconi, J. Zimmer, Y. Liang, N.S. Gray, M. Tarsounas, S. Knapp, A unique inhibitor binding site in ERK1/2 is associated with slow binding kinetics, Nat Chem Biol. 10 (2014) 853–860. https://doi.org/10.1038/nchembio.1629.

[37] J. Batson, H.D. Toop, C. Redondo, R. Babaei-Jadidi, A. Chaikuad, S.F. Wearmouth, B. Gibbons, C. Allen, C. Tallant, J. Zhang, C. Du, J.C. Hancox, T. Hawtrey, J. Da Rocha, R. Griffith, S. Knapp, D.O. Bates, J.C. Morris, Development of Potent, Selective SRPK1 Inhibitors as Potential Topical Therapeutics for Neovascular Eye Disease, ACS Chem. Biol. 12 (2017) 825–832. https://doi.org/10.1021/acschembio.6b01048.

[38] Q. Zhang, H. Liu, X. Liu, D. Jiang, B. Zhang, H. Tian, C. Yang, L.W. Guddat, H. Yang, K. Mi, Z. Rao, Discovery of the first macrolide antibiotic binding protein in Mycobacterium tuberculosis: a new antibiotic resistance drug target, Protein Cell. 9 (2018) 971–975. https://doi.org/10.1007/s13238-017-0502-7.

[39] J. Kallen, C. Bergsdorf, B. Arnaud, M. Bernhard, M. Brichet, A. Cobos-Correa, A. Elhajouji, F. Freuler, I. Galimberti, C. Guibourdenche, S. Haenni, S. Holzinger, J. Hunziker, A. Izaac, M. Kaufmann, L. Leder, H.-J. Martus, P. von Matt, V. Polyakov, P. Roethlisberger, G. Roma, N. Stiefl, M. Uteng, A. Lerchner, X-ray Structures and Feasibility Assessment of CLK2 Inhibitors for Phelan–McDermid Syndrome, ChemMedChem. 13 (2018) 1997–2007. https://doi.org/10.1002/cmdc.201800344.

[40] A. Walter, A. Chaikuad, R. Helmer, N. Loaëc, L. Preu, I. Ott, S. Knapp, L. Meijer, C. Kunick, Molecular structures of cdc2-like kinases in complex with a new inhibitor chemotype, PLOS ONE. 13 (2018) e0196761. https://doi.org/10.1371/journal.pone.0196761.

[41] K. Kim, J.S. Cha, Y.-S. Cho, H. Kim, N. Chang, H.-J. Kim, H.-S. Cho, Crystal Structure of Human Dual-Specificity Tyrosine-Regulated Kinase 3 Reveals New Structural Features and Insights into its Auto-phosphorylation, Journal of Molecular Biology. 430 (2018) 1521–1530. https://doi.org/10.1016/j.jmb.2018.04.001.

[42] A. Chaikuad, J. Diharce, M. Schröder, A. Foucourt, B. Leblond, A.-S. Casagrande, L. Désiré, P. Bonnet, S. Knapp, T. Besson, An Unusual Binding Model of the Methyl 9-Anilinothiazolo[5,4-f] quinazoline-2-carbimidates (EHT 1610 and EHT 5372) Confers High Selectivity for Dual-Specificity Tyrosine Phosphorylation-Regulated Kinases, J. Med. Chem. 59 (2016) 10315–10321. https://doi.org/10.1021/acs.jmedchem.6b01083.

[43] Q. Gao, I. Mechin, N. Kothari, Z. Guo, G. Deng, K. Haas, J. McManus, D. Hoffmann, A. Wang, D. Wiederschain, J. Rocnik, W. Czechtizky, X. Chen, L. McLean, H. Arlt, D. Harper, F. Liu, T. Majid, V. Patel, C. Lengauer, C. Garcia-Echeverria, B. Zhang, H. Cheng, M. Dorsch, S.-M.A. Huang, Evaluation of Cancer Dependence and Druggability of PRP4 Kinase Using Cellular, Biochemical, and Structural Approaches, Journal of Biological Chemistry. 288 (2013) 30125–30138. https://doi.org/10.1074/jbc.M113.473348.

[44] C. Agnew, L. Liu, S. Liu, W. Xu, L. You, W. Yeung, N. Kannan, D. Jablons, N. Jura, The crystal structure of the protein kinase HIPK2 reveals a unique architecture of its CMGC-insert region, Journal of Biological Chemistry. 294 (2019) 13545–13559. https://doi.org/10.1074/jbc.RA119.009725.

[45] N. Han, C. Chen, Z. Shi, D. Cheng, Structure of the kinase domain of Gilgamesh from Drosophila melanogaster, Acta Cryst F. 70 (2014) 438–443. https://doi.org/10.1107/S2053230X14004774.

[46] L. Liu, X. Li, J. Wang, M. Wang, P. Chen, M. Yin, J. Li, G. Sheng, Y. Wang, Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities, Cell. 168 (2017) 121-134.e12. https://doi.org/10.1016/j.cell.2016.12.031.

[47] A. Ishchenko, L. Zhang, J.-Y. Le Brazidec, J. Fan, J.H. Chong, A. Hingway, A. Raditsis, L. Singh, B. Elenbaas, V.S. Hong, D. Marcotte, L. Silvian, I. Enyedy, J. Chao, Structure-based design of low-nanomolar PIM kinase inhibitors, Bioorganic & Medicinal Chemistry Letters. 25 (2015) 474–480. https://doi.org/10.1016/j.bmcl.2014.12.041.

[48] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The Protein Kinase Complement of the Human Genome, Science. 298 (2002) 1912–1934. https://doi.org/10.1126/science.1075762.

[49] K.S. Metz, E.M. Deoudes, M.E. Berginski, I. Jimenez-Ruiz, B.A. Aksoy, J. Hammerbacher, S.M. Gomez, D.H. Phanstiel, Coral: Clear and Customizable Visualization of Human Kinome Data, Cell Systems. 7 (2018) 347-350.e1. https://doi.org/10.1016/j.cels.2018.07.001.

[50] W. Qiu, A. Wernimont, K. Tang, S. Taylor, V. Lunin, M. Schapira, S. Fentress, R. Hui, L.D. Sibley, Novel structural and regulatory features of rhoptry secretory kinases in Toxoplasma gondii, EMBO J. 28 (2009) 969–979. https://doi.org/10.1038/emboj.2009.24.

[51] E. Sajnaga, K. Kubiński, R. Szyszka, Catalytic activity of mutants of yeast protein kinase CK2alpha, Acta Biochim Pol. 55 (2008) 767–776.

[52] A.N. Bullock, S. Das, J.É. Debreczeni, P. Rellos, O. Fedorov, F.H. Niesen, K. Guo, E. Papagrigoriou, A.L. Amos, S. Cho, B.E. Turk, G. Ghosh, S. Knapp, Kinase Domain Insertions Define Distinct Roles of CLK Kinases in SR Protein Phosphorylation, Structure. 17 (2009) 352–362. https://doi.org/10.1016/j.str.2008.12.023.

[53] T. Adayev, M.-C. Chen-Hwang, N. Murakami, E. Lee, D.C. Bolton, Y.-W. Hwang, Dual-Specificity Tyrosine Phosphorylation-Regulated Kinase 1A Does Not Require Tyrosine Phosphorylation for Activity in Vitro, Biochemistry. 46 (2007) 7614–7624. https://doi.org/10.1021/bi700251n.

[54] M. Soundararajan, A.K. Roos, P. Savitsky, P. Filippakopoulos, A.N. Kettenbach, J.V. Olsen, S.A. Gerber, J. Eswaran, S. Knapp, J.M. Elkins, Structures of Down Syndrome Kinases, DYRKs, Reveal Mechanisms of Kinase Activation and Substrate Recognition, Structure. 21 (2013) 986–996. https://doi.org/10.1016/j.str.2013.03.012.

[55] R.M. Couñago, C.K. Allerston, P. Savitsky, H. Azevedo, P.H. Godoi, C.I. Wells, A. Mascarello, F.H. de Souza Gama, K.B. Massirer, W.J. Zuercher, C.R.W. Guimarães, O. Gileadi, Structural characterization of human Vaccinia-Related Kinases (VRK) bound to small-molecule inhibitors identifies different P-loop conformations, Sci Rep. 7 (2017) 7501. https://doi.org/10.1038/s41598-017-07755-y.

[56] E. Talevich, N. Kannan, Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors, BMC Evolutionary Biology. 13 (2013) 117. https://doi.org/10.1186/1471-2148-13-117.

[57] J.G. Tate, S. Bamford, H.C. Jubb, Z. Sondka, D.M. Beare, N. Bindal, H. Boutselakis, C.G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S.C. Jupe, C.Y. Kok, K. Noble, L. Ponting, C.C. Ramshaw, C.E. Rye, H.E. Speedy, R. Stefancsik, S.L. Thompson, S. Wang, S. Ward, P.J. Campbell, S.A. Forbes, COSMIC: the Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Research. 47 (2019) D941–D947. https://doi.org/10.1093/nar/gky1015.

[58] H. Chen, J. Ma, W. Li, A.V. Eliseenkova, C. Xu, T.A. Neubert, W.T. Miller, M. Mohammadi, A Molecular Brake in the Kinase Hinge Region Regulates the Activity of Receptor Tyrosine Kinases, Molecular Cell. 27 (2007) 717–730. https://doi.org/10.1016/j.molcel.2007.06.028.

[59] L.H. Gallo, K.N. Nelson, A.N. Meyer, D.J. Donoghue, Functions of Fibroblast Growth Factor Receptors in cancer defined by novel translocations and mutations, Cytokine & Growth Factor Reviews. 26 (2015) 425–449. https://doi.org/10.1016/j.cytogfr.2015.03.003.

[60] E.D. Lew, C.M. Furdui, K.S. Anderson, J. Schlessinger, The Precise Sequence of FGF Receptor Autophosphorylation Is Kinetically Driven and Is Disrupted by Oncogenic Mutations, Science Signaling. 2 (2009) ra6–ra6. https://doi.org/10.1126/scisignal.2000021.

[61] H. Patani, T.D. Bunney, N. Thiyagarajan, R.A. Norman, D. Ogg, J. Breed, P. Ashford, A. Potterton, M. Edwards, S.V. Williams, G.S. Thomson, C.S.M. Pang, M.A. Knowles, A.L. Breeze, C. Orengo, C. Phillips, M. Katan, Landscape of activating cancer mutations in FGFR kinases and their differential responses to inhibitors in clinical use, Oncotarget. 7 (2016) 24252–24268. https://doi.org/10.18632/oncotarget.8132.

[62] V. Rand, J. Huang, T. Stockwell, S. Ferriera, O. Buzko, S. Levy, D. Busam, K. Li, J.B. Edwards, C. Eberhart, K.M. Murphy, A. Tsiamouri, K. Beeson, A.J.G. Simpson, J.C. Venter, G.J. Riggins, R.L. Strausberg, Sequence survey of receptor tyrosine kinases reveals mutations in glioblastomas, Proc Natl Acad Sci U S A. 102 (2005) 14344–14349. https://doi.org/10.1073/pnas.0507200102.

[63] C.L. Corless, A. Schroeder, D. Griffith, A. Town, L. McGreevey, P. Harrell, S. Shiraga, T. Bainbridge, J. Morich, M.C. Heinrich, PDGFRA Mutations in Gastrointestinal Stromal Tumors: Frequency, Spectrum and In Vitro Sensitivity to Imatinib, JCO. 23 (2005) 5357–5364. https://doi.org/10.1200/JCO.2005.14.068.

[64] C. Elling, P. Erben, C. Walz, M. Frickenhaus, M. Schemionek, M. Stehling, H. Serve, N.C.P. Cross, A. Hochhaus, W.-K. Hofmann, W.E. Berdel, C. Müller-Tidow, A. Reiter, S. Koschmieder, Novel imatinib-sensitive PDGFRA-activating point mutations in hypereosinophilic syndrome induce growth factor independence and leukemia-like disease, Blood. 117 (2011) 2935–2943. https://doi.org/10.1182/blood-2010-05-286757.

[65] K. Huang, M. Yang, Z. Pan, F.H. Heidel, M. Scherr, M. Eder, T. Fischer, G. Büsche, K. Welte, N. von Neuhoff, A. Ganser, Z. Li, Leukemogenic potency of the novel FLT3-N676K mutant, Ann Hematol. 95 (2016) 783–791. https://doi.org/10.1007/s00277-016-2616-z.

[66] S. Opatz, H. Polzer, T. Herold, N.P. Konstandin, B. Ksienzyk, E. Zellmeier, S. Vosberg, A. Graf, S. Krebs, H. Blum, K.-P. Hopfner, P.M. Kakadia, S. Schneider, A. Dufour, J. Braess, M.C. Sauerland, W.E. Berdel, T. Büchner, B.J. Woermann, W. Hiddemann, K. Spiekermann, S.K. Bohlander, P.A. Greif, Exome sequencing identifies recurring FLT3 N676K mutations in core-binding factor leukemia, Blood. 122 (2013) 1761–1769. https://doi.org/10.1182/blood-2013-01-476473.

[67] Z. Ruan, N. Kannan, Mechanistic Insights into R776H Mediated Activation of Epidermal Growth Factor Receptor Kinase, Biochemistry. 54 (2015) 4216–4225. https://doi.org/10.1021/acs.biochem.5b00444.

[68] C.B. Gambacorti-Passerini, R.H. Gunby, R. Piazza, A. Galietta, R. Rostagno, L. Scapozza, Molecular mechanisms of resistance to imatinib in Philadelphia-chromosome-positive leukaemias, The Lancet Oncology. 4 (2003) 75–85. https://doi.org/10.1016/S1470-2045(03)00979-3.

[69] R. Roskoski, Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes, Pharmacological Research. 103 (2016) 26–48. https://doi.org/10.1016/j.phrs.2015.10.021.

[70] K.S. Gajiwala, J.C. Wu, J. Christensen, G.D. Deshmukh, W. Diehl, J.P. DiNitto, J.M. English, M.J. Greig, Y.-A. He, S.L. Jacques, E.A. Lunney, M. McTigue, D. Molina, T. Quenzer, P.A. Wells, X. Yu, Y. Zhang, A. Zou, M.R. Emmett, A.G. Marshall, H.-M. Zhang, G.D. Demetri, KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients, Proc Natl Acad Sci U S A. 106 (2009) 1542–1547. https://doi.org/10.1073/pnas.0812413106.

[71] A.P. Garner, J.M. Gozgit, R. Anjum, S. Vodala, A. Schrock, T. Zhou, C. Serrano, G. Eilers, M. Zhu, J. Ketzer, S. Wardwell, Y. Ning, Y. Song, A. Kohlmann, F. Wang, T. Clackson, M.C. Heinrich, J.A. Fletcher, S. Bauer, V.M. Rivera, Ponatinib Inhibits Polyclonal Drug-Resistant KIT Oncoproteins and Shows Therapeutic Potential in Heavily Pretreated Gastrointestinal Stromal Tumor (GIST) Patients, Clinical Cancer Research. 20 (2014) 5745–5755. https://doi.org/10.1158/1078-0432.CCR-14-1397.

[72] C.C. Smith, C. Zhang, K.C. Lin, E.A. Lasater, Y. Zhang, E. Massi, L.E. Damon, M. Pendleton, A. Bashir, R. Sebra, A. Perl, A. Kasarskis, R. Shellooe, G. Tsang, H. Carias, B. Powell, E.A. Burton, B. Matusow, J. Zhang, W. Spevak, P.N. Ibrahim, M.H. Le, H.H. Hsu, G. Habets, B.L. West, G. Bollag, N.P. Shah, Characterizing and Overriding the Structural Mechanism of the Quizartinib-Resistant FLT3 "Gatekeeper" F691L Mutation with PLX3397, Cancer Discovery. 5 (2015) 668–679. https://doi.org/10.1158/2159-8290.CD-15-0060.

[73] J.A. Zorn, Q. Wang, E. Fujimura, T. Barros, J. Kuriyan, Crystal Structure of the FLT3 Kinase Domain Bound to the Inhibitor Quizartinib (AC220), PLOS ONE. 10 (2015) e0121177. https://doi.org/10.1371/journal.pone.0121177.

[74] A. Galanis, T. Rajkhowa, C. Muralidhara, A. Ramachandran, M.J. Levis, Crenolanib Is A Highly Potent, Selective, FLT3 TKI with Activity Against D835 Mutations, Blood. 120 (2012) 1341. https://doi.org/10.1182/blood.V120.21.1341.1341.

[75] M. Azam, M.A. Seeliger, N.S. Gray, J. Kuriyan, G.Q. Daley, Activation of tyrosine kinases by mutation of the gatekeeper threonine, Nat Struct Mol Biol. 15 (2008) 1109–1118. https://doi.org/10.1038/nsmb.1486.

[76] C.-H. Yun, K.E. Mengwasser, A.V. Toms, M.S. Woo, H. Greulich, K.-K. Wong, M. Meyerson, M.J. Eck, The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP, Proc Natl Acad Sci U S A. 105 (2008) 2070–2075. https://doi.org/10.1073/pnas.0709662105.

[77] E. Jabbour, V. Morris, H. Kantarjian, C.C. Yin, E. Burton, J. Cortes, Characteristics and outcomes of patients with V299L BCR-ABL kinase domain mutation after therapy with tyrosine kinase inhibitors, Blood. 120 (2012) 3382–3383. https://doi.org/10.1182/blood-2012-04-424192.

[78] K.G. Roberts, A.F. Odell, E.M. Byrnes, R.M. Baleato, R. Griffith, A.B. Lyons, L.K. Ashman, Resistance to c-KIT kinase inhibitors conferred by V654A mutation, Molecular Cancer Therapeutics. 6 (2007) 1159–1166. https://doi.org/10.1158/1535-7163.MCT-06-0641.

[79] Y. Hirotsu, H. Nakagomi, K. Amemiya, T. Oyama, M. Inoue, H. Mochizuki, M. Omata, Intrinsic HER2 V777L mutation mediates resistance to trastuzumab in a breast cancer patient, Med Oncol. 34 (2016) 3. https://doi.org/10.1007/s12032-016-0857-2.

[80] H. Yasuda, S. Kobayashi, D.B. Costa, EGFR exon 20 insertion mutations in non-small-cell lung cancer: preclinical data and clinical implications, The Lancet Oncology. 13 (2012) e23–e31. https://doi.org/10.1016/S1470-2045(11)70129-2.

[81] H. Yasuda, E. Park, C.-H. Yun, N.J. Sng, A.R. Lucena-Araujo, W.-L. Yeo, M.S. Huberman, D.W. Cohen, S. Nakayama, K. Ishioka, N. Yamaguchi, M. Hanna, G.R. Oxnard, C.S. Lathan, T. Moran, L.V. Sequist, J.E. Chaft, G.J. Riely, M.E. Arcila, R.A. Soo, M. Meyerson, M.J. Eck, S.S. Kobayashi, D.B. Costa, Structural, Biochemical, and Clinical Characterization of Epidermal Growth Factor Receptor (EGFR) Exon 20 Insertion Mutations in Lung Cancer, Science Translational Medicine. 5 (2013) 216ra177-216ra177. https://doi.org/10.1126/scitranslmed.3007205.

[82] A. Chaikuad, I. Alfano, G. Kerr, C.E. Sanvitale, J.H. Boergermann, J.T. Triffitt, F. von Delft, S. Knapp, P. Knaus, A.N. Bullock, Structure of the Bone Morphogenetic Protein Receptor ALK2 and Implications for Fibrodysplasia Ossificans Progressiva*, Journal of Biological Chemistry. 287 (2012) 36990–36998. https://doi.org/10.1074/jbc.M112.365932.

[83] M. Huse, Y.-G. Chen, J. Massagué, J. Kuriyan, Crystal Structure of the Cytoplasmic Domain of the Type I TGF β Receptor in Complex with FKBP12, Cell. 96 (1999) 425–436. https://doi.org/10.1016/S0092-8674(00)80555-3.

[84] M. Huse, T.W. Muir, L. Xu, Y.-G. Chen, J. Kuriyan, J. Massagué, The TGFβ Receptor Activation Process: An Inhibitor- to Substrate-Binding Switch, Molecular Cell. 8 (2001) 671–682. https://doi.org/10.1016/S1097-2765(01)00332-X.

[85] F.S. Kaplan, J.A. Kobori, C. Orellana, I. Calvo, M. Rosello, F. Martinez, B. Lopez, M. Xu, R.J. Pignolo, E.M. Shore, J.C. Groppe, Multi-system involvement in a severe variant of fibrodysplasia ossificans progressiva (ACVR1 c.772G>A; R258G): A report of two patients, American Journal of Medical Genetics Part A. 167 (2015) 2265–2271. https://doi.org/10.1002/ajmg.a.37205.

[86] E. Kovacs, R. Das, Q. Wang, T.S. Collier, A. Cantor, Y. Huang, K. Wong, A. Mirza, T. Barros, P. Grob, N. Jura, R. Bose, J. Kuriyan, Analysis of the Role of the C-Terminal Tail in the Regulation of the Epidermal Growth Factor Receptor, Molecular and Cellular Biology. 35 (2015) 3083–3102. https://doi.org/10.1128/MCB.00248-15.

[87] D.I. McSkimming, S. Dastgheib, E. Talevich, A. Narayanan, S. Katiyar, S.S. Taylor, K. Kochut, N. Kannan, ProKinO: A Unified Resource for Mining the Cancer Kinome, Human Mutation. 36 (2015) 175–186. https://doi.org/10.1002/humu.22726.

[88] R. Diskin, M. Lebendiker, D. Engelberg, O. Livnah, Structures of p38α Active Mutants Reveal Conformational Changes in L16 Loop that Induce Autophosphorylation and Activation, Journal of Molecular Biology. 365 (2007) 66–76. https://doi.org/10.1016/j.jmb.2006.08.043.

[89] J.M. Salvador, P.R. Mittelstadt, T. Guszczynski, T.D. Copeland, H. Yamaguchi, E. Appella, A.J. Fornace, J.D. Ashwell, Alternative p38 activation pathway mediated by T

cell receptor–proximal tyrosine kinases, Nat Immunol. 6 (2005) 390–395. https://doi.org/10.1038/ni1177.

[90] G.M. Kelly, D.A. Buckley, P.A. Kiely, D.R. Adams, R. O'Connor, Serine Phosphorylation of the Insulin-like Growth Factor I (IGF-1) Receptor C-terminal Tail Restrains Kinase Activity and Cell Growth*, Journal of Biological Chemistry. 287 (2012) 28180–28194. https://doi.org/10.1074/jbc.M112.385757.

[91] C.M. Emery, K.G. Vijayendran, M.C. Zipser, A.M. Sawyer, L. Niu, J.J. Kim, C. Hatton, R. Chopra, P.A. Oberholzer, M.B. Karpova, L.E. MacConaill, J. Zhang, N.S. Gray, W.R. Sellers, R. Dummer, L.A. Garraway, MEK1 mutations confer resistance to MEK and B-RAF inhibition, Proc Natl Acad Sci U S A. 106 (2009) 20411–20416. https://doi.org/10.1073/pnas.0905833106.

[92] Q. Huang, T.W. Johnson, S. Bailey, A. Brooun, K.D. Bunker, B.J. Burke, M.R. Collins, A.S. Cook, J.J. Cui, K.N. Dack, J.G. Deal, Y.-L. Deng, D. Dinh, L.D. Engstrom, M. He, J. Hoffman, R.L. Hoffman, P.S. Johnson, R.S. Kania, H. Lam, J.L. Lam, P.T. Le, Q. Li, L. Lingardo, W. Liu, M.W. Lu, M. McTigue, C.L. Palmer, P.F. Richardson, N.W. Sach, H. Shen, T. Smeal, G.L. Smith, A.E. Stewart, S. Timofeevski, K. Tsaparikos, H. Wang, H. Zhu, J. Zhu, H.Y. Zou, M.P. Edwards, Design of Potent and Selective Inhibitors to Overcome Clinical Anaplastic Lymphoma Kinase Mutations Resistant to Crizotinib, J. Med. Chem. 57 (2014) 1170–1187. https://doi.org/10.1021/jm401805h.

[93] T. Berry, W. Luther, N. Bhatnagar, Y. Jamin, E. Poon, T. Sanda, D. Pei, B. Sharma, W.R. Vetharoy, A. Hallsworth, Z. Ahmad, K. Barker, L. Moreau, H. Webber, W. Wang, Q. Liu, A. Perez-Atayde, S. Rodig, N.-K. Cheung, F. Raynaud, B. Hallberg, S.P. Robinson, N.S. Gray, A.D.J. Pearson, S.A. Eccles, L. Chesler, R.E. George, The ALKF1174L Mutation Potentiates the Oncogenic Activity of MYCN in Neuroblastoma, Cancer Cell. 22 (2012) 117–130. https://doi.org/10.1016/j.ccr.2012.06.001.

[94] A. Alexa, G. Gógl, G. Glatz, Á. Garai, A. Zeke, J. Varga, E. Dudás, N. Jeszenői, A. Bodor, C. Hetényi, A. Reményi, Structural assembly of the signaling competent ERK2-RSK1 heterodimeric protein kinase complex, Proc Natl Acad Sci U S A. 112 (2015) 2711–2716. https://doi.org/10.1073/pnas.1417571112.

[95] L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N.S. Persky, C. Zhu, M. Bagul, E.M. Goetz, A.B. Burgin, L.A. Garraway, G. Getz, T.S. Mikkelsen, F. Piccioni, D.E. Root, C.M. Johannessen, Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants, Cell Reports. 17 (2016) 1171–1183. https://doi.org/10.1016/j.celrep.2016.09.061.

[96] P.V. Hornbeck, B. Zhang, B. Murray, J.M. Kornhauser, V. Latham, E. Skrzypek, PhosphoSitePlus, 2014: mutations, PTMs and recalibrations, Nucleic Acids Research. 43 (2015) D512–D520. https://doi.org/10.1093/nar/gku1267.

[97] W. Bi, L. Xiao, Y. Jia, J. Wu, Q. Xie, J. Ren, G. Ji, Z. Yuan, c-Jun N-terminal Kinase Enhances MST1-mediated Pro-apoptotic Signaling through Phosphorylation at Serine 82*,

Journal of Biological Chemistry. 285 (2010) 6259–6264. https://doi.org/10.1074/jbc.M109.038570.

[98] W. Liu, J. Wu, L. Xiao, Y. Bai, A. Qu, Z. Zheng, Z. Yuan, Regulation of Neuronal Cell Death by c-Abl-Hippo/MST2 Signaling Pathway, PLOS ONE. 7 (2012) e36562. https://doi.org/10.1371/journal.pone.0036562.

[99] L. Wells, S.A. Whelan, G.W. Hart, O-GlcNAc: a regulatory post-translational modification, Biochemical and Biophysical Research Communications. 302 (2003) 435–441. https://doi.org/10.1016/S0006-291X(03)00175-X.

[100] M. Robles-Flores, L. Meléndez, W. García, G. Mendoza-Hernández, T.T. Lam, C. Castañeda-Patlán, H. González-Aguilar, Posttranslational modifications on protein kinase c isozymes. Effects of epinephrine and phorbol esters, Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 1783 (2008) 695–712. https://doi.org/10.1016/j.bbamcr.2007.07.011.

[101] G.W. Hart, C. Slawson, G. Ramirez-Correa, O. Lagerlof, Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease, Annu Rev Biochem. 80 (2011) 825–858. https://doi.org/10.1146/annurev-biochem-060608-102511.

[102] D.T. Hess, A. Matsumoto, S.-O. Kim, H.E. Marshall, J.S. Stamler, Protein S-nitrosylation: purview and parameters, Nat Rev Mol Cell Biol. 6 (2005) 150–166. https://doi.org/10.1038/nrm1569.

[103] E. Schmid, A. Hotz-Wagenblatt, W. Dröge, Inhibition of the Insulin Receptor Kinase Phosphorylation by Nitric Oxide: Functional and Structural Aspects, Antioxidants & Redox Signaling. 1 (1999) 45–53. https://doi.org/10.1089/ars.1999.1.1-45.

[104] T. Rajakulendran, M. Sahmi, M. Lefrançois, F. Sicheri, M. Therrien, A dimerization-dependent mechanism drives RAF catalytic activation, Nature. 461 (2009) 542–545. https://doi.org/10.1038/nature08314.

[105] P.T.C. Wan, M.J. Garnett, S.M. Roe, S. Lee, D. Niculescu-Duvaz, V.M. Good, C.G. Project, C.M. Jones, C.J. Marshall, C.J. Springer, D. Barford, R. Marais, Mechanism of Activation of the RAF-ERK Signaling Pathway by Oncogenic Mutations of B-RAF, Cell. 116 (2004) 855–867. https://doi.org/10.1016/S0092-8674(04)00215-6.

[106] J. Hu, E.C. Stites, H. Yu, E.A. Germino, H.S. Meharena, P.J.S. Stork, A.P. Kornev, S.S. Taylor, A.S. Shaw, Allosteric Activation of Functionally Asymmetric RAF Kinase Dimers, Cell. 154 (2013) 1036–1046. https://doi.org/10.1016/j.cell.2013.07.046.

[107] H. Lavoie, M. Sahmi, P. Maisonneuve, S.A. Marullo, N. Thevakumaran, T. Jin, I. Kurinov, F. Sicheri, M. Therrien, MEK drives BRAF activation through allosteric control of KSR proteins, Nature. 554 (2018) 549–553. https://doi.org/10.1038/nature25478.

[108] A.C. Dar, T.E. Dever, F. Sicheri, Higher-Order Substrate Recognition of eIF2α by the RNA-Dependent Protein Kinase PKR, Cell. 122 (2005) 887–900. https://doi.org/10.1016/j.cell.2005.06.044.

[109] A. Joshi, Y. Newbatt, P.C. McAndrew, M. Stubbs, R. Burke, M.W. Richards, C. Bhatia, J.J. Caldwell, T. McHardy, I. Collins, R. Bayliss, Molecular mechanisms of human IRE1 activation through dimerization and ligand binding, Oncotarget. 6 (2015) 13019–13035. https://doi.org/10.18632/oncotarget.3864.

[110] T. Haq, M.W. Richards, S.G. Burgess, P. Gallego, S. Yeoh, L. O'Regan, D. Reverter, J. Roig, A.M. Fry, R. Bayliss, Mechanistic basis of Nek7 activation through Nek9 binding and induced dimerization, Nat Commun. 6 (2015) 8771. https://doi.org/10.1038/ncomms9771.

[111] T.N. Lombana, N. Echols, M.C. Good, N.D. Thomsen, H.-L. Ng, A.E. Greenstein, A.M. Falick, D.S. King, T. Alber, Allosteric Activation Mechanism of the Mycobacterium tuberculosis Receptor Ser/Thr Protein Kinase, PknB, Structure. 18 (2010) 1667–1677. https://doi.org/10.1016/j.str.2010.09.019.

[112] T.A. Young, B. Delagoutte, J.A. Endrizzi, A.M. Falick, T. Alber, Structure of Mycobacterium tuberculosis PknB supports a universal activation mechanism for Ser/Thr protein kinases, Nat Struct Mol Biol. 10 (2003) 168–174. https://doi.org/10.1038/nsb897.

[113] L.M. Gay, H.-L. Ng, T. Alber, A Conserved Dimer and Global Conformational Changes in the Structure of apo-PknE Ser/Thr Protein Kinase from Mycobacterium tuberculosis, Journal of Molecular Biology. 360 (2006) 409–420. https://doi.org/10.1016/j.jmb.2006.05.015.

[114] M. Taipale, I. Krykbaeva, M. Koeva, C. Kayatekin, K.D. Westover, G.I. Karras, S. Lindquist, Quantitative Analysis of Hsp90-Client Interactions Reveals Principles of Substrate Recognition, Cell. 150 (2012) 987–1001. https://doi.org/10.1016/j.cell.2012.06.047.

[115] W. Xu, X. Yuan, Z. Xiang, E. Mimnaugh, M. Marcu, L. Neckers, Surface charge and hydrophobicity determine ErbB2 binding to the Hsp90 chaperone complex, Nat Struct Mol Biol. 12 (2005) 120–126. https://doi.org/10.1038/nsmb885.

[116] W. Xu, S. Soga, K. Beebe, M.-J. Lee, Y.S. Kim, J. Trepel, L. Neckers, Sensitivity of epidermal growth factor receptor and ErbB2 exon 20 insertion mutants to Hsp90 inhibition, Br J Cancer. 97 (2007) 741–744. https://doi.org/10.1038/sj.bjc.6603950.

[117] A. Citri, D. Harari, G. Shohat, P. Ramakrishnan, J. Gan, S. Lavi, M. Eisenstein, A. Kimchi, D. Wallach, S. Pietrokovski, Y. Yarden, Hsp90 Recognizes a Common Surface on Client Kinases, Journal of Biological Chemistry. 281 (2006) 14361–14369. https://doi.org/10.1074/jbc.M512613200.

[118] E. Hikri, S. Shpungin, U. Nir, Hsp90 and a tyrosine embedded in the Hsp90 recognition loop are required for the Fer tyrosine kinase activity, Cellular Signalling. 21 (2009) 588–596. https://doi.org/10.1016/j.cellsig.2008.12.011.

[119] K.A. Verba, D.A. Agard, How Hsp90 and Cdc37 Lubricate Kinase Molecular Switches, Trends in Biochemical Sciences. 42 (2017) 799–811. https://doi.org/10.1016/j.tibs.2017.07.002.

Chapter 3

EVOLUTION OF FUNCTIONAL DIVERSITY IN THE HOLOZOAN TYROSINE KINOME

W. Yeung, A. Kwon, R. Taujale, C. Bunn, A. Venkat, N. Kannan, Evolution of Functional Diversity in the Holozoan Tyrosine Kinome, Molecular Biology and Evolution. 38 (2021) 5625–5639.

Reprinted here with permission of publisher.

Abstract

The emergence of multicellularity is strongly correlated with the expansion of tyrosine kinases, a conserved family of signaling enzymes that regulates pathways essential for cell-to-cell communication. Although tyrosine kinases have been classified from several model organisms, a molecular-level understanding of tyrosine kinase evolution across all holozoans is currently lacking. Using a hierarchical sequence constraint-based classification of diverse holozoan tyrosine kinases, we construct a new phylogenetic tree that identifies two ancient clades of cytoplasmic and receptor tyrosine kinases separated by the presence of an extended insert segment in the kinase domain connecting the D and E-helices. Present in nearly all receptor tyrosine kinases, this fast-evolving insertion imparts diverse functionalities such as post-translational modification sites and regulatory interactions. Eph and EGFR receptor tyrosine kinases are two exceptions which lack this insert, each forming an independent lineage characterized by unique functional features. We also identify common constraints shared across multiple tyrosine kinase families which warrant the designation of three new subgroups: Src Module (SrcM), Insulin Receptor Kinase-Like (IRKL), and Fibroblast, Platelet-derived, Vascular, and growth factor Receptors (FPVR). Subgroup-specific constraints reflect shared autoinhibitory interactions involved in kinase conformational regulation. Conservation analyses describe how diverse tyrosine kinase signaling functions arose through the addition of family-specific motifs upon subgroup-specific features and co-evolving protein domains. We propose the oldest tyrosine kinases, IRKL, SrcM, and Csk, originated from unicellular pre-metazoans and were co-opted for complex multicellular functions. The increased frequency of oncogenic variants in more recent tyrosine kinases suggests that lineage-specific functionalities are selectively altered in human cancers.

3.1 Introduction

Tyrosine kinases propagate cellular signals through the phosphorylation of tyrosine residues on protein substrates. Forming a monophyletic group within the larger protein kinase superfamily, tyrosine kinases diverged from serine-threonine kinases prior to the emergence of opisthokonts (animals and fungi) [1,2], which are estimated to be over a billion years old [3]. While their detection in unicellular pre-metazoans such as choanoflagellates and filastereans has indicated the fundamental roles of tyrosine kinases in the evolution of multicellularity [4–6], their subsequent expansion throughout metazoan evolution is associated with the evolution of diverse metazoan body plans and complex biological systems such as the nervous, vascular, and immune systems [5,7]. Given the vast diversity of tyrosine kinases, the diversification events that gave rise to the functional repertoire of tyrosine kinases and the evolutionary timeline of such events has not been fully explored.

A classification of the protein kinome into evolutionarily and functionally related families (here on referred to as the KinBase classification) was achieved two decades ago following the sequencing and comparative genomic analyses of model organism genomes including human [8], mouse [9], sea urchin [10], fly [11], nematode [12], sponge [13], choanoflagellate [4], and yeast [11]. In addition, tyrosine kinases can be broadly classified as cytoplasmic or receptor tyrosine kinases based on the presence of transmembrane and extracellular ligand binding domains; however, unlike kinase groups and families defined in the KinBase classification, cytoplasmic and receptor tyrosine kinases do not form monophyletic clades in the kinome tree since receptor tyrosine kinases are believed to have independently emerged multiple times throughout tyrosine kinase evolution [1,14]. The KinBase classification has subsequently become a foundation for comparative studies to study the conservation and

divergence of kinase sequence, structure, and function. For example, previous studies of the patterns of sequence conservation and variation across kinase families and groups have provided important insights into the unique regulatory spine of tyrosine kinases relative to serine/threonine kinases [15,16], as well as into regulatory mechanisms that evolved uniquely in the EGFR [17], Eph [18], and Tec [19] families of tyrosine kinases.

In addition to the uniquely evolved features across different tyrosine kinase families, similarities across some tyrosine kinase families have also been noted. For example, the recently termed “Src module”, which consists of a tyrosine kinase domain and N-terminal SH3 and SH2 domains, is found across the Src, Abl, Tec, and Csk families, and structural and solution studies have determined that a similar autoinhibitory configuration of the Src module is shared across members of the Src, Abl, and Tec families [20]. Because previous classifications of protein kinases were determined by analyzing branching points in phylogenetic trees of diverse kinase domain sequences, along with analysis of common domain structures and known biological functions, the existence of evolutionarily related and functionally relevant higher order groupings of families within the kinase classification has not yet been systematically explored.

Here, we determine a novel hierarchical, constraint-based classification of the tyrosine kinome that newly identifies three evolutionary subgroupings of tyrosine kinase families based on the selective conservation of sequence motifs in the kinase domain, which encode common autoinhibitory conformations. In addition, we illustrate an evolutionary timeline of how unique kinase functions have expanded on shared subgroup-specific features through duplication events, evolutionary selection of family-specific motifs, and domain shuffling to give rise to the vast repertoire of tyrosine kinase signaling observed throughout metazoans. A closer examination of tyrosine kinase phylogeny in light of constraint-based tyrosine kinase subgroups reveals new

insights into the evolutionary conservation or divergence of subgroups, as well as the unique signaling features that may have emerged from three separate monophyletic clades of receptor tyrosine kinases. In particular, we note the early emergence of two major clades of holozoan tyrosine kinases distinguished by the presence (or absence) of an insert between the αD and αE helices of the kinase domain, where tyrosine kinases containing the insert comprise the majority of metazoan receptor tyrosine kinases. Our classification of the tyrosine kinome and the approach used in this study set a new precedent for the classification and evolutionary study of protein kinases and other large protein families.

3.2 Results

3.2.1 A hierarchical, constraint-based classification of the holozoan tyrosine kinome reveals new tyrosine kinase subgroups

To generate a comprehensive classification of the holozoan tyrosine kinome, we generated a multiple sequence alignment of 44,639 tyrosine kinase sequences spanning 586 species (see methods for details). We then used a Bayesian Partitioning with Pattern Selection (BPPS) algorithm to classify aligned sequences into hierarchical clusters based on the patterns of conservation and variation in aligned tyrosine kinase domain sequences **(Figure 3.1A)** [21,22]. Each cluster is distinguished by co-occurring sequence motifs which are highly conserved within the cluster, but strikingly different outside of the cluster. By sampling different clustering hierarchies and highly distinguishing sequence motifs, we defined an optimal hierarchy for holozoan tyrosine kinases based on the log-probability ratio (LPR) scores, which quantifies the contribution of conserved sequence patterns to the classification/clustering measured in natural units of information (nats) [21]. The total LPR score for the optimized holozoan tyrosine kinome

classification was 459825.35 nats, which is higher than the LPR score for KinBase classification of tyrosine kinases (443759.95 nats).

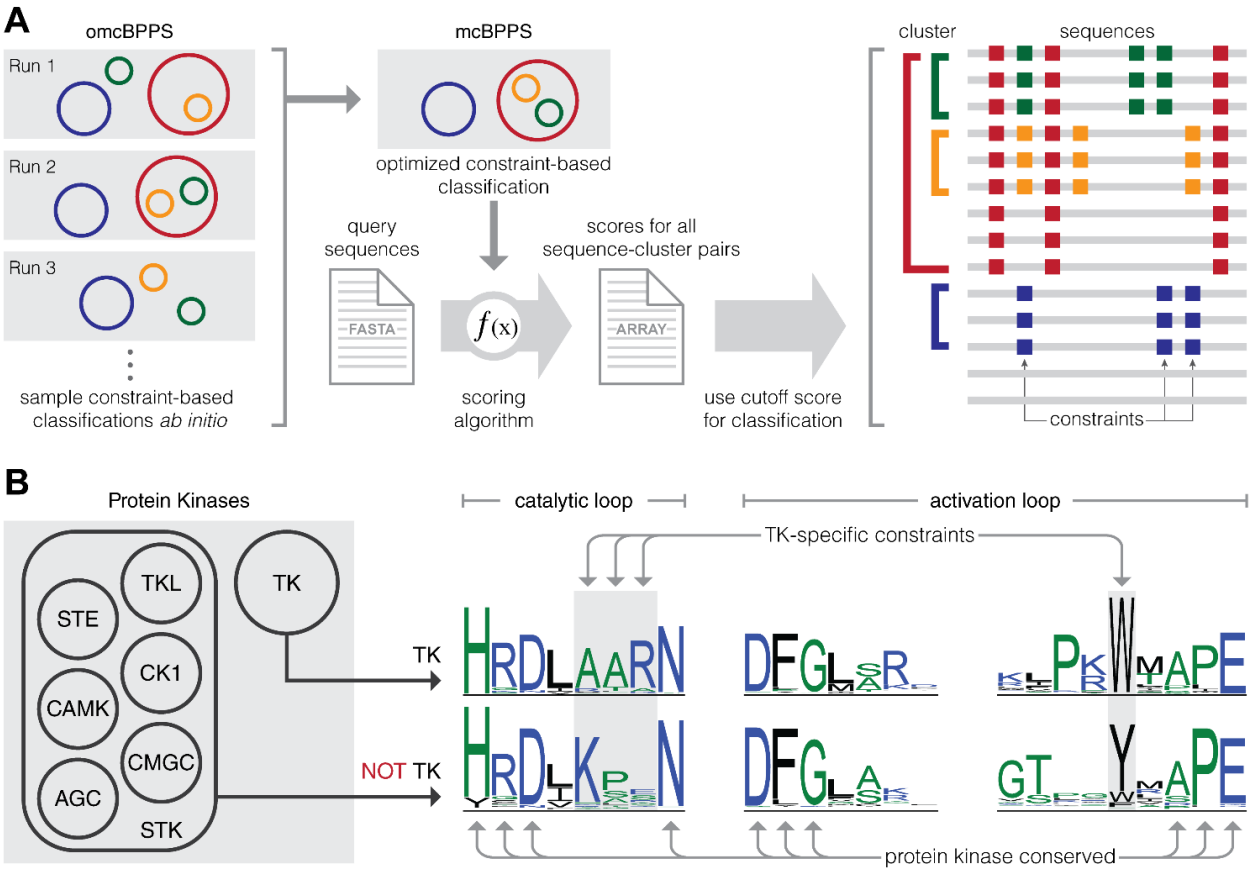


Figure 3.1: Workflow for constraint-based hierarchical clustering of tyrosine kinase sequences

(A) Multiple omcBPPS runs were used to sample various constraint-based hierarchical classifications for tyrosine kinases. An optimal constraint-based classification was determined based on LPR scores calculated using mcBPPS. Next, a scoring algorithm was used to score sequence-cluster pairs and to include or exclude tyrosine kinase sequences from each cluster defined in the optimal classification. An example constraint-based hierarchical classification is shown on the right. Clusters are represented as colored brackets, sequences are represented as grey lines, and constraints specific to each cluster are represented by squares colored according to the cluster to which they belong. For example, sequences in the green cluster share sequence constraints denoted by green squares, which are not found in sequences outside of the green cluster, as well as sequence constraints denoted by red squares, which are not found in sequences outside of the red cluster. The last two sequences are not included in any cluster as they lack any of the cluster-specific constraints defined in the constraint-based classification. (B) A visual

representation of cluster-specific constraints is shown using previously published data on tyrosine kinase-specific constraints [16]. Seven clusters of protein kinases are shown on the left, where tyrosine kinases are clustered separately from other protein kinases. A sequence logo of tyrosine kinase sequences is shown alongside a sequence logo of all other protein kinases. Sequence motifs such as HRD, DFG, and APE are conserved throughout all protein kinases, whereas the catalytic loop AAR motif and the activation loop tryptophan are defined as tyrosine kinase-specific constraints because their conservation is specific to the tyrosine kinase cluster.

Next, we re-classified 34,954 tyrosine kinase sequences from the UniProt reference proteomes database into the optimized hierarchy by quantifying the extent to which individual sequences match cluster-specific motifs (see methods for details). We define this re-classification as a constraint-based classification because this post-processing step eliminates spurious or divergent sequences from clusters that do not score over an optimal cut-off score due to their lack of cluster-specific patterns. The spurious sequences eliminated from each cluster are categorized within the Unclassified family **(Figure 3.2)**.

The new constraint-based hierarchical classification of tyrosine kinases, which is broadly similar to the KinBase classification of the tyrosine kinome, reveals several novel sub-groupings. In particular, the constraint-based classification defines three new subgroupings of tyrosine kinase families which account for nearly half of the tyrosine kinome: the Src Module (SrcM) subgroup, the Insulin Receptor Kinase-Like (IRKL) subgroup, and the Fibroblast, Platelet-derived, and Vascular growth factor Receptors (FPVR) subgroup **(Figure 3.2)**. The SrcM subgroup differs significantly from the KinBase classification in that it clusters the SrcA, SrcB, and Frk subfamilies of the KinBase-defined Src family within the same subgroup as the Tec and Abl families. Furthermore, SrcM does not include the SRM and sponge-specific Src (Src-Aque1) families. The FPVR subgroup includes seven distinct receptor tyrosine kinase families, where a Platelet-derived, and Vascular growth factor Receptor (PVR) subgroup sub-classifies the

VEGFR, PDGFR, Kit, CSF1R, and Flt3 families as a distinct cluster separate from the FGFR and Ret families. Notably, the new classification separates the KinBase PDGFR family into four families (PDGFR, Kit, CSF1R, and Flt3) due to statistically significant sequence constraints that define each of these families, warranting their designation as distinct families. The IRKL subgroup is the largest subgroup, comprising roughly 16% of tyrosine kinase sequences, and encompasses nine receptor tyrosine kinase families, including the insulin receptor kinase family as well as other poorly studied tyrosine kinases such as the CCK4 family of pseudokinases [23,24] and the Lmr family which exhibits serine/threonine kinase activity despite its placement into the tyrosine kinase clade [25,26]. We note that some organism-specific tyrosine kinase families defined in the KinBase classification, such as the unique receptor tyrosine kinase families in choanoflagellates (e.g. RTKA, RTKB, etc.) [4,27], were not detected due to the limited number of detectable homologs in current sequence databases (see methods).

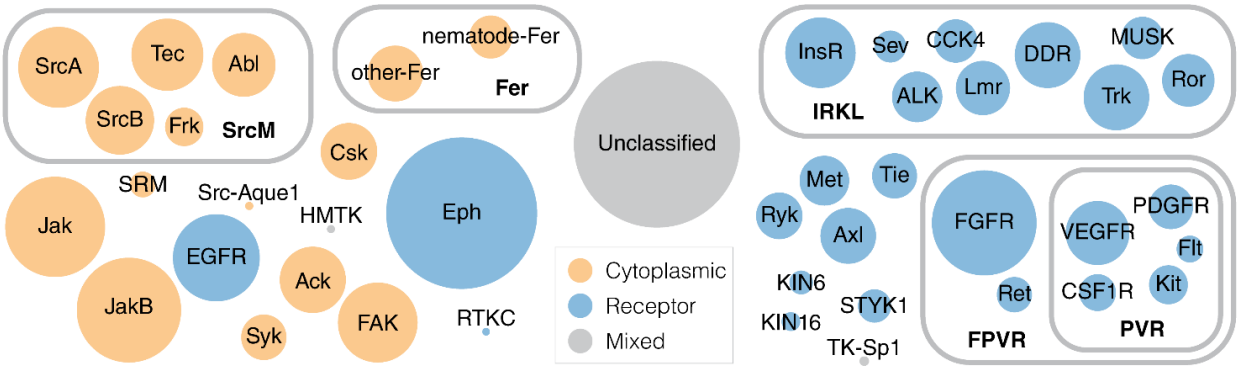


Figure 3.2: An evolutionary constraint-based hierarchical classification of the tyrosine kinome

The constraint-based hierarchical classification of tyrosine kinases is depicted as a Euler diagram. Each circle represents a distinct cluster of tyrosine kinases and is scaled to the number of sequences in each cluster. Clusters containing cytoplasmic tyrosine kinases are indicated with orange circles, while clusters containing receptor tyrosine kinases are indicated with blue circles. Clusters containing both cytoplasmic and receptor tyrosine kinases are colored grey.

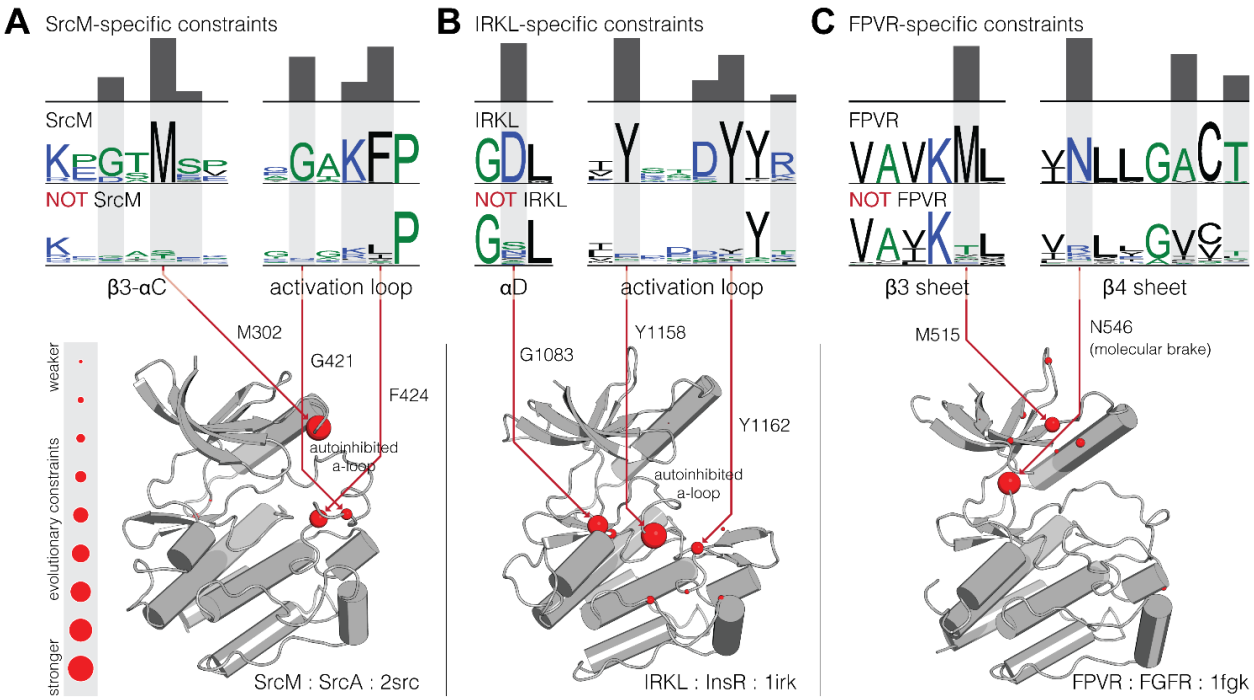


Figure 3.3: Structural locations of sequence motifs defining the SrcM, IRKL, and FPVR subgroups

Comparative sequence logos and structural mappings of subgroup-specific motifs are shown for the (A) SrcM, (B) IRKL, and (C) FPVR subgroups of tyrosine kinases. Sequence logos for the strongest evolutionary constraints corresponding to each subgroup are shown on top, with comparative sequence logos for sequences outside each subgroup provided below. Evolutionary constraints are highlighted in gray, with the height of the histogram reflecting the degree of divergence at that position between the subgroup and sequences outside the subgroup. Evolutionary constraints are shown as red balls on representative inactive structures of SrcM (Src) [28], IRKL (IRK) [32], and FPVR (FGFR1) [88]. The size of the red balls represents the strength of the evolutionary constraint. The strongest constraints are labeled with residue numbers corresponding to the position in the representative structures.

3.2.2 Subgroup-specific motifs localize to known autoregulatory sites in the kinase domain

By examining the sequence constraints that define each of the three novel subgroups in light of existing crystal structures, we observe that subgroup-specific motifs are located in known regulatory regions of the kinase domain. For example, the SrcM subgroup conserves a highly distinguishing GxM motif in the β3-αC loop and a GxKF motif in the activation loop that both form important interactions associated with a common Src-like inactive conformation in the activation loop **(Figure 3.3A)** [20,28]. This Src-like inactive conformation has been observed across diverse SrcM families such as SrcA [28], SrcB [29], Abl [30], and Tec [31], and similarities between their inactive structures have been previously noted [20]. That SrcM-specific sequence motifs are located in key regions in the Src-like inactive conformation, suggesting that these residues play key roles the conformational control of SrcM kinase activity. Likewise, the strongest sequence constraints on IRKL tyrosine kinases are associated with a common autoinhibitory conformation of the activation loop **(Figure 3.3B)**, which has been observed across crystal structures of diverse IRKL members [26,32–34], and is distinct from the autoinhibitory activation loop conformation of SrcM tyrosine kinases. Lastly, the FPVR subgroup of tyrosine kinases is defined by a highly conserved asparagine in the hinge region of the kinase domain **(Figure 3.3C)**, which engages an autoinhibitory 'molecular brake' [35] shared across these kinases. Other FPVR-specific sequence motifs are structurally located near the juxtamembrane, which is an important regulatory segment for many receptor tyrosine kinases [36], and may play common structural and functional roles in juxtamembrane-mediated regulation across FPVR tyrosine kinases.

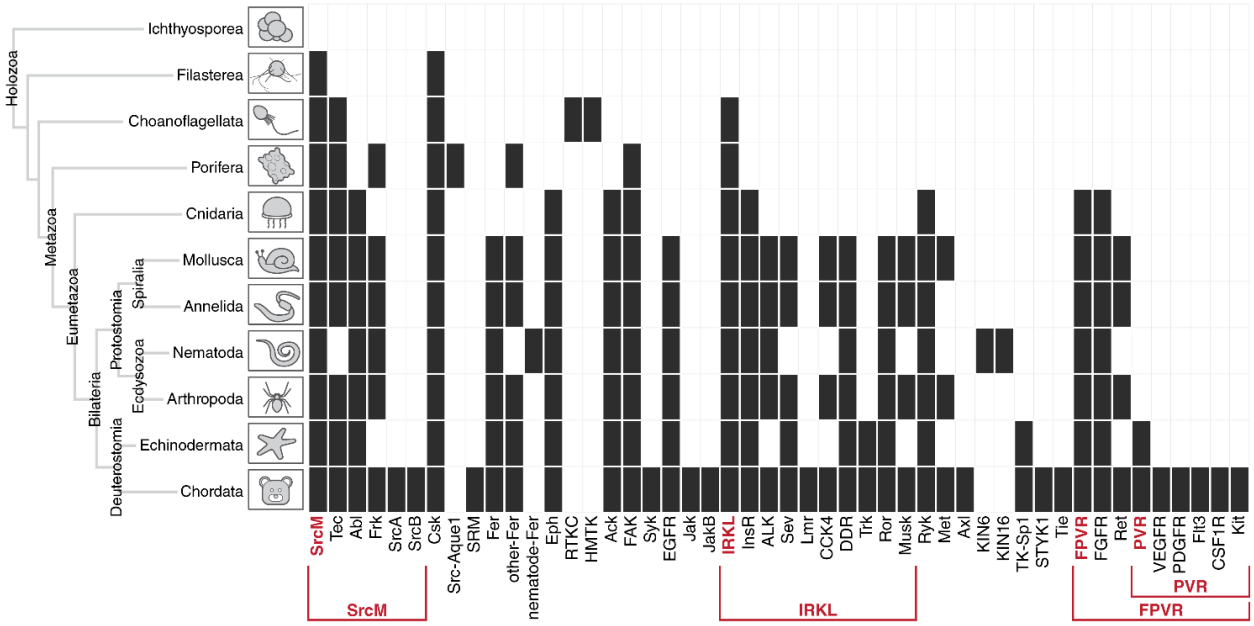


Figure 3.4: Emergence of tyrosine kinase subgroups and families throughout diverse holozoan taxa

The detection of each tyrosine kinase subgroup and family defined in the constraint-based classification is shown across diverse holozoan taxa, including single-celled relatives of metazoa. Constraint-based tyrosine kinase subgroups and families are shown across the x-axis, and diverse holozoan taxa, and their evolutionary relationships are shown on the y-axis. Cells are marked black if one or more members of a subgroup or family could be detected within a given taxa.

3.2.3 Tyrosine kinase subgroups are anciently conserved across diverse holozoan taxa

In order to infer when each of these tyrosine kinase subgroups and families emerged in evolution, we organized tyrosine kinase subgroups and families based on taxonomic conservation **(Figure 3.4)**. Tyrosine kinases from the SrcM subgroup, IRKL subgroup, and Csk family are detected across the most diverse holozoan taxa, including in unicellular relatives of metazoans (pre-metazoans) such as filastereans and choanoflagellates. The conservation of these early-emerging tyrosine kinases suggests that they likely played important roles in the evolution of metazoan multicellularity. The FPVR subgroup appears to have emerged later in eumetazoans,

following the divergence from early metazoans such as poriferans, which lack the organized tissues observed across eumetazoans. Interestingly, the PVR subgroup within the FPVR subgroup evolved much later in metazoan evolution and can only be detected in deuterostomes. The fact that tyrosine kinases from the SrcM, IRKL, and FPVR subgroups emerged in early stages of metazoan evolution suggests that these tyrosine kinases, their defining sequence motifs, and the regulatory functions associated with the motifs **(Figure 3.3)** were important in the evolution of metazoan morphologies such as multicellularity and organized tissues. We also note that, as found in previous kinome studies, our constraint-based classification defines several organism-specific tyrosine kinase families, such as the sponge-specific Src-Aque1 family [13], the nematode-specific KIN6 and KIN16 families [12], and the choanoflagellate-specific HMTK and RTKC families [4] **(Figure 3.4)**.

3.2.4 Domain shuffling contributed to diverse functions of the SrcM, IRKL, and FPVR kinase domains

In order to further explore the functional diversity of SrcM, IRKL, and FPVR tyrosine kinases, we surveyed the diversity of protein domains present across these subgroups and analyzed their conservation across holozoan taxa **(Figure 3.5)**. As previously noted, the SrcM tyrosine kinases, as well as the SRM, Csk, and Src-Aque1 families of tyrosine kinases share a core SH3-SH2-kinase domain organization, an anciently conserved [20], co-evolving unit [37] which can be detected in SrcM orthologs in unicellular pre-metazoans such as choanoflagellates and filastereans **(Figure 3.5D)**. The Tec family domain architecture, which includes an N-terminal lipid-targeting PH-domain (except in the Tec family member Txk), can be detected in choanoflagellates and filastereans, therefore the SH3-SH2-kinase and PH-SH3-SH2-kinase domain architectures represent the most anciently conserved tyrosine kinase domain structures.

Because sequence motifs defining the SrcM subgroup and the Tec family are also anciently conserved **(Figure 3.4)**, we suggest that these motifs have co-evolved with the SH3-SH2 and PH domains, respectively, and play key functional roles that link the kinase domain with their associated domains. Interestingly, the Abl family domain architecture, which includes a C-terminal F-actin binding domain appended to the SH3-SH2-kinase domain structure, emerged later in metazoan evolution after the divergence of bilaterians from other eumetazoans. However, evolutionary sequence constraints that distinguish the Abl family kinase domain from other SrcM members emerged earlier in metazoan evolution after the emergence of eumetazoans **(Figure 3.4)**. This suggests that Abl-specific functions of the kinase domain, perhaps substrate-specificity or Abl-specific regulation of catalysis, predated the additional functions imparted by the F-actin binding domain.

The IRKL and FPVR subgroups encompass the majority of receptor tyrosine kinase families and, despite sharing common kinase domain mechanisms within these subgroups **(Figure 3.3)**, have diversified functions through the incorporation of various extracellular domains. In fact, the diverse assortment of extracellular domains architectures found throughout the IRKL subgroup may be consistent with extracellular domain shuffling throughout holozoan evolution. Interestingly, the emergence of family-specific extracellular domains often precedes the emergence of family-specific motifs that define receptor tyrosine kinase families. For example, the extracellular fibronectin domain, which is associated with InsR and Sev family receptor tyrosine kinases, can be detected in diverse holozoan taxa, from chordates to choanoflagellates **(Figure 3.5D)**, however, InsR and Sev-specific sequence motifs in the kinase domain emerged later in metazoan evolution during the emergence of eumetazoans and bilaterians, respectively **(Figure 3.4)**. Similarly, Frizzled cysteine-rich (CRD-FZ), Coagulation

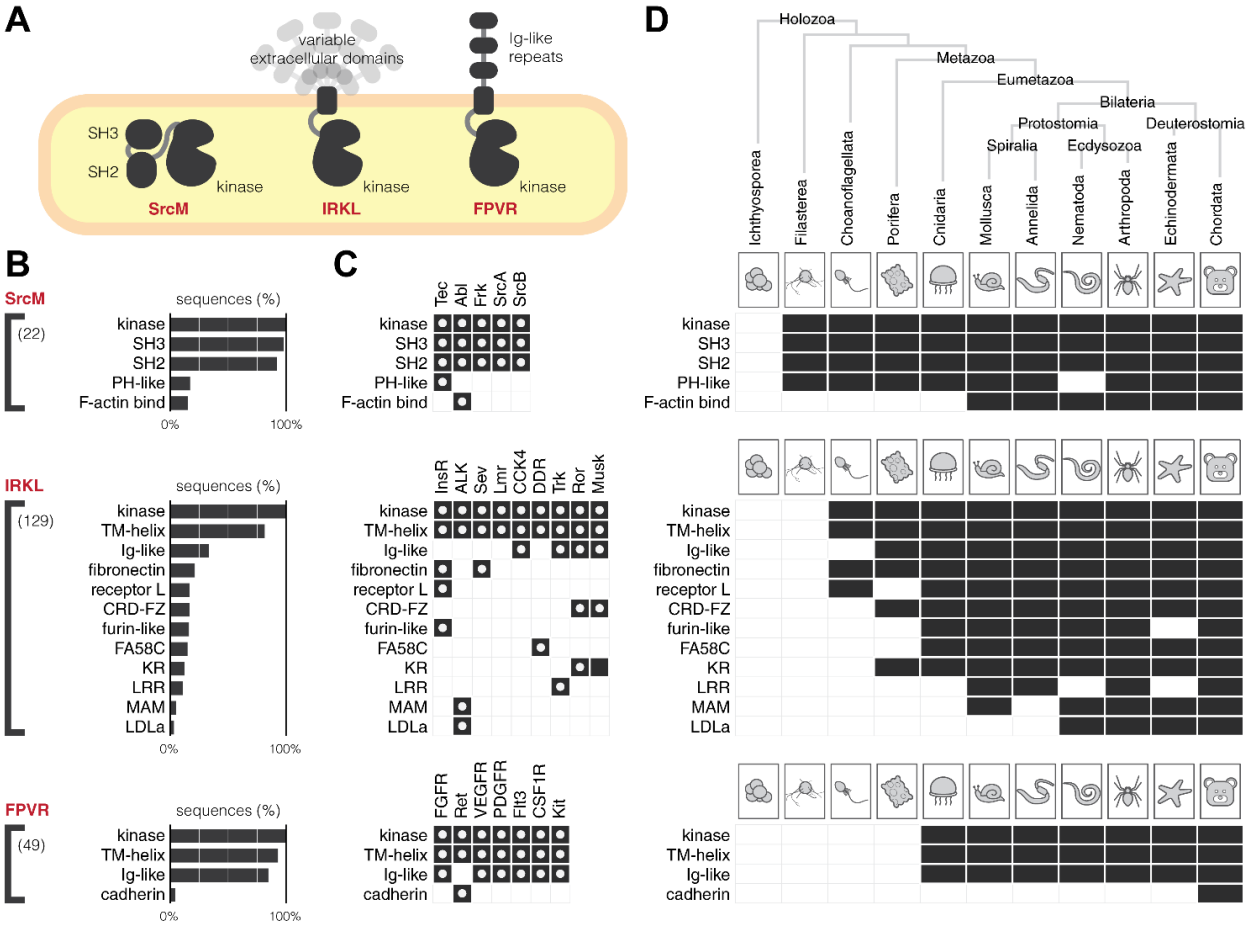


Figure 3.5: Protein domains associated with SrcM, IRKL, and FPVR tyrosine kinases
(A) A graphical depiction of common protein domain architectures observed across the SrcM, IRKL, and FPVR subgroups. (B) Bar graphs show the frequencies of protein domains detected in SrcM, IRKL, and FPVR sequences. Protein domains that occur in at least 3% of sequences are shown. The value in parentheses denotes the total number of unique protein domains found for sequences of a given subgroup. Consecutive repeat domains have been compressed into a single domain for simplicity. (C) Protein domains found across individual tyrosine kinase families within the SrcM, IRKL, and FPVR subgroups. White dots indicate domains found in human tyrosine kinases in each family. (D) The conservation of protein domains associated with SrcM, IRKL, and FPVR tyrosine kinases across diverse holozoan taxa.

factor 5/8 C-terminal (FA58C), and leucine-rich repeat (LRR) domains, which are associated with the Ror/Musk, DDR, and Trk families, respectively, emerged before their respective family-specific motifs. The evolutionary emergence of extracellular domains before family-specific motifs in the kinase domain suggests that evolution first diversified extracellular ligand binding before the fine-tuning of family-specific kinase domain functions such as downstream substrate specificity, regulation of catalysis, or intracellular protein-protein interactions. In contrast, in the ALK and Ret families of receptor tyrosine kinases, which uniquely contain LDLa and cadherin extracellular domains, respectively, the emergence of their family-specific sequence motifs predated the addition of their distinctive extracellular domains. These cases suggest that unique family-specific functions in the intracellular portion of receptor tyrosine kinases can also be expanded upon by subsequent shuffling of extracellular domains such that intracellular signaling functions are newly adapted to alternative extracellular ligands. Generally, despite the high degree of conservation of family-specific core domain structures **(Figure 3.5)**, the extreme diversification of SrcM, IRKL, and FPVR kinase domains across holozoans is evident in the huge number of unique protein domains that can be detected across sequences.

3.2.5 A representative phylogeny of the holozoan tyrosine kinome reveals new insights into tyrosine kinase evolution

To better understand the evolutionary relationships between tyrosine kinase subgroups and families, we constructed a phylogenetic tree using maximum likelihood, which models the natural process of sequence variation and finds a tree that best describes the evolutionary history of diverse protein sequences (see methods for details). By integrating our constraint-based classification of tyrosine kinases with our phylogenetic tree **(Figure 3.6A)**, we can now infer the evolutionary history of sequence constraints imposed on tyrosine kinase subgroups and families

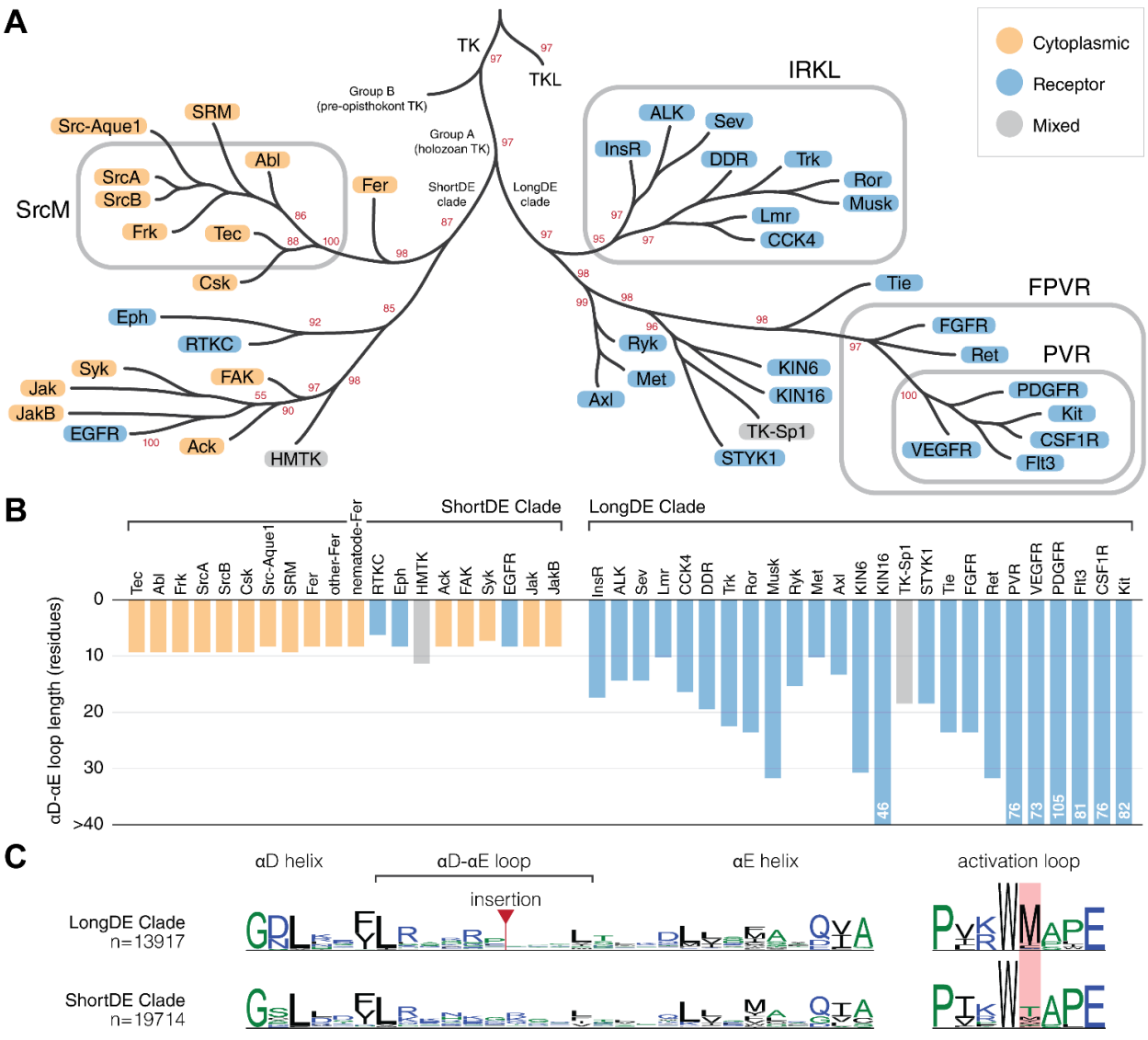


Figure 3.6: Evolution of sequence constraint-defined subgroups/families in the holozoan tyrosine kinome

(A) An abridged depiction of our representative phylogeny for the holozoan tyrosine kinome. Branch tips represent clades of sequence constraint-defined families. Cytoplasmic tyrosine kinase families are indicated with orange circles, while receptor tyrosine kinase families are indicated with blue circles. Subgroups are indicated by rounded rectangles which encompass their respective constituents. Bootstrap support values for select clades are shown in red. Branch lengths are not drawn to scale. (B) A bar chart showing the median αD-αE loop length of each tyrosine kinase family, which is shown on the x-axis separated by LongDE or ShortDE clade. On the y-axis, loop length is truncated at 30 residues. Any αD-αE loop lengths surpassing this limit are designated as >30. (C) Comparative sequence logos show differences between ShortDE and LongDE clade kinases.

defined in our constraint-based classification. For example, the IRKL, FPVR, and PVR subgroups, which were defined in our constraint-based classification due to their conservation of a set of subgroup-specific motifs, each form monophyletic clades in the phylogenetic tree, demonstrating that all tyrosine kinases within these subgroups have both maintained their respective subgroup-specific motifs and have descended from a common evolutionary ancestor. In contrast, the SrcM subgroup does not form a monophyletic clade in the phylogenetic tree. Instead, families within the SrcM subgroup share a monophyletic clade with the SRM, Src-Aque1, and Csk families, all of which likely descended from a common ancestor which conserved SrcM-specific motifs. However, that the SRM, Src-Aque1, and Csk families were not included in our constraint-based definition of the SrcM subgroup signifies that these families independently diverged from the rest of the SrcM subgroup through variations in the canonical SrcM-specific motifs, as well as accumulating additional variations contributing to functional divergence. Furthermore, examining SRM-specific, Src-Aque1-specific, and Csk-specific motifs unique to each of these families alongside SrcM subgroup-specific motifs will reveal how a common SH3-SH2-kinase domain organization **(Figure 3.5)** [20] has diverged along these various lineages to innovate divergent regulatory functions on a shared domain architecture. Our phylogenetic tree also confirms, along with our constraint-based classification, that the Lmr family belongs within the IRKL subgroup/clade despite detectable serine/threonine activity. Further sequence analysis shows that the Lmr kinases have regained a serine/threonine kinase-specific histidine (LMTK3^{His260}) in the αE helix which tyrosine kinases selectively lost upon diverging from the serine/threonine kinases [16]. The reemergence of this histidine may explain why Lmr kinases have regained serine/threonine activity. In addition, the evolutionary

relationships described by our phylogeny of the holozoan tyrosine kinome are independently support by conserved intron and phase positions in the kinase domain [38,39].

A further examination of our phylogenetic tree of tyrosine kinases reveals several new insights about tyrosine kinase evolution. As noted in previous phylogenetic studies of tyrosine kinase sequences, tyrosine kinases form a monophyletic clade separate from the closely related tyrosine kinase-like (TKL) group of serine/threonine kinases. Holozoan tyrosine kinases (Group A) also form a monophyletic clade that is distinct from a paraphyletic group of divergent tyrosine kinase sequences found in pre-opisthokonts (Group B), such as those found in the amoebozoans *Dictyostelium discoideum* or in the green algae *Chlamydomonas reinhardtii* **(Figure 3.6)** [1]. We note for the first time another major branching point in the evolution of tyrosine kinases which is associated with the presence or absence of an insert between the αD and αE helices of the kinase domain that we refer to as the DE insert, historically referred to as the kinase insert domain [40]. Although the DE insert segment was not explicitly used in building the phylogenetic tree due to lack of detectable sequence similarity in this region across families, the phylogenetic tree exhibits a clear division of two clades, one containing kinases with a short αD-αE loop, which we refer to as the shortDE clade, and one predominantly containing kinases with a long αD-αE loop, which we refer to as the longDE clade **(Figure 3.6A)**. The evolutionary separation of the longDE clade is also independently supported by a common phase-2 intron at the αH helix. While the variation in the length of the DE insert has been previously noted **(Figure 3.6B)**, and previous studies have shown the functional significance of the insert on downstream signaling and kinase activation [40,41], the evolutionary history of the DE insert has not been examined.

In light of the evolutionary divergence between longDE and shortDE clades, we also note that the longDE clade primarily contains receptor tyrosine kinase families, while the shortDE clade contains predominantly cytoplasmic receptor tyrosine kinases (with the exception of the EGFR, Eph, and choanoflagellate-specific RTKC families). This correlation between the presence of the longDE insert with the presence of transmembrane and extracellular domains, alongside evidence that the insert plays important roles in kinase activation and protein recruitment for downstream signaling, suggests that the longDE insert evolved as a means to facilitate downstream intracellular signaling upon the activation of receptor tyrosine kinases by extracellular signals. In addition, though the DE insert is difficult to align across families due to the lack of sequence conservation, the DE insert is alignable within families and often conserves sequence motifs including phosphorylatable tyrosine, serine, or threonine residues, suggesting that individual receptor tyrosine kinase families along the longDE clade have rapidly and frequently evolved the longDE insert in family-specific contexts, presumably to carry out family-specific downstream signaling functions. We also note that the longDE tyrosine kinases highly conserve a unique activation loop methionine, which is not observed in shortDE tyrosine kinases **(Figure 3.6C)**; however, the role of this methionine, or whether it is significant for DE insert function or for receptor tyrosine kinase function is unknown.

Interestingly, the shortDE clade in the tyrosine kinase phylogeny, which predominantly consists of cytoplasmic tyrosine kinases, includes two monophyletic clades of receptor tyrosine kinases: the EGFR family of receptor tyrosine kinases and a separate monophyletic clade that includes the Eph and choanoflagellate-specific RTKC families of receptor tyrosine kinases. Thus, our phylogeny suggests at least three independent origins of highly expanded receptor tyrosine clades, with the majority of receptor tyrosine kinases emerging from the longDE clade.

That these disparate branches along the tyrosine kinase phylogeny have convergently evolved to include transmembrane and extracellular domains highlights the importance of relaying extracellular signals into intracellular responses across various signaling niches. While the longDE receptor tyrosine kinases are distinguished by extra functionalities imparted by the longDE insert, the Eph and EGFR families also exhibit unusual signaling functions so far unobserved in other longDE receptor tyrosine kinases. Eph receptor tyrosine kinases have a unique capacity for bi-directional signaling, where the binding of ephrin ligands, which are also membrane bound, can activate signaling both in the receptor-bearing cell, as well as in the ligand-bearing cell [42]. Furthermore, our tree suggests that the RTKC family may also share this unique capacity for bi-directional signaling. The EGFR family is also a unique family of receptor tyrosine kinases in that ligand binding induces dramatic conformational changes in the dimerization arm extracellularly, also inducing a unique allosterically activating dimer in the intracellular portion [43,44]. These mechanisms of EGFR family kinases, as well as their activating (rather than autoinhibitory) juxtamembrane and their lack of activation via activation loop phosphorylation distinguishes the EGFR family from receptor tyrosine kinases in the longDE clade [45,46].

3.3 Discussion

The classification of protein kinases into evolutionarily related families has provided the foundation for decades of comparative sequence-structure-function studies on protein kinases [8,11,47,48]. Here, we propose a new constraint-based classification of tyrosine kinases that newly defines the SrcM, IRKL, and FPVR subgroups **(Figure 3.2)** each of which maintains core subgroup-specific sequence motifs associated with subgroup-specific auto-inhibited conformations **(Figure 3.3)**. Subsequent taxonomic conservation analysis suggests that

expansion of tyrosine kinase subgroups and evolution of family-specific motifs within these subgroups, along with domain shuffling, elaborated on subgroup-specific functions to diversify cell signaling functions **(Figure 3.4, 3.5)**. However, these core regulatory motifs are likely conserved because they ensure that these kinases are activated at the right place and time. For instance, the strongest SrcM-specific constraint is a methionine in the αC-β4 loop which packs against the auto-inhibited activation loop conformation, suggesting an anciently conserved role in stabilizing the Src-like inactive conformation **(Figure 3.3A)** that may be relieved in various manners such as the binding of substrates to the co-evolved SH3-SH2 domains. Furthermore, the ancient conservation of Csk and SrcM-specific constraints supports the pre-metazoan origins of SrcM inhibition via C-terminal tail phosphorylation by Csk [49]. Further study of family-specific motifs across various lineages of tyrosine kinases is expected to reveal unique regulatory mechanisms across distinct tyrosine kinase families.

We constructed a new representative phylogenetic tree of the holozoan tyrosine kinome which revealed larger evolutionarily-related clades of tyrosine kinases associated with additional defining features **(Figure 3.6A)**. Dividing the holozoan tyrosine kinome into two roughly equal halves, the basal longDE and shortDE clades are distinguished by the presence or absence (respectively) of a fast-evolving kinase domain insertion in the αD-αE loop **(Figure 3.6B)**. Functionally important insertion regions shared by large groups of evolutionarily-related protein kinases have also been described in the CMGC group which conserves a kinase domain insertion between the αH and αI helices [50]. The diverse functions of the longDE insertion remains understudied; however, the region is documented to possess many functionally-important phosphorylation sites in multiple tyrosine kinase families [40]. Our tree also reveals three distinct evolutionary lines of receptor tyrosine kinases: longDE, RTKC-Eph, and EGFR, each of which

are distinguished by unique lineage-specific variations on receptor tyrosine kinase signaling and regulation [40,42,46]. Overall, the definition of these evolutionarily-related tyrosine kinases will enable the inference of sequence-structure-function relationships in the understudied kinases based on the known functions of well-studied kinase families.

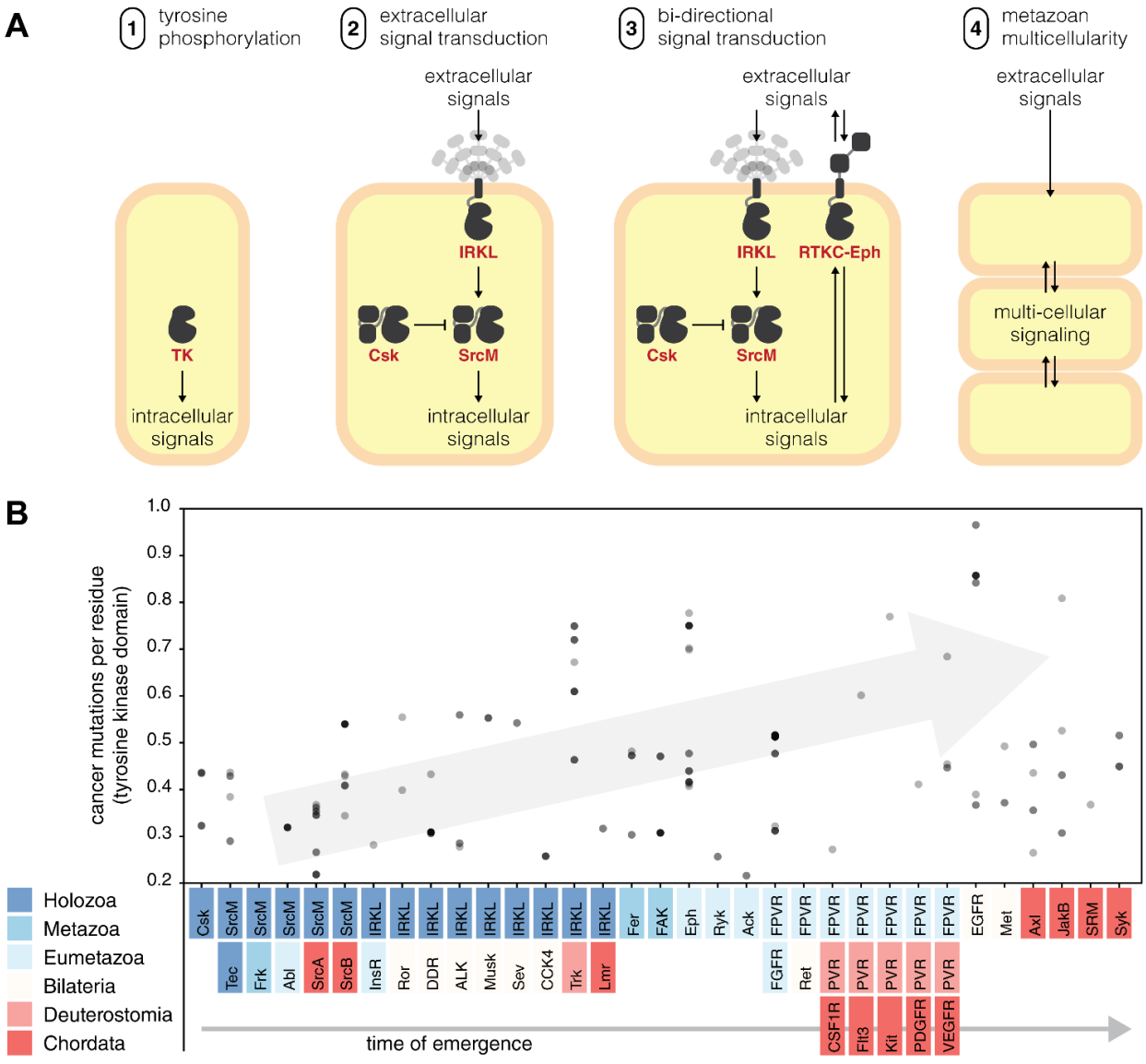


Figure 3.7: Tyrosine kinases in the evolution of multicellularity and cancer
(A) We propose a series of evolutionary innovations of the tyrosine kinome signaling which potentially contributed to the emergence of metazoan multicellularity. (B) Disease-related mutations tend to occur in more recent tyrosine kinase families. A scatter plot shows how

frequently human tyrosine kinases are found to be mutated in genome-wide cancer sequencing studies from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database [89]. On the y-axis, mutation frequency is measured by the average number of mutations per residue in the tyrosine kinase domain. On the x-axis, each kinase is sorted into a cluster defined by our hierarchy and sorted by their time of emergence. All points were drawn with transparency; as a result, points which appear darker indicate multiple points falling in the same location.

The expansion and diversification of the tyrosine kinome across the animal kingdom highlights its central role in metazoan biology. While many previous studies have speculated on the role of tyrosine kinases in the evolution of multicellularity [5,39,51,52], our findings suggest key evolutionary innovations which likely contributed to the adoption of tyrosine kinase signaling for multicellular functions **(Figure 3.7A)**. While elaborate tyrosine kinase signaling networks have been discovered in unicellular pre-metazoans, they generally display low orthology to tyrosine signaling networks in metazoans [27]. Our analyses identify sparse similarities between pre-metazoan and metazoan tyrosine kinase signaling in that SrcM, Tec, Csk, and IRKL tyrosine kinases originated in pre-metazoans and have remained conserved throughout diverse metazoan taxa **(Figure 3.4)**. These components may represent a core phospho-tyrosine signaling machinery that have been expanded through the addition of taxa-specific tyrosine kinases to suit unique organismal needs. We further speculate that additional capacity for bi-directional signaling emerged in the RTKC-Eph lineage in an extinct ancestral pre-metazoan organism, as previous studies have identified distant Eph orthologs in various pre-metazoan organisms [6,53] which are not shown in our conservation table because they lack Eph-specific sequence constraints. These core signaling functions likely enabled the evolution of primordial Metazoa, a mass of cells capable of moving and growing in a coordinated fashion, lacking much of the complexity found in modern day metazoans. Throughout the course of metazoan evolution, the emergence of new complex biological systems such as the nervous

system, circulatory system, and adaptive immunity appears to coincide with the emergence of functionally associated tyrosine kinase families **(Table 3.1)**. Additional complexity was gained through whole genome duplication and small-scale duplication events which played a major role in tyrosine kinase evolution in vertebrates, especially the PVR kinases whose evolution has been heavily influenced by both types of duplication events [38,39,54,55].

family / subgroup	kinase functions	time of emergence	predicted association to metazoan evolution	references
FAK	dynamically regulation of cell adhesion and motility	Porifera (Metazoa)	organized cell layer (found in all metazoans)	[76,77]
Eph	regulation of cell migration during development	Cnidaria (Eumetazoa)	organized germ layers (emerged in cnidarians)	[78]
FGFR	regulation of early metazoan development, such as gastrulation and neural induction	Cnidaria (Eumetazoa)	simple nervous system (emerged in eumetazoans)	[79–81]
FPVR	many FPVR kinases play roles in vasculature-related functions	Cnidaria (Eumetazoa)	vasculature (emerged in bilaterians)	[11]
EGFR	regulated cell growth via the EGFR pathway	Bilateria	EGFR pathway (emerged in bilaterians)	[82]
Ror, Musk, DDR, Sev, ALK, CCK4	many IRKL kinase families play roles in neural development	Bilateria	true brain characterized by neuropile (emerged in bilaterians)	[11,83]
VEGFR, PDGFR, Kit, CSF1R, Flt3	many PVR kinase families play roles in blood vasculature-related functions and are implicated in blood cancers	Chordata	closed circulatory system lined with endothelium (emerged in vertebrates)	[84,85]
JAK, JAK-b, Syk	immune signaling via the JAK-STAT pathway	Chordata	adaptive immunity (emerged in vertebrates)	[86,87]

Table 3.1: The emergence of tyrosine kinase families is associated with the metazoan phenotypes

The taxa in which many tyrosine kinase families first emerged **(Figure 3.4)** is correlated with the emergence of various taxa-specific phenotypes which are associated with functions performed by the corresponding tyrosine kinase family. Families whose functions do not seem to exhibit a clear connection to the emergence of a metazoan phenotype are not shown.

Given the important roles of tyrosine kinases in multicellular metazoan biology, it comes as no surprise that sequencing efforts have identified many disease-related variants across the tyrosine kinome [56]. Multiple studies have proposed reversal of cancer cells to a more unicellular-like state through the release of "multicellular constraints" [57]. While many oncogenes predate the origin of multicellularity, evolutionary studies have identified a second burst of oncogene emergence which co-occurred with the appearance of multicellular metazoans, and is comprised of genes (including tyrosine kinases) involved in cellular signaling and growth processes [58]. In support of this finding, our data offers a closer look at this event, suggesting that cancer variants occur more frequently in recently-evolved tyrosine kinases **(Figure 3.7B)**. Overall, the deep connection between phospho-tyrosine signaling and cancer encourages further studies on the tyrosine kinome and the unique sequence constraints that guide their evolution and function. Focused analyses are expected to reveal new insights on metazoan multicellular signaling and its malfunctions in disease states.

3.4 Methods

3.4.1 Evolutionary constraint-based clustering

We sampled alternative classification hierarchies *ab initio*, using the omcBPPS algorithm [22] which employs a Markov Chain Monte Carlo (MCMC) sampling strategy to classify aligned sequences into hierarchical categories/clusters based on shared constraints, i.e. slow evolving sites. Category/cluster-specific constraint refers to alignment positions that are highly conserved within a given cluster, but divergent in sequences outside of the cluster. The omcBPPS algorithm iteratively optimizes for two interdependent criteria: (1) which alignment positions should be

defined as cluster-specific constraints and (2) what is the optimal hierarchical clustering scheme based on cluster-specific constraints.

We ran omcBPPS on two different sequence sets from UniProt reference proteomes (retrieved 2/13/2019) and NCBI non-redundant (nr) proteins database (retrieved 2/13/2019) [59]. Within these sequence sets, we identified and aligned tyrosine kinase sequences using the Multiply-Aligned Profiles for Global Alignment of Protein Sequences (MAPGAPS) algorithm [60]. This alignment was restricted to the protein kinase domain starting from the β1 strand (PKA^{Phe43}) and ending at the αI helix (PKA^{Lys292}). Kinase sequences which did not span from at least the β3-Lys to the DFG-Asp were deemed fragmentary and removed from the alignment. The UniProt sequence set contained 12,137 tyrosine kinase sequences. The nr sequence set was further purged at 98% sequence identity and contained 17,071 tyrosine kinase sequences. We performed hierarchical clustering on both sequence sets using omcBPPS. For both sequence sets, we optimized the "minnats" parameters (minnat=1 and minnat=5) which changed the minimum log-likelihood required to form a cluster. All runs were performed twice. To create a consensus of the hierarchical classification schemes found by multiple runs of omcBPPS, we used the mcBPPS algorithm. Clusters which were consistently identified throughout multiple runs were refined using the mcBPPS algorithm [21]. We ran mcBPPS on a maximally diverse set of 33,769 sequences containing all tyrosine kinase detectable from nr to generate an optimal model that is consistent with the existing data.

3.4.2 Fitting new sequences to a constraint-based clustering model

Using our consensus model, we developed quantitative means of evaluating how well any given sequence fit into each of the clusters defined by our model. Within our model, each cluster

was associated with a list of constraints that dictated which amino acid(s) were likely to be found at a given alignment position. Furthermore, each constraint was associated with a log-likelihood score which described how specific a constraint was to its respective cluster. To score a sequence against a cluster, we added the log-likelihoods of all the constraints which were true for the query sequence. In order to make this score comparable across clusters, we divided this number by the total log-likelihood of all the cluster's constraints. This resulted in a range from 0 to 1. For example, a sequence which followed all of a given cluster's constraints would have received a maximum score of 1 for that cluster, while a sequence which did not follow any of a given cluster's constraints would have received a minimum score of 0 for that cluster. For the purposes of discrete classification, we defined a cut-off score for classifying a sequence into a cluster. Based on the distribution of scores from all possible sequence-cluster pairs across multiple test datasets, we defined the optimal cut-off at the global minima of 0.7.

We scored a representative set of sequences from UniProt proteomes (retrieved 4/2/2020) and estimated the size of each cluster. A representative set of 44,639 tyrosine kinase sequences spanning 586 species were identified and aligned to a common alignment profile using the MAPGAPS algorithm [60]. We determined the size of each cluster based on how many sequences scored above the cut-off. We also quantified the similarity between clusters by evaluating how well sequences in a given cluster scored against all other clusters using an all-vs-all comparison. We defined a non-symmetric similarity metric for any two given clusters, A and B, as the average score of cluster A sequences when classified against cluster B constraints. As a consequence of our cut-off, the average score when A and B were the same cluster was always greater than 0.7 which we observe across the diagonal. Indicative of our hierarchical classification scheme, we also observed a distinct signature for subclusters (such as Tec) which

were classified under a larger supercluster (such as SrcM): subclusters would score highly (>0.7) for the constraints of its respective supercluster but not vice versa. Our comparison matrix also indicated that evolutionarily-related sequences within clusters share more constraints in common than sequences outside of the cluster. Finally, we observed that the majority of values outside of the diagonal were quite low which indicated that our model defined well-separated sequence clusters, each defined by their own unique sequence constraints.

3.4.3 Comparative sequence analysis

All comparative sequence analyses were performed in Python 3 using the HelperBunny library (provided with our computational notebooks) which implements NumPy array-based sequence alignment manipulation. All sequence features (including features pertaining to motifs, evolutionary constraints, domain composition, taxonomic descriptors, insertions, and deletions) were represented as Boolean arrays as a function of the full sequence alignment. More complex queries pertaining to multiple sequence features (such as the presence of a given domain in a given taxon) were constructed by Boolean algebra. These Boolean arrays were applied as filters to our sequence alignment using NumPy indexing routines [61]. Sequence logos were generated using the WebLogo 3 API [62].

3.4.4 Taxonomic conservation analysis

After we classified our representative set of tyrosine kinase into discrete clusters, we determined the taxonomic conservation of each cluster. We determined the source of each tyrosine kinase sequence using the organism identifier number (OX) provided in the FASTA header of UniProtKB sequences. OX numbers were traced back to their parent node identifiers using the nodes dump file provided in the NCBI taxdump database. All node identifiers were

translated to their respective scientific names using the taxonomy names dump file. We determined the distribution of taxa across each cluster of tyrosine kinases and selected an optimal mix to depict diverse taxa ranging from unicellular pre-metazoans to more complex metazoans such as chordates. Because our definition of protein family is based on sequences constraints, our methods may sometimes yield differential results compared to more traditional methods which utilize sequence homology [63,64]. We report taxonomic conservation using our optimized cut-off score of 0.7 **(Figure 3.4)** as well as a relaxed cut-off score of 0.6.

3.4.5 Intron and phase position analysis

Intron/exon annotations were mapped using Scipio [65]. Intron phases were determined by calculating the modulus of three (representing the codon size) of the cumulative sum of lengths of each exon which preceded an intron within an open reading frame. A phase-0 intron is located between two consecutive codons; a phase-1 intron is located between the first and second nucleotides of a codon; and a phase-2 intron is located between the second and third nucleotides of a codon.

3.4.6 Protein domain conservation analysis

We produced protein domain annotations for each full-length tyrosine kinase sequence using the NCBI Conserved Domain Database (CDD v3.18 - 55,570 PSSMs) database of conserved protein domains [66,67]. Queries to the database were programmatically submitted using the bwrpsb PERL script which was provided in the Batch CD-Search API. Search parameters included an expected value threshold of 0.01 with only the best scoring domain model being returned. Transmembrane (TM) helix annotations were identified and appended to our domain annotation results using TMHMM 2.0 [68]. Synonymous domain names were

manually identified and merged in post-processing. We combined these domain annotations with previously generated data to evaluate the conservation of protein domains across various constraint-defined clusters and taxonomic clades.

3.4.7 Phylogenetic analysis

We inferred the evolution of tyrosine kinase families/subgroups using a maximum-likelihood approach. In order to create a representative phylogeny of the holozoan tyrosine kinome, we sampled a taxonomically diverse set of sequences from each constraint-defined cluster of tyrosine kinase sequences. Our representative set of sequences consisted of (1) one randomly selected sequence from each cluster-taxon pair excluding Chordata, (2) all human tyrosine kinases sequences which represented the chordate taxon, (3) three early-diverging, pre-opisthokont tyrosine kinase sequences from amoebozoan, *Acanthamoeba castellanii* [1], and (4) eight human TKL group kinases which was used as an outgroup. We produced multiple representative sequence sets with different random samples of taxonomically diverse sequences. Using these sequence sets, we generated multiple phylogenies using IQTREE v1.6.11 [69], with ModelFinder [70]. Branch support values were generated using ultrafast bootstrap with 1000 resamples [71]. Results indicated that sequences from the same clusters consistently formed paraphyletic groups or monophyletic clades with high bootstrap support. We compared topologies generated from different random samples and found no major changes in the evolutionary relationships between evolutionary constraint-defined sequence clusters. Furthermore, our topology was robust to the inclusion of unclassified tyrosine kinase sequences.

The consensus tree with the highest bootstrap support values for the three major tyrosine subgroups was selected as the final tree. The optimal substitution model for our final topology

was determined to be LG+R8 based on the Bayesian Information Criterion as determined by ModelFinder [70]. We rooted our final tree against the TKL outgroup using ETE Toolkit v3.1.1 [72]. Consistent with previous studies, the most divergent tyrosine kinases in our tree were the paraphyletic pre-opisthokont "Group B" tyrosine kinases which diverged prior to the emergence of the monophyletic "Group A" holozoan tyrosine kinases [1]. We observed high concordance between our evolutionary constraint-based clustering and phylogenetic inference; this allowed us to simplify our tree topology by condensing monophyletic clades and pruning paraphyletic groups **(Figure 3.6A)**. The simplified representation describes evolutionary relationships between constraint-defined families. Furthermore, we represented each of the three major tyrosine kinase subgroups by drawing an enclosure around each subgroup's constituent families.

We compared our tree to previously published phylogenies which also sampled diverse tyrosine kinase families [1,8,14,38,73,74]. However, many of these studies focused on vertebrate, basal metazoan, and pre-metazoan kinase sequences leaving out many diverse protostome sequences. We observed key differences with the current widely-accepted phylogeny of the human tyrosine kinome [8] and found the most similarities to a tree published by Robinson et al [14]. Previously placed near the base of the tyrosine kinase clade, our placement of STYK1 was supported by common introns shared with Tie and various IRKL families, while our placement of Lmr was supported by common introns shared by various IRKL kinases [38]. Furthermore, the majority of human longDE kinases share a common phase-2 intron in the genomic region which codes for the αH helix. Overall, our tree has high support for all major clades, including historically difficult-to-place families such as JAK [75]. Providing additional support, we note that our tree is highly concordant with the evolutionary progression of holozoan

taxa in that anciently conserved tyrosine kinase families tend to diverge first, while more recently diverged tyrosine kinase families appear closer to the tips.

3.4.8 Data Availability

The data sets and code for our analyses are freely available for download from GitHub at: https://github.com/esbgkannan/holozoan_tk_evolution.

Bibliography

[1] H. Suga, M. Dacre, A. de Mendoza, K. Shalchian-Tabrizi, G. Manning, I. Ruiz-Trillo, Genomic Survey of Premetazoans Shows Deep Conservation of Cytoplasmic Tyrosine Kinases and Multiple Radiations of Receptor Tyrosine Kinases, Sci. Signal. 5 (2012) ra35–ra35. https://doi.org/10.1126/scisignal.2002733.

[2] T. Hunter, The Genesis of Tyrosine Phosphorylation, Cold Spring Harb Perspect Biol. 6 (2014) a020644. https://doi.org/10.1101/cshperspect.a020644.

[3] L.W. Parfrey, D.J.G. Lahr, A.H. Knoll, L.A. Katz, Estimating the timing of early eukaryotic diversification with multigene molecular clocks, PNAS. 108 (2011) 13624–13629. https://doi.org/10.1073/pnas.1110633108.

[4] N. King, M.J. Westbrook, S.L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K.J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J.B. Lyons, A. Morris, S. Nichols, D.J. Richter, A. Salamov, J.G.I. Sequencing, P. Bork, W.A. Lim, G. Manning, W.T. Miller, W. McGinnis, H. Shapiro, R. Tjian, I.V. Grigoriev, D. Rokhsar, The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans, Nature. 451 (2008) 783–788. https://doi.org/10.1038/nature06617.

[5] W.T. Miller, Tyrosine kinase signaling and the emergence of multicellularity, Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 1823 (2012) 1053–1057. https://doi.org/10.1016/j.bbamcr.2012.03.009.

[6] K. Tong, Y. Wang, Z. Su, Phosphotyrosine signalling and the origin of animal multicellularity, Proceedings of the Royal Society B: Biological Sciences. 284 (2017) 20170681. https://doi.org/10.1098/rspb.2017.0681.

[7] B.A. Liu, E. Shah, K. Jablonowski, A. Stergachis, B. Engelmann, P.D. Nash, The SH2 Domain–Containing Proteins in 21 Species Establish the Provenance and Scope of Phosphotyrosine Signaling in Eukaryotes, Sci. Signal. 4 (2011) ra83–ra83. https://doi.org/10.1126/scisignal.2002105.

[8] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The Protein Kinase Complement of the Human Genome, Science. 298 (2002) 1912–1934. https://doi.org/10.1126/science.1075762.

[9] S. Caenepeel, G. Charydczak, S. Sudarsanam, T. Hunter, G. Manning, The mouse kinome: Discovery and comparative genomics of all mouse protein kinases, PNAS. 101 (2004) 11707–11712. https://doi.org/10.1073/pnas.0306880101.

[10] C.A. Bradham, K.R. Foltz, W.S. Beane, M.I. Arnone, F. Rizzo, J.A. Coffman, A. Mushegian, M. Goel, J. Morales, A.-M. Geneviere, F. Lapraz, A.J. Robertson, H. Kelkar, M. Loza-Coll, I.K. Townley, M. Raisch, M.M. Roux, T. Lepage, C. Gache, D.R. McClay,

G. Manning, The sea urchin kinome: A first look, Developmental Biology. 300 (2006) 180–193. https://doi.org/10.1016/j.ydbio.2006.08.074.

[11] G. Manning, G.D. Plowman, T. Hunter, S. Sudarsanam, Evolution of protein kinase signaling from yeast to man, Trends in Biochemical Sciences. 27 (2002) 514–520. https://doi.org/10.1016/S0968-0004(02)02179-5.

[12] G.D. Plowman, S. Sudarsanam, J. Bingham, D. Whyte, T. Hunter, The protein kinases of Caenorhabditis elegans: A model for signal transduction in multicellular organisms, PNAS. 96 (1999) 13603–13610. https://doi.org/10.1073/pnas.96.24.13603.

[13] M. Srivastava, O. Simakov, J. Chapman, B. Fahey, M.E.A. Gauthier, T. Mitros, G.S. Richards, C. Conaco, M. Dacre, U. Hellsten, C. Larroux, N.H. Putnam, M. Stanke, M. Adamska, A. Darling, S.M. Degnan, T.H. Oakley, D.C. Plachetzki, Y. Zhai, M. Adamski, A. Calcino, S.F. Cummins, D.M. Goodstein, C. Harris, D.J. Jackson, S.P. Leys, S. Shu, B.J. Woodcroft, M. Vervoort, K.S. Kosik, G. Manning, B.M. Degnan, D.S. Rokhsar, The Amphimedon queenslandica genome and the evolution of animal complexity, Nature. 466 (2010) 720–726. https://doi.org/10.1038/nature09201.

[14] D.R. Robinson, Y.-M. Wu, S.-F. Lin, The protein tyrosine kinase family of the human genome, Oncogene. 19 (2000) 5548–5557. https://doi.org/10.1038/sj.onc.1203957.

[15] K. Oruganty, N.S. Talathi, Z.A. Wood, N. Kannan, Identification of a hidden strain switch provides clues to an ancient structural mechanism in protein kinases, PNAS. 110 (2013) 924–929. https://doi.org/10.1073/pnas.1207104110.

[16] S. Mohanty, K. Oruganty, A. Kwon, D.P. Byrne, S. Ferries, Z. Ruan, L.E. Hanold, S. Katiyar, E.J. Kennedy, P.A. Eyers, N. Kannan, Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization, PLOS Genetics. 12 (2016) e1005885. https://doi.org/10.1371/journal.pgen.1005885.

[17] A. Mirza, M. Mustafa, E. Talevich, N. Kannan, Co-Conserved Features Associated with cis Regulation of ErbB Tyrosine Kinases, PLOS ONE. 5 (2010) e14310. https://doi.org/10.1371/journal.pone.0014310.

[18] A. Kwon, M. John, Z. Ruan, N. Kannan, Coupled regulation by the juxtamembrane and sterile α motif (SAM) linker is a hallmark of ephrin tyrosine kinase evolution, J Biol Chem. 293 (2018) 5102–5116. https://doi.org/10.1074/jbc.RA117.001296.

[19] N. Amatya, T.E. Wales, A. Kwon, W. Yeung, R.E. Joseph, D.B. Fulton, N. Kannan, J.R. Engen, A.H. Andreotti, Lipid-targeting pleckstrin homology domain turns its autoinhibitory face toward the TEC kinases, PNAS. 116 (2019) 21539–21544. https://doi.org/10.1073/pnas.1907566116.

[20] N.H. Shah, J.F. Amacher, L.M. Nocka, J. Kuriyan, The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases, Critical Reviews in Biochemistry and Molecular Biology. 53 (2018) 535–563. https://doi.org/10.1080/10409238.2018.1495173.

[21] A.F. Neuwald, Surveying the Manifold Divergence of an Entire Protein Class for Statistical Clues to Underlying Biochemical Mechanisms, Statistical Applications in Genetics and Molecular Biology. 10 (2011). https://doi.org/10.2202/1544-6115.1666.

[22] A.F. Neuwald, A Bayesian Sampler for Optimization of Protein Domain Hierarchies, Journal of Computational Biology. 21 (2014) 269–286. https://doi.org/10.1089/cmb.2013.0099.

[23] J.-W. Jung, W.-S. Shin, J. Song, S.-T. Lee, Cloning and characterization of the full-length mouse Ptk7 cDNA encoding a defective receptor protein tyrosine kinase, Gene. 328 (2004) 75–84. https://doi.org/10.1016/j.gene.2003.12.006.

[24] J.M. Murphy, Q. Zhang, S.N. Young, M.L. Reese, F.P. Bailey, P.A. Eyers, D. Ungureanu, H. Hammaren, O. Silvennoinen, L.N. Varghese, K. Chen, A. Tripaydonis, N. Jura, K. Fukuda, J. Qin, Z. Nimchuk, M.B. Mudgett, S. Elowe, C.L. Gee, L. Liu, R.J. Daly, G. Manning, J.J. Babon, I.S. Lucet, A robust methodology to subclassify pseudokinases based on their nucleotide-binding properties, Biochemical Journal. 457 (2013) 323–334. https://doi.org/10.1042/BJ20131174.

[25] H. Wang, D.L. Brautigan, A Novel Transmembrane Ser/Thr Kinase Complexes with Protein Phosphatase-1 and Inhibitor-2*, Journal of Biological Chemistry. 277 (2002) 49605–49612. https://doi.org/10.1074/jbc.M209335200.

[26] A. Ditsiou, C. Cilibrasi, N. Simigdala, A. Papakyriakou, L. Milton-Harris, V. Vella, J.E. Nettleship, J.H. Lo, S. Soni, G. Smbatyan, P. Ntavelou, T. Gagliano, M.C. Iachini, S. Khurshid, T. Simon, L. Zhou, S. Hassell-Hart, P. Carter, L.H. Pearl, R.L. Owen, R.J. Owens, S.M. Roe, N.E. Chayen, H.-J. Lenz, J. Spencer, C. Prodromou, A. Klinakis, J. Stebbing, G. Giamas, The structure-function relationship of oncogenic LMTK3, Science Advances. 6 (2020) eabc3099. https://doi.org/10.1126/sciadv.abc3099.

[27] G. Manning, S.L. Young, W.T. Miller, Y. Zhai, The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan, PNAS. 105 (2008) 9674–9679. https://doi.org/10.1073/pnas.0801314105.

[28] W. Xu, A. Doshi, M. Lei, M.J. Eck, S.C. Harrison, Crystal Structures of c-Src Reveal Features of Its Autoinhibitory Mechanism, Molecular Cell. 3 (1999) 629–638. https://doi.org/10.1016/S1097-2765(00)80356-1.

[29] T. Schindler, F. Sicheri, A. Pico, A. Gazit, A. Levitzki, J. Kuriyan, Crystal Structure of Hck in Complex with a Src Family—Selective Tyrosine Kinase Inhibitor, Molecular Cell. 3 (1999) 639–648. https://doi.org/10.1016/S1097-2765(00)80357-3.

[30] N.M. Levinson, O. Kuchment, K. Shen, M.A. Young, M. Koldobskiy, M. Karplus, P.A. Cole, J. Kuriyan, A Src-Like Inactive Conformation in the Abl Tyrosine Kinase Domain, PLOS Biology. 4 (2006) e144. https://doi.org/10.1371/journal.pbio.0040144.

[31] Q. Wang, E.M. Vogan, L.M. Nocka, C.E. Rosen, J.A. Zorn, S.C. Harrison, J. Kuriyan, Autoinhibition of Bruton's tyrosine kinase (Btk) and activation by soluble inositol hexakisphosphate, ELife. 4 (2015) e06074. https://doi.org/10.7554/eLife.06074.

[32] S.R. Hubbard, L. Wei, W.A. Hendrickson, Crystal structure of the tyrosine kinase domain of the human insulin receptor, Nature. 372 (1994) 746–754. https://doi.org/10.1038/372746a0.

[33] S.C. Artim, J.M. Mendrola, M.A. Lemmon, Assessing the range of kinase autoinhibition mechanisms in the insulin receptor family, Biochemical Journal. 448 (2012) 213–220. https://doi.org/10.1042/BJ20121365.

[34] P. Canning, L. Tan, K. Chu, S.W. Lee, N.S. Gray, A.N. Bullock, Structural Mechanisms Determining Inhibition of the Collagen Receptor DDR1 by Selective and Multi-Targeted Type II Kinase Inhibitors, Journal of Molecular Biology. 426 (2014) 2457–2470. https://doi.org/10.1016/j.jmb.2014.04.014.

[35] H. Chen, J. Ma, W. Li, A.V. Eliseenkova, C. Xu, T.A. Neubert, W.T. Miller, M. Mohammadi, A Molecular Brake in the Kinase Hinge Region Regulates the Activity of Receptor Tyrosine Kinases, Molecular Cell. 27 (2007) 717–730. https://doi.org/10.1016/j.molcel.2007.06.028.

[36] J. Griffith, J. Black, C. Faerman, L. Swenson, M. Wynn, F. Lu, J. Lippke, K. Saxena, The Structural Basis for Autoinhibition of FLT3 by the Juxtamembrane Domain, Molecular Cell. 13 (2004) 169–178. https://doi.org/10.1016/S1097-2765(03)00505-7.

[37] M. Nars, M. Vihinen, Coevolution of the Domains of Cytoplasmic Tyrosine Kinases, Molecular Biology and Evolution. 18 (2001) 312–321. https://doi.org/10.1093/oxfordjournals.molbev.a003807.

[38] F.G. Brunet, J.-N. Volff, M. Schartl, Whole Genome Duplications Shaped the Receptor Tyrosine Kinase Repertoire of Jawed Vertebrates, Genome Biology and Evolution. 8 (2016) 1600–1613. https://doi.org/10.1093/gbe/evw103.

[39] F.G. Brunet, T. Lorin, L. Bernard, Z. Haftek-Terreau, D. Galiana, M. Schartl, J.-N. Volff, Case Studies of Seven Gene Families with Unusual High Retention Rate Since the Vertebrate and Teleost Whole-Genome Duplications, in: P. Pontarotti (Ed.), Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts, Springer International Publishing, Cham, 2017: pp. 369–396. https://doi.org/10.1007/978-3-319-61569-1_19.

[40] L.E. Locascio, D.J. Donoghue, KIDs rule: regulatory phosphorylation of RTKs, Trends in Biochemical Sciences. 38 (2013) 75–84. https://doi.org/10.1016/j.tibs.2012.12.001.

[41] S. Manni, K. Kisko, T. Schleier, J. Missimer, K. Ballmer-Hofer, Functional and structural characterization of the kinase insert and the carboxy terminal domain in VEGF receptor 2 activation, The FASEB Journal. 28 (2014) 4914–4923. https://doi.org/10.1096/fj.14-256206.

[42] E.B. Pasquale, Eph receptors and ephrins in cancer: bidirectional signalling and beyond, Nature Reviews Cancer. 10 (2010) 165–180. https://doi.org/10.1038/nrc2806.

[43] X. Zhang, J. Gureasko, K. Shen, P.A. Cole, J. Kuriyan, An Allosteric Mechanism for Activation of the Kinase Domain of Epidermal Growth Factor Receptor, Cell. 125 (2006) 1137–1149. https://doi.org/10.1016/j.cell.2006.05.013.

[44] N. Jura, N.F. Endres, K. Engel, S. Deindl, R. Das, M.H. Lamers, D.E. Wemmer, X. Zhang, J. Kuriyan, Mechanism for Activation of the EGF Receptor Catalytic Domain by the Juxtamembrane Segment, Cell. 137 (2009) 1293–1307. https://doi.org/10.1016/j.cell.2009.04.025.

[45] M.A. Lemmon, J. Schlessinger, Cell Signaling by Receptor Tyrosine Kinases, Cell. 141 (2010) 1117–1134. https://doi.org/10.1016/j.cell.2010.06.011.

[46] M.A. Lemmon, J. Schlessinger, K.M. Ferguson, The EGFR Family: Not So Prototypical Receptor Tyrosine Kinases, Cold Spring Harb Perspect Biol. 6 (2014) a020768. https://doi.org/10.1101/cshperspect.a020768.

[47] S.K. Hanks, T. Hunter, The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification1, The FASEB Journal. 9 (1995) 576–596. https://doi.org/10.1096/fasebj.9.8.7768349.

[48] A. Kwon, S. Scott, R. Taujale, W. Yeung, K.J. Kochut, P.A. Eyers, N. Kannan, Tracing the origin and evolution of pseudokinases across the tree of life, Sci. Signal. 12 (2019). https://doi.org/10.1126/scisignal.aav3810.

[49] B. Taskinen, E. Ferrada, D.M. Fowler, Early emergence of negative regulation of the tyrosine kinase Src by the C-terminal Src kinase, Journal of Biological Chemistry. 292 (2017) 18518–18529. https://doi.org/10.1074/jbc.M117.811174.

[50] N. Kannan, A.F. Neuwald, Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha, Protein Sci. 13 (2004) 2059–2077. https://doi.org/10.1110/ps.04637904.

[51] H. Suga, G. Torruella, G. Burger, M.W. Brown, I. Ruiz-Trillo, Earliest Holozoan Expansion of Phosphotyrosine Signaling, Molecular Biology and Evolution. 31 (2014) 517–528. https://doi.org/10.1093/molbev/mst241.

[52] T. Hunter, G. Manning, The Eukaryotic Protein Kinase Superfamily and the Emergence of Receptor Tyrosine Kinases, in: D.L. Wheeler, Y. Yarden (Eds.), Receptor Tyrosine Kinases: Structure, Functions and Role in Human Disease, Springer, New York, NY, 2015: pp. 1–15. https://doi.org/10.1007/978-1-4939-2053-2_1.

[53] D.J. Richter, N. King, The Genomic and Cellular Foundations of Animal Origins, Annu. Rev. Genet. 47 (2013) 509–537. https://doi.org/10.1146/annurev-genet-111212-133456.

[54] J. Grassot, M. Gouy, G. Perrière, G. Mouchiroud, Origin and Molecular Evolution of Receptor Tyrosine Kinases with Immunoglobulin-Like Domains, Molecular Biology and Evolution. 23 (2006) 1232–1241. https://doi.org/10.1093/molbev/msk007.

[55] S. D'Aniello, M. Irimia, I. Maeso, J. Pascual-Anaya, S. Jiménez-Delgado, S. Bertrand, J. Garcia-Fernàndez, Gene Expansion and Retention Leads to a Diverse Tyrosine Kinase Superfamily in Amphioxus, Molecular Biology and Evolution. 25 (2008) 1841–1854. https://doi.org/10.1093/molbev/msn132.

[56] T. Hunter, Tyrosine phosphorylation: thirty years and counting, Current Opinion in Cell Biology. 21 (2009) 140–146. https://doi.org/10.1016/j.ceb.2009.01.028.

[57] H. Chen, F. Lin, K. Xing, X. He, The reverse evolution from multicellularity to unicellularity during carcinogenesis, Nature Communications. 6 (2015) 6367. https://doi.org/10.1038/ncomms7367.

[58] T. Domazet-Lošo, D. Tautz, Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa, BMC Biology. 8 (2010) 66. https://doi.org/10.1186/1741-7007-8-66.

[59] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Research. 35 (2007) D61–D65. https://doi.org/10.1093/nar/gkl842.

[60] A.F. Neuwald, Rapid detection, classification and accurate alignment of up to a million or more related protein sequences, Bioinformatics. 25 (2009) 1869–1875. https://doi.org/10.1093/bioinformatics/btp342.

[61] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature. 585 (2020) 357–362. https://doi.org/10.1038/s41586-020-2649-2.

[62] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: A Sequence Logo Generator, Genome Res. 14 (2004) 1188–1190. https://doi.org/10.1101/gr.849004.

[63] S.R. Fairclough, Z. Chen, E. Kramer, Q. Zeng, S. Young, H.M. Robertson, E. Begovic, D.J. Richter, C. Russ, M.J. Westbrook, G. Manning, B.F. Lang, B. Haas, C. Nusbaum, N. King, Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta, Genome Biology. 14 (2013) R15. https://doi.org/10.1186/gb-2013-14-2-r15.

[64] C. Junqueira Alves, K. Yotoko, H. Zou, R.H. Friedel, Origin and evolution of plexins, semaphorins, and Met receptor tyrosine kinases, Sci Rep. 9 (2019) 1970. https://doi.org/10.1038/s41598-019-38512-y.

[65] O. Keller, F. Odronitz, M. Stanke, M. Kollmar, S. Waack, Scipio: Using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species, BMC Bioinformatics. 9 (2008) 278. https://doi.org/10.1186/1471-2105-9-278.

[66] A. Marchler-Bauer, S. Lu, J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, J.D. Jackson, Z. Ke, C.J. Lanczycki, F. Lu, G.H. Marchler, M. Mullokandov, M.V. Omelchenko, C.L. Robertson, J.S. Song, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S.H. Bryant, CDD: a Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Research. 39 (2011) D225–D229. https://doi.org/10.1093/nar/gkq1189.

[67] A. Marchler-Bauer, Y. Bo, L. Han, J. He, C.J. Lanczycki, S. Lu, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng, L.Y. Geer, S.H. Bryant, CDD/SPARCLE: functional classification of proteins via subfamily domain architectures, Nucleic Acids Research. 45 (2017) D200–D203. https://doi.org/10.1093/nar/gkw1129.

[68] A. Krogh, B. Larsson, G. von Heijne, E.L.L. Sonnhammer, Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, Journal of Molecular Biology. 305 (2001) 567–580. https://doi.org/10.1006/jmbi.2000.4315.

[69] L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies, Molecular Biology and Evolution. 32 (2015) 268–274. https://doi.org/10.1093/molbev/msu300.

[70] S. Kalyaanamoorthy, B.Q. Minh, T.K.F. Wong, A. von Haeseler, L.S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates, Nature Methods. 14 (2017) 587–589. https://doi.org/10.1038/nmeth.4285.

[71] D.T. Hoang, O. Chernomor, A. von Haeseler, B.Q. Minh, L.S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation, Molecular Biology and Evolution. 35 (2018) 518–522. https://doi.org/10.1093/molbev/msx281.

[72] J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data, Molecular Biology and Evolution. 33 (2016) 1635–1638. https://doi.org/10.1093/molbev/msw046.

[73] H. Suga, G. Sasaki, K. Kuma, H. Nishiyori, N. Hirose, Z.-H. Su, N. Iwabe, T. Miyata, Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes, FEBS Letters. 582 (2008) 815–818. https://doi.org/10.1016/j.febslet.2008.02.002.

[74] V. Modi, R.L. Dunbrack, A Structurally-Validated Multiple Sequence Alignment of 497 Human Protein Kinase Domains, Scientific Reports. 9 (2019) 19790. https://doi.org/10.1038/s41598-019-56499-4.

[75] S.-H. Shiu, W.-H. Li, Origins, Lineage-Specific Expansions, and Multiple Losses of Tyrosine Kinases in Eukaryotes, Molecular Biology and Evolution. 21 (2004) 828–840. https://doi.org/10.1093/molbev/msh077.

[76] S.K. Mitra, D.A. Hanson, D.D. Schlaepfer, Focal adhesion kinase: in command and control of cell motility, Nature Reviews Molecular Cell Biology. 6 (2005) 56–68. https://doi.org/10.1038/nrm1549.

[77] J.M. Mitchell, S.A. Nichols, Diverse cell junctions with unique molecular composition in tissues of a sponge (Porifera), EvoDevo. 10 (2019) 26. https://doi.org/10.1186/s13227-019-0139-0.

[78] A. Poliakov, M. Cotrina, D.G. Wilkinson, Diverse Roles of Eph Receptors and Ephrins in the Regulation of Cell Migration and Tissue Assembly, Developmental Cell. 7 (2004) 465–480. https://doi.org/10.1016/j.devcel.2004.09.006.

[79] P.J. Green, F.S. Walsh, P. Doherty, Promiscuity of fibroblast growth factor receptors, BioEssays. 18 (1996) 639–646. https://doi.org/10.1002/bies.950180807.

[80] D.Q. Matus, G.H. Thomsen, M.Q. Martindale, FGF signaling in gastrulation and neural development in Nematostella vectensis, an anthozoan cnidarian, Dev Genes Evol. 217 (2007) 137–148. https://doi.org/10.1007/s00427-006-0122-3.

[81] S. Bertrand, T. Iwema, H. Escriva, FGF Signaling Emerged Concomitantly with the Origin of Eumetazoans, Molecular Biology and Evolution. 31 (2014) 310–318. https://doi.org/10.1093/molbev/mst222.

[82] S. Barberán, J.M. Martín-Durán, F. Cebrià, Evolution of the EGFR pathway in Metazoa and its diversification in the planarian Schmidtea mediterranea, Sci Rep. 6 (2016) 28071. https://doi.org/10.1038/srep28071.

[83] S.-L. Chiu, H.T. Cline, Insulin receptor signaling in the development of neuronal structure and function, Neural Development. 5 (2010) 7. https://doi.org/10.1186/1749-8104-5-7.

[84] R. Monahan-Earley, A.M. Dvorak, W.C. Aird, Evolutionary origins of the blood vascular system and endothelium, Journal of Thrombosis and Haemostasis. 11 (2013) 46–66. https://doi.org/10.1111/jth.12253.

[85] R. Berenstein, Class III Receptor Tyrosine Kinases in Acute Leukemia – Biological Functions and Modern Laboratory Analysis, Biomarker Insights. 10s3 (2015) BMI.S22433. https://doi.org/10.4137/BMI.S22433.

[86] A. Mócsai, J. Ruland, V.L.J. Tybulewicz, The SYK tyrosine kinase: a crucial player in diverse biological functions, Nature Reviews Immunology. 10 (2010) 387–402. https://doi.org/10.1038/nri2765.

[87] C. Liongue, R. Sertori, A.C. Ward, Evolution of Cytokine Receptor Signaling, The Journal of Immunology. 197 (2016) 11–18. https://doi.org/10.4049/jimmunol.1600372.

[88] M. Mohammadi, J. Schlessinger, S.R. Hubbard, Structure of the FGF Receptor Tyrosine Kinase Domain Reveals a Novel Autoinhibitory Mechanism, Cell. 86 (1996) 577–587. https://doi.org/10.1016/S0092-8674(00)80131-2.

[89] J.G. Tate, S. Bamford, H.C. Jubb, Z. Sondka, D.M. Beare, N. Bindal, H. Boutselakis, C.G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S.C. Jupe, C.Y. Kok, K. Noble, L. Ponting, C.C. Ramshaw, C.E. Rye, H.E. Speedy, R. Stefancsik, S.L. Thompson, S. Wang, S. Ward, P.J. Campbell, S.A. Forbes, COSMIC: the Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Research. 47 (2019) D941–D947. https://doi.org/10.1093/nar/gky1015.

Chapter 4

ALIGNMENT-FREE EVOLUTIONARY ANALYSIS AND SEQUENCE CLASSIFICATION USING TRANSFORMER PROTEIN LANGUAGE MODELS

W. Yeung, Z. Zhou, L. Mathew, N. Gravel, R. Taujale, A. Venkat, W. Lanzilotta, S. Li, N. Kannan, Alignment-free evolutionary analysis and classification using Transformer protein language models.

Submitted to PNAS, 2022.

Abstract

Protein sequence analysis is a foundational aspect of biology which is traditionally facilitated by the alignment-based comparison of primary sequences. While alignment-based approaches work well on closely related sequences, they pose major challenges towards the classification of highly divergent or fast-evolving protein families. Here, we developed methods for alignment-free evolutionary analyses using embedding vectors generated from pre-trained protein language models which capture underlying protein structural-functional properties from unsupervised training on millions of biologically-observed sequences. Comparisons between many pre-trained language models reveal that embedding vectors generated from ESM-1b can be used to infer embedding-based phylogenies with branch support in a completely alignment-independent manner. Results remain highly consistent with equivalent alignment-based trees while also inferring new evolutionary relationships. The placement of proteins on embedding-based trees can be explained using projection vectors which highlight fast/slow-evolving sequence regions. We showcase the application of embedding-based sequence analyses in benchmark studies across three diverse protein superfamilies. Within the protein kinase superfamily, the embedding-based tree reveals Casein Kinase 1 (CK1) as the most ancestral protein kinase group linking canonical protein kinases with evolutionarily distant small molecule and lipid kinases. Embedding-based hierarchical clustering also yields a biologically-meaningful organization of phosphatase enzymes spanning ten different structural folds. Finally, we infer the first phylogeny of the radical S-Adenosyl-L-Methionine (SAM) superfamily which have been challenging to align due to the high degree of structural variations between families. We propose embedding-based trees as an orthogonal approach for the evolutionary classification of divergent protein families.

4.1 Introduction

Alignment-based biological sequence comparison is a foundational aspect of bioinformatics. High-quality sequence alignments are critical for accurate protein classification [1], function prediction [2], structure prediction [3], and evolutionary inference [4]. While alignments excel at comparing closely-related sequences, comparing divergent sequences, especially beyond the “twilight zone” (~25% sequence identity) [5] requires sophisticated methods. Profile-based methods such as PSI-BLAST [6], HMMER [7], and MMseqs [8] are capable of comparing sequences within the twilight zone; however, performance depends on alignment parameters such as substitution matrices and gap penalties, derived from prior assumptions about protein evolution. Alignment-free strategies based on word-frequency [9] or information theory [10] have been proposed; however, these methods suffer from high false positive rates and cannot capture co-evolutionary information in primary sequences [11].

Recent advances in representation learning offer a powerful alternative for alignment-free comparison of protein sequences. Using the Transformer neural network architecture [12], protein language models (LM) such as ESM-1b [13] and ProtBERT [14] capture the underlying grammar of biological sequences by training on large, universal proteome databases such as UniProt [15]. These models are trained by masked language modeling in which a random subset of residues in each sequence is replaced with blanks and the model is trained to fill in these blanks using contextual information. During this process, the model translates protein sequences into embedding vectors, which serve as a numerical matrix representation of the original sequence. Sequence embeddings are typically used as input features for machine learning to facilitate supervised predictions of various structure-functional properties [16–18]. Although useful, these methods utilize pre-trained LMs as a black-box feature extractor, resulting in

limited interpretability and biological insight. Furthermore, these methods require labeled data, demanding additional labor for curation as well as introducing a potential source of error and bias. Placing an emphasis on unsupervised methods, we developed a set of analytical methods which utilize sequence embeddings as a proxy for protein sequences.

We present a generalized, unsupervised protocol for hierarchical clustering on protein sequence embedding vectors. Benchmark studies across diverse protein superfamilies reveal that sequence embeddings can quantify long-distance evolutionary relationships, beyond the twilight zone of sequence similarity. Visualization of sequence projection vectors reveals cluster-specific sequence motifs, which enable explainability and provide additional support for embedding-based classification. Evaluation of multiple language models reveals that ESM-1b best captures the complexities of protein sequence space. We conclude that embedding-based, alignment-free evolutionary analyses offer a unique set of strengths — well-suited as an orthogonal, complementary approach to traditional alignment-based techniques for protein sequence analysis.

4.2 Results

4.2.1 Sequence embeddings enable comparisons across long evolutionary distances

We evaluate the ability of protein LMs to model distances between highly divergent protein sequences using the encoder of ESM-1b [13]. LM encoders contain a variable number of Attention blocks [12] where the majority of interpretable information accumulates at the last Attention block of the encoder **(Figure 4.1A)**. While previous work has shown that pairwise structural contacts can be inferred as a mathematical function of the attention matrix [19], we gained additional explainability through sequence projections derived from the embedding

vector, calculated downstream to the attention matrix. Sequence projections vectors assign normalized weights to each residue in a given protein sequence which infers important catalytic motifs and fast/slow evolving sites. We later demonstrate this in three diverse protein superfamilies.

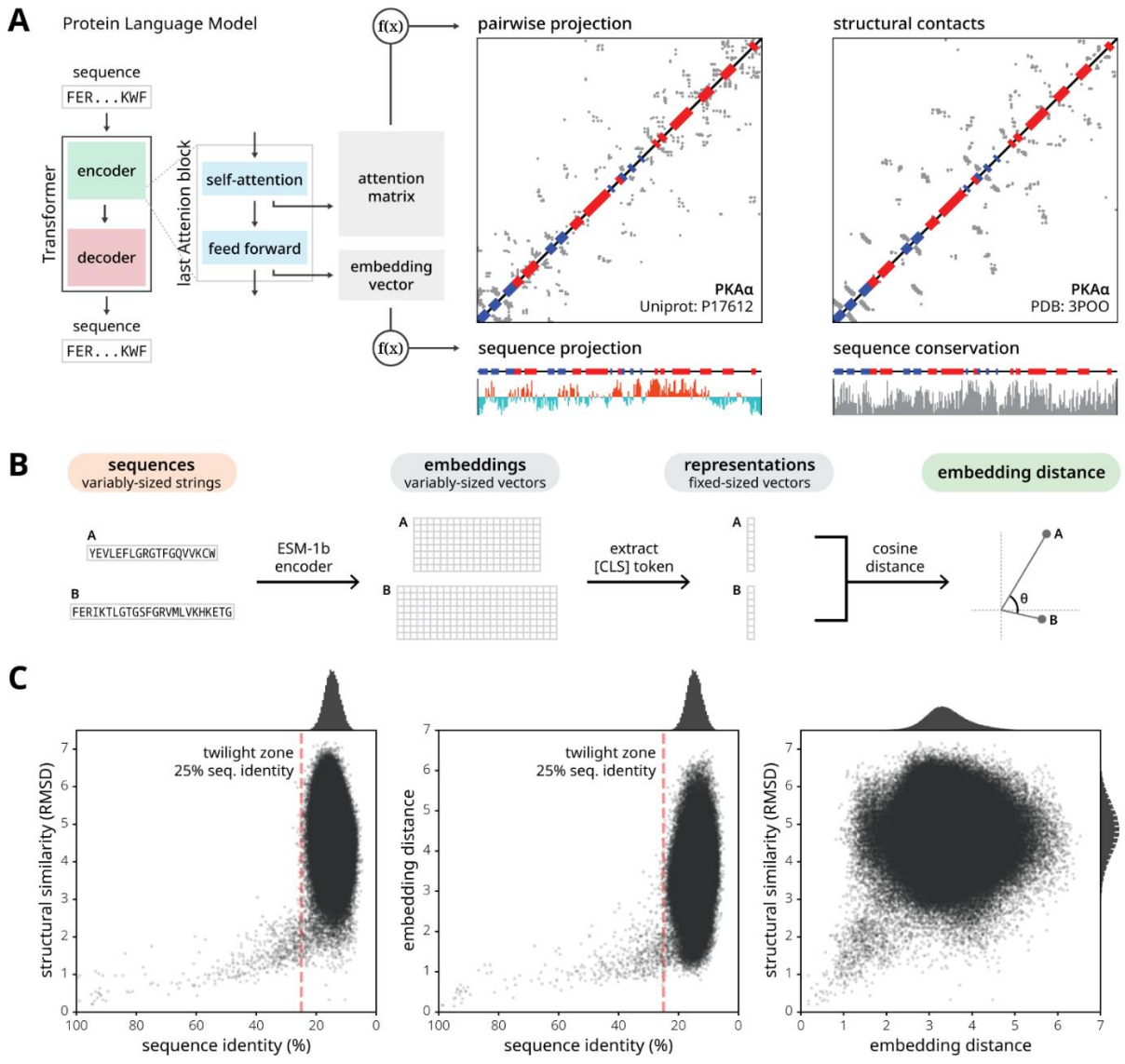


Figure 4.1: Embedding vectors encode a nuanced description of protein sequence
(A) On the far left, we show a graphical representation of a Transformer-based protein Language Model (LM) consisting of an encoder and a decoder module. A zoomed inset depicts major components within the last attention block in the encoder stack. Pairwise contacts can be inferred as a function of the attention matrix [19]. Sequence projections, calculated as a function of the

embedding vector, correspond to fast/slow evolving regions. The example was generated from the kinase domain sequence of PKAα (UniProt: P17612) using the ESM-1b model [13]. On the far right, pairwise structural contacts are shown for the crystal structure of PKAα (PDB: 3POO) defined by C-alpha contacts <7.5 Å. Sequence conservation was calculated by the Jensen-Shannon divergence. (C) A graphical overview shows how to calculate embedding distance which provides an embedding-based protocol for pairwise comparison. (D) Three scatter plots depict the relationship between sequence and structure (left), sequence and embedding (middle), and embedding and structure (right). Histograms show the distribution of sequence identities, embedding distances, and structural similarities along their respective axes. The twilight zone of sequence identity is marked by a dotted red line at 25% identity [5]. Each point denotes a pairwise comparison between two protein domains. This data was collected by randomly selecting 1,000 proteins from the SCOP (Structural Classification of Proteins) database [20], provided they were 80-250 residues long with a resolution of 2.3 Å or better.

Embedding vectors facilitate meaningful pairwise comparisons because they encode a nuanced description of protein sequence information. The distance between two embedding vectors can be measured by calculating cosine similarity using the [CLS] special token which is appended before each sequence to capture the sequence-level information during standard preprocessing **(Figure 4.1B)**. We measured embedding distances between 1000 randomly selected protein domains against standard measures of sequence similarity (percent identity) and structural similarity (RMSD).

As expected, scatterplots show a close relationship between protein sequence and structural similarity that abruptly fades in the twilight zone **(Figure 4.1C, left)** (sequences below 25% identity)[5,21,22]. Notably, embedding distance is also correlated with sequence identity, displaying a similar boundary at ~25% **(Figure 4.1C, middle)**. In contrast, embedding distance and structural similarity (RMSD) display a positive correlation **(Figure 4.1C, right)**, but instead of a twilight zone, larger variance in embedding distance is observed with increased structural divergence (larger RMSD). This is because sequence embeddings capture a wide range of

protein properties beyond 3D structure. Together, these comparisons suggest that protein sequence embeddings can be used as a proxy for sequence and structural similarity metrices and are suitable for comparing sequences in the twilight zone, where traditional alignment-based approaches have proven difficult.

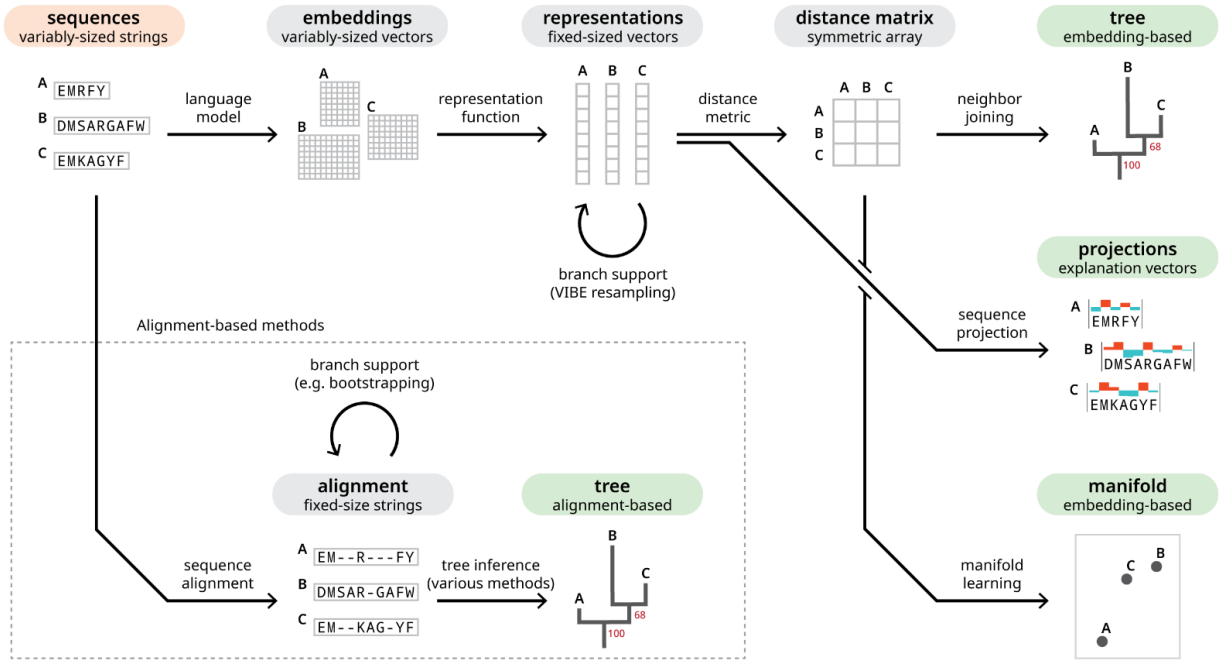


Figure 4.2: Workflows for alignment and embedding-based sequence analysis
A graphical overview of analysis workflows starting from unaligned protein sequences (label highlighted in red) can lead to four possible endpoints (each label highlighted in green). The top row describes the protocol for creating an embedding tree. Under the representations, the circular arrow denotes our variational autoencoder (VAE)-based strategy for resampling the representation vectors. Resampled representations are used to build replicate trees to calculate branch support, represented by the red number underneath each fork on the tree. Representations, generated from embedding vectors, can also be used to create sequence projections (middle-right) or clustered using manifold learning algorithms such as UMAP (bottom-right). The branching route in the bottom row depicts a more traditional protocol for creating trees using multiple sequence alignments. There are many diverse algorithms for inferring trees using sequence alignments [4]. There are also various methods for resampling data to build replicate trees (such as bootstrapping) which is required for branch support calculations [23].

4.2.2 Unsupervised hierarchical clustering of the protein embedding manifold

Harnessing the unique advantages of protein sequence embeddings, we developed orthogonal methods to facilitate alignment-free evolutionary analyses. We define a hierarchical clustering protocol for constructing embedding trees **(Figure 4.2, top row)** which provide meaningful organizations of protein sequence datasets and in some instances (see below) can also reflect evolutionary relationships. This protocol has three hyperparameters: the pre-trained LM, representation function, and distance metric.

We systematically assessed all hyperparameter combinations across diverse case studies using three enzyme superfamilies which will be individually discussed over the next three sections. We used a variety of quantitative measures such as Sackin's index [24], treeness [25], and silhouette coefficient [26] to evaluate embedding trees which are also a general strategy for visualizing high-dimensional datasets — representing pairwise relationships using cophenetic distance. To measure how well the tree preserves all pairwise distances observed in the original data, we quantify the Pearson's correlation of the tree's distance matrix versus the representation's distance matrix. Using the same method, we also compared against manifold learning algorithms such as UMAP (Uniform Manifold Approximation and Projection) [27,28].

Across all sequence datasets, the ESM-1b model consistently produced trees that agree with previously established protein classifications schemes based on silhouette coefficient while also proposing new relationships. Although some LMs such as ProtBERT can be fine-tuned to gain better performance for specific tasks, fine-tuned LMs did not yield significant improvements in embedding trees. Given the overall performance of ESM-1b, all analyses throughout this study utilized this LM. Meaningfully compressing embedding vectors [29] and

defining a unified distance metric [30] are both non-trivial problems. Consequently, the optimal representation function and distance metric varied across different protein datasets.

Upon identifying an optimal tree, we quantify clustering confidence using a variational autoencoder (VAE)-based strategy. The confidence of each split is measured using VIBE (VAE-Implemented Branch support Estimation) **(Figure 4.2, top-middle)** — conceptually similar to bootstrap support, used in alignment-based phylogenies. As a generative model, the VAE learns the latent distribution of a given set of representations [31], then resamples the distribution to generate replicate trees. We assign a value to each branch of the original tree, indicating the percentage of replicate trees which also exhibited the same corresponding bipartition. This is a particularly stringent metric which does not consider similar bipartitions if an exact match is not present.

4.2.3 Embedding trees infer the earliest diverging protein kinase group.

We applied our methods towards the protein kinase superfamily — an important gene family which plays diverse roles in cellular signaling and disease. Most protein kinases are classified into nine major groups based on sequence similarity [33]. Outside the protein kinase superfamily, lipid and small molecule kinases are distant relatives which conserve a similar bilobal structure [34,35]. Although structure-functional similarities strongly imply evolutionary relationships between all kinase-fold enzymes, further characterization has eluded traditional phylogenetic methods.

We built an embedding tree of ~550 human kinase-fold enzymes using unaligned protein sequences, trimmed to the conserved catalytic domain. The optimal tree organizes sequences into nine major groups **(Figure 4.3A)** where the inferred between-group relationships are largely

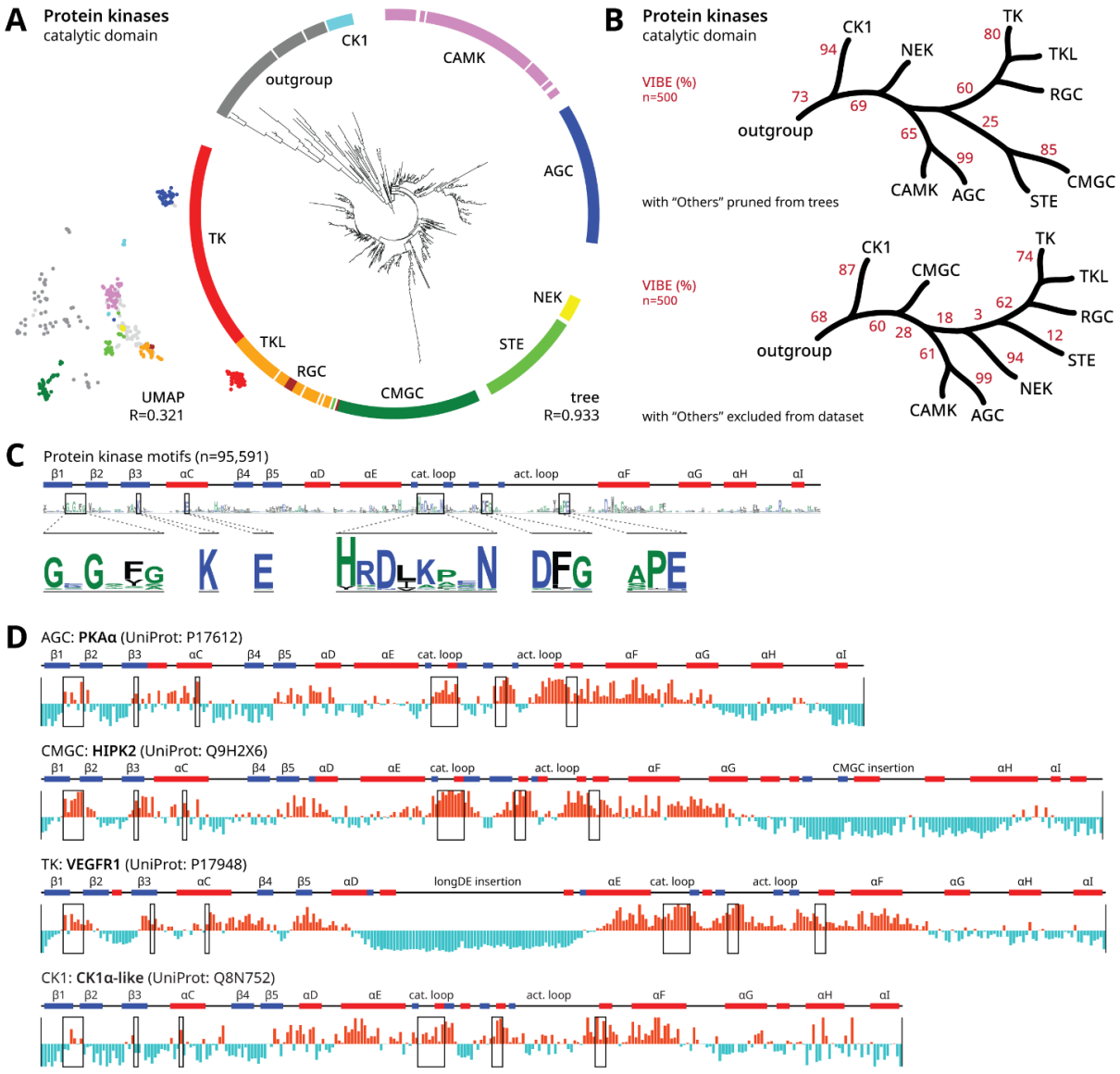


Figure 4.3: Embedding-based analysis of the human protein kinases
(A) An embedding tree of the human protein kinase domains in a circular layout with major groups labeled. This tree was generated using the sum_spec representation function and TS-SS (triangle similarity sector similarity) distance [32]. To the left of the tree, we plot a UMAP projection using the same dataset. At the bottom of each graph, we provide the correlation coefficient which quantifies how well the all-vs-all pairwise distances denoted by each visualization reflects the pairwise distances from the original dataset. (B) Stylized trees showing the major kinase groups with VIBEs indicated by the red percentage values. The top topology was inferred by pruning unclassified kinases ("Others" group). The bottom topology was inferred by excluding the unclassified from the sequence set. (C) Six major protein kinase motifs are shown within zoomed insets of a sequence logo, generated from an alignment of 95,591

kinase sequences. Above the logo plot, we show secondary structure elements across the kinase domain with α-helices (red), β-sheets (blue), and loops (black). (D) Sequence projections for three diverse protein kinase sequences: PKAα, HIPK2, VEGFR1, and CK1α-like. Positive peaks are shown in bright red and negative peaks are shown in icy blue. Sequence regions corresponding to the six major protein kinase motifs are designated by boxes. We note that CK1 kinases lack the APE motif and instead conserve a CK1-specific SIN motif at the equivalent position. Based on the optimal tree parameters, sequence projections were calculated using the sum_spec representation function.

consistent with the widely-accepted alignment-based phylogeny [33]. In comparison, untrimmed sequences yield a trivial topology, indicating that meaningful evolutionary analyses require a common frame of reference. To further evaluate confidence, we generated 500 replicates for the kinase domain tree. Sequences from the "Others" category (not belonging to the major protein kinase groups) showed unstable placement across replicates. These rogue taxa are known to decrease branch support [36], thus we used two common strategies to resolve this issue. The first strategy was to prune rogues from all trees prior to calculating VIBEs, while the second was to rebuild the tree excluding rogue sequences from the dataset **(Figure 4.3B)**. For both trees, at least 60% of replicates place RGC, TKL, and TK into a monophyletic clade, also placing CAMK and AGC as sister clades — consistent with the existing phylogeny [33]. Extending beyond the existing model, we included an evolutionary outgroup of lipid and small molecule kinases. The placement of CK1 kinases in both topologies infer that CK1 is the earliest diverging protein kinase group, which is further supported by CK1-specific divergence in the substrate binding lobe [37] and its apparent substrate promiscuity and constitutive activity [38].

Relationships between sequence embeddings can also be visualized by manifold learning algorithms such as UMAP [27]. We compare against our tree-based method by creating a UMAP projection from the same dataset. The tree-based layout is superior at preserving pairwise

distance information, facilitating a more accurate depiction of the underlying manifold. In the kinase domain dataset, all-vs-all pairwise distances from the UMAP projection are weakly correlated to the original data, quantified by a Pearson's correlation coefficient of 0.366, compared to 0.926 for the tree **(Figure 4.3A)**. While pairwise distances in UMAP scatterplots are represented by Euclidean distance, pairwise distances in circular trees are represented by cophenetic distance, the sum of branch lengths along the shortest path between two points. Branch length is solely represented by distance across the radial axis, while the circular axis and number of edges do not matter [39].

Sequence projections provide further explainability for embedding-based analyses. The sequence projection quantifies how strongly a given representation vector weights each residue of a protein sequence. Weights are correlated to fast/slow evolving sites. Most kinases share a common set of sequence motifs such as the nucleotide-binding G-loop motif and catalytic motifs **(Figure 4.3C)** [40]. A projection of archetypical kinase PKA-α reveals positive peaks for kinase-conserved motifs **(Figure 4.3D)**. We observe similar peaks for HIPK2 which has a CMGC-specific insertion region towards the C-terminal of the kinase domain [41] and VEGFR1 which has the longDE insertion towards the center of the kinase domain [42]. These fast-evolving insertion regions correspond to negative peaks. A sequence projection of CK1α-like kinase also highlights protein kinase motifs, albeit with its own unique variations. While determining fast/slow evolving sites typically require a sequence alignment, protein LMs delineate this information without an alignment, functioning as unsupervised learners for fast/slow evolving sites.

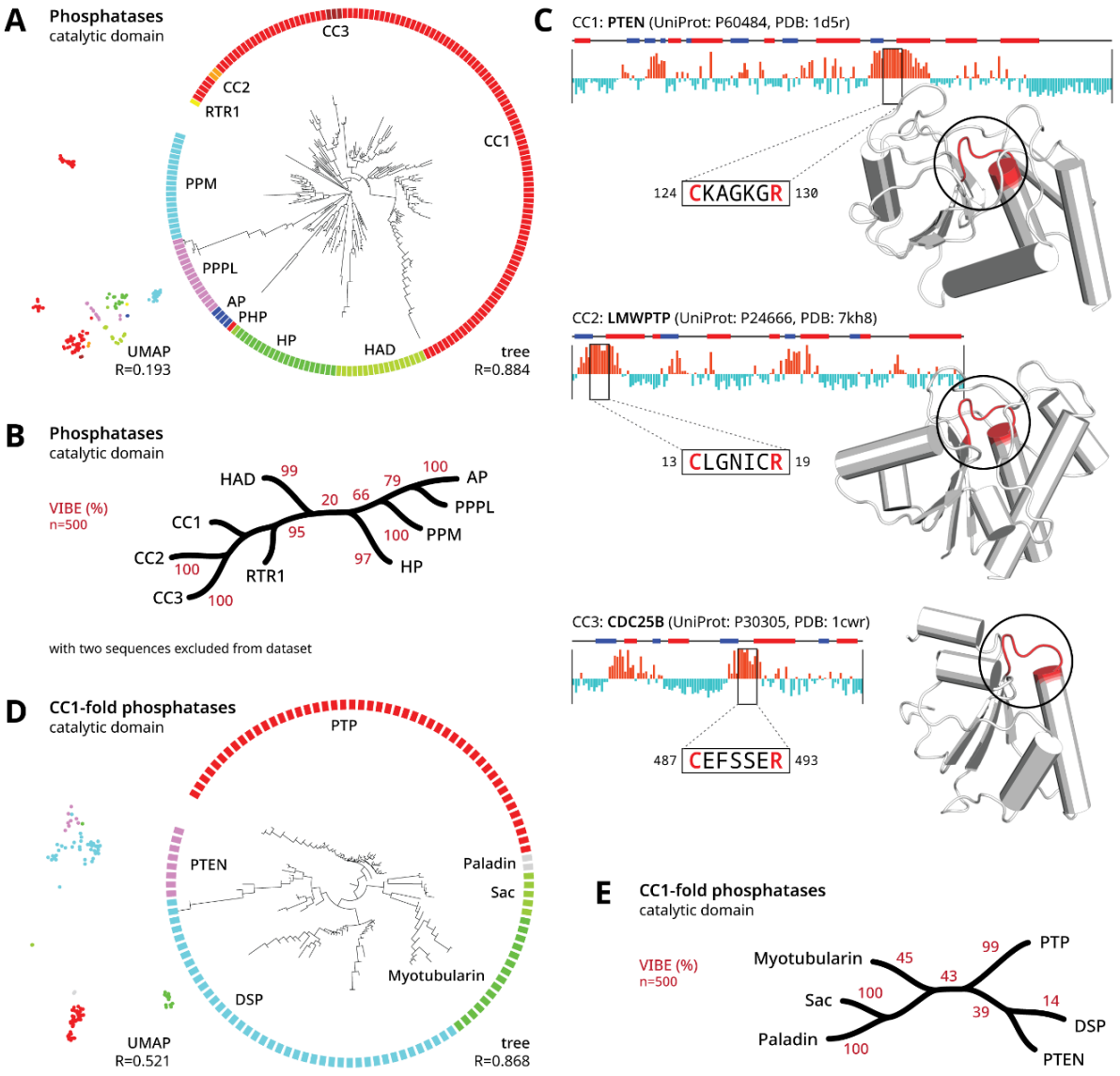


Figure 4.4: Embedding-based analysis of the human phosphatases
(A) An embedding tree of the human phosphatases in circular layout, generated from all human phosphatases enzymes spanning ten structural folds. This tree was generated using the avg_seq representation function and cosine distance. To the left of the tree, we plot a UMAP projection using the same dataset. (B) A stylized tree showing the phosphatase folds with VIBEs indicated by the red percentage values. This topology was inferred by excluding two rogue taxa, O60729 and Q9NRX4, from the sequence set. (C) Sequence projections for a representative CC1, CC2, and CC3-fold phosphatase. The conserved CxxxxxR motif corresponds to a positive peak in all three enzymes, despite adopting different protein folds. Crystal structures of each enzyme are shown with the CxxxxxR motif circled and highlighted red. Based on the optimal tree parameters, sequence projections were calculated using the avg_seq representation function. (D) An embedding tree and UMAP projection of the human CC1-fold phosphatases. This tree was

generated using the [CLF] representation function and TS-SS distance. (E) A stylized tree showing the CC1 phosphatase families with VIBEs indicated by the red percentage values.

4.2.4 Embedding trees capture similarities between protein folds in protein phosphatases

To further demonstrate the applicability of embedding trees, we generated trees for phosphatase enzymes, which, unlike kinases, adopt distinct structural folds [43]. Out of ~200 human phosphatases, roughly half adopt the CC1 fold, while only one adopts the RTR1 or PHP fold. The salient heterogeneity of structural folds suggests that phosphatases emerged independently multiple times throughout evolution [44,45].

We constructed an embedding tree spanning all ten structural folds using the catalytic domain sequences **(Figure 4.4A)**. VIBEs were calculated using a filtered dataset which excludes two rogue taxa **(Figure 4.4B)**. The grouping of the three cysteine-based phosphatase folds (CC1, CC2, and CC3) was supported by 95% of replicates. Within all three structural folds, sequence projections revealed a positive peak at the shared CxxxxxR catalytic motif **(Figure 4.4C)**. Catalytic similarities between CC1, CC2, and CC3 phosphatases likely arose via convergent evolution — CC2 is more structurally related to bacterial arsenate reductases, while CC3 emerged from bacterial rhodanese-like enzymes [46]. At the opposite end of the tree, PPPL, PPM, and AP were placed into a distinct cluster supported by 66% of replicates. PPPL and PPM phosphatases are phosphoserine/threonine-specific [47], while AP phosphatases act on phosphotyrosine [48] with possible phosphoserine/threonine activity based on substrate binding specificities [49]. While these enzymes share similar substrates, the embedding-based similarities between these three folds are not immediately obvious.

We also constructed another embedding tree from a reduced dataset only containing CC1-fold phosphatases which adopt a conserved structural fold, implying a common evolutionary origin. Consistent with an alignment-based phylogeny[43], the embedding tree identified five major clades across the six families. Notably, DSP is a paraphyletic group and shares a clade with PTEN **(Figure 4.4D)**. VIBEs of the five major clades ranged from 39-100% **(Figure 4.4E)**. While the CC1-fold tree showed evolutionary relationships, embedding trees for highly divergent sequence sets should be interpreted with caution as similarities can arise from alternative sources such as convergent evolution. Even if evolutionary inferences cannot be made, results can still be interpreted as hierarchical clustering.

4.2.5 An initial characterization of the radical SAM enzyme superfamily

We apply embedding-based methods towards an initial evolutionary characterization of the radical S-Adenosyl-L-Methionine (SAM) enzyme superfamily. SAM enzymes are present in all domains of life, catalyzing radical chemistry towards a wide variety of essential biological functions [50]. The catalytic core domain of radical SAM enzymes adopts a TIM barrel (α/β barrel) fold with varying numbers of α/β pairs, and a conserved iron-sulfur cluster binding motif, CxxxCxΦC, where Φ denotes an aromatic residue [51]. Family-specific insertions and deletions add additional structural variance, making a superfamily-scale alignment difficult. We curated a dataset of diverse radical SAM enzymes using available protein structures and the AlphaFold2 database [52]. To establish a common frame of reference, we trimmed each sequence to the core catalytic domain, removing any domain extensions or accessory domains.

Despite only utilizing the core domain, an embedding tree of the radical SAM superfamily organized enzymes into structure-functionally similar groups **(Figure 4.5A)** with good VIBEs **(Figure 4.5B)**. For instance, families which specialize in methyl or sulfur transfer

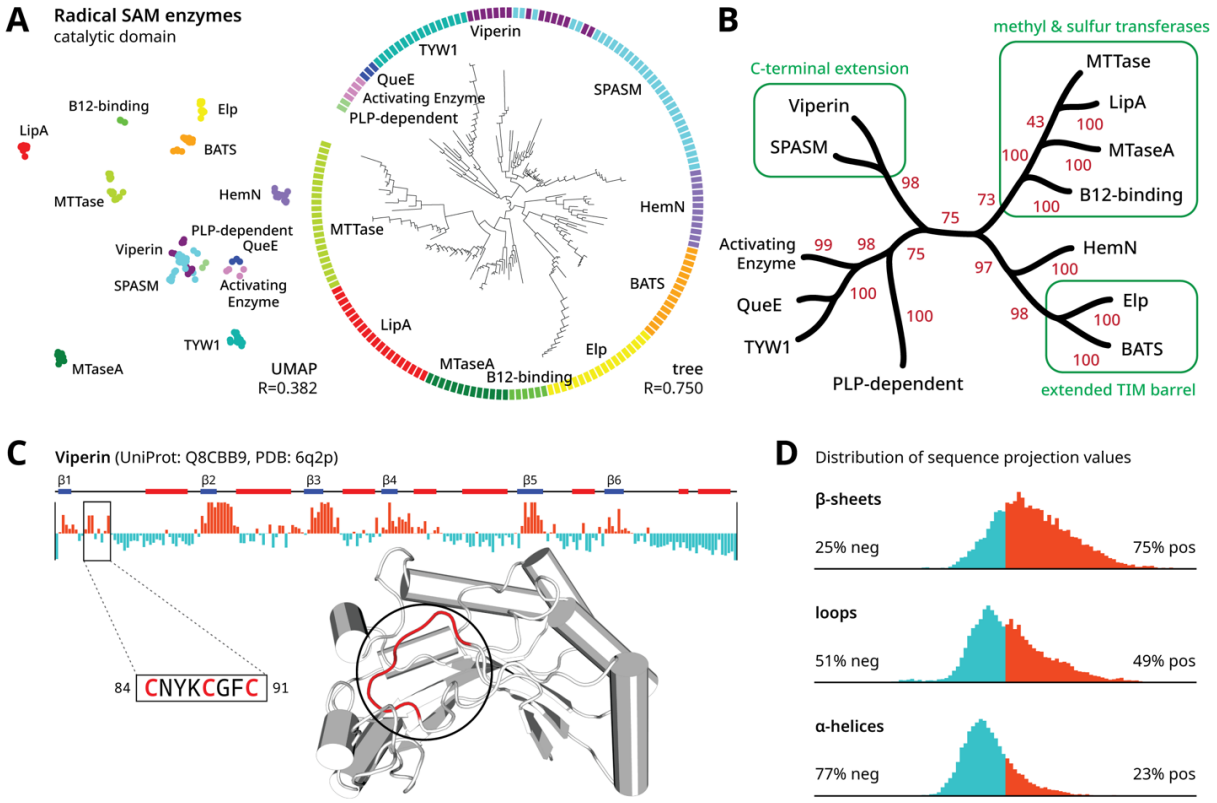


Figure 4.5: Embedding-based analysis of the diverse radical SAM enzymes
(A) An embedding tree of radical SAM enzymes in a circular layout with major groups labeled. This tree was generated using the [CLF] representation function and cosine distance. To the left of the tree, we plot a UMAP projection using the same dataset. At the bottom of each graph, we provide the correlation coefficient which quantifies how well the all-vs-all pairwise distances denoted by each visualization reflect the pairwise distances from the underlying dataset. (B) A condensed tree showing various families of radical SAM enzymes. VIBEs are indicated by the red percentage values. Select structure-functional annotations are shown in green. (C) A sequence projection for a representative radical SAM enzyme. The conserved iron-sulfur cluster binding motif, CxxxCxΦC, corresponds to a positive peak, designated by the zoomed inset. A crystal structure of mouse Viperin is shown with the CxxxCxΦC motif circled and highlighted red. Based on the optimal tree parameters, sequence projections were calculated using the [CLF] representation function. (D) We plot histograms of sequence projection values across our dataset of diverse radical SAM enzymes, stratified by the secondary structure at each residue.

(B12-binding, MTaseA, LipA, and MTTase families) [51] were placed in a single clade. Placed in the neighboring clade, some HemN enzymes also catalyze methyl transfer [53,54]. The HemN and Elp families have reported sequence similarity [55], while Elp and BAT families both conserve extended TIM barrel folds. Additionally, many Elp and BATS enzymes contain alterations to the canonical CxxxCxΦC motif [56]. Viperin and SPASM families both conserve a C-terminal extension which facilitates family-specific functionalities [57]. Viperin is placed closest to the MoaA subfamily (within the SPASM family); both of which act on nucleotide substrates [58]. Activating enzyme and QueE family members sometimes adopt a “Tiny TIM" minimal core fold [59]. QueE and TYW1 families are also closely grouped together; both families are involved in tRNA biosynthesis and hypermodification [60].

Sequence projections across diverse radical SAM enzymes place a positive peak at the conserved CxxxCxΦC motif **(Figure 4.5C)**. Positive peaks also tend to fall on β-sheets **(Figure 4.5D)** extending onto each proceeding loop. These regions correspond to previously identified SAM binding sites found in all radical SAM enzymes, as well as family-specific motifs which facilitate unique family-specific chemistry [61,62]. This trend suggests that the usage of β-sheets in substrate binding and catalysis may be a shared feature across the radical SAM superfamily. Although β-sheets are more conserved than loops and helices [63,64], sequence projections on other globular protein superfamilies show comparatively weaker association with β-sheets.

4.3 Discussion

We present an arsenal of orthogonal techniques, listed in **Table 4.1**, for alignment-free protein sequence analysis by utilizing sequence embeddings as a proxy for actual amino acid sequences. Throughout diverse case studies, embedding vectors appear most suited for modeling long-distance evolutionary relationships **(Figure 4.1D)**, allowing us to infer a single tree

containing all human kinase-fold enzymes **(Figure 4.3A)**, identify similarities between divergent phosphatases structural folds which likely arise by convergent evolution **(Figure 4.4A-C)**, and infer the initial tree of the radical SAM enzyme superfamily **(Figure 4.5A-B)**. Across all case studies, closely-related proteins had a tendency towards unbalanced, ladder-like topologies with zero branch length tips, suggesting that embedding trees do not have the capacity to resolve closely-related proteins within the same family. While our analyses only utilized shared catalytic domains, a focused analysis on closely-related sequences may benefit from embeddings that include shared regions beyond the catalytic domain. In comparison, while sequence alignments-based approaches are not well suited for long-range evolutionary inference, they work well on closely related sequences. A combination of alignment and alignment-free embedding approaches are expected to advance the frontiers of sequence analysis.

Embedding trees indicate that the protein LM provides a reasonable model of the theoretical evolutionary landscape. Although it is technically possible to build embedding trees from any sequence dataset, evolutionary inference should only be invoked if common ancestry can be supported by orthogonal evidence. Common ancestry between kinase fold enzymes is supported by a highly conserved structural fold and sequence motifs [34]. Although phosphatase enzymes share a common catalytic function, different folds utilize different mechanisms which indicate that these enzymes independently emerged multiple times throughout evolution [43]. Consequently, the phosphatase fold tree only should be interpreted as clustering. Despite methodological differences, many established principles in phylogenetic analyses remain relevant such as generating replicate trees and being vigilant towards confounds such as long-branch attraction [73].

	alignment-based methods	embedding-based methods
sequence comparison	sequence alignment [65]	embedding distance
residue conservation	statistical entropy (e.g. Shannon entropy, Kullback-Leibler divergence, Jensen-Shannon divergence) [2] sequence logo [66]	sequence projections
sequence clustering	sequence similarity networks[67]	embedding trees manifold learning (e.g. t-SNE [68], UMAP [27,28])
tree inference	probabilistic methods (e.g. maximum likelihood [69], Bayesian inference [70]) distance matrix methods (e.g. neighbor-joining [71])	embedding trees (assuming sufficient orthogonal evidence)
branch support	Bootstrapping [72]	VIBEs

Table 4.1: Comparison of alignment and embedding-based methods for protein sequence analysis
For a diverse range of sequence alignment-based methods, we define an analogous embedding-based approach.

Beyond biological applications, our study provides useful methods for explainable machine learning. Tree-based visualizations more accurately capture the global data structure of high-dimensional data compared to manifold learning algorithms such as UMAP. Citing a major difference, our tree-based method does not frame manifold visualization as a dimensionality reduction problem; trees are inherently capable of depicting high dimensional relationships without assuming an underlying geometry. Sequence projections can also be used as explainability vectors. Not requiring backward gradient calculations, our method demonstrates superior computational efficiency and simplicity. By showcasing these new applications, we hope to promote the development of better LMs. Recent results have proposed mechanisms for

generating fixed-sized embeddings from variable-size inputs [74,75] which would potentially exclude the need for representation functions. Further advances in the field of representation learning are expected to improve the unsupervised classification of large protein families.

4.4 Methods

4.4.1 Data collection and preprocessing

The sequence dataset of 558 human kinase-fold enzymes [76] and 204 human phosphatase sequences [43] were derived from previously published studies. Our dataset of 179 taxonomically diverse radical SAM enzymes was manually curated based on a previous sequence clustering study [50]. Core domain segments were manually identified and trimmed based on all available crystal structures and AlphaFold2 [52] models. Secondary structure annotations were assigned based on AlphaFold2 models using the DSSP algorithm [77].

4.4.2 Calculating embedding trees

Following sequence dataset curation, the sequences were converted into embedding vectors using a Transformer-based protein LM. Specifically, the embedding vector is the final hidden state generated from the last layer of the encoder module.

$$Embeddings = Transformer(ProteinSequences)$$

Each embedding is a two-dimensional matrix. The token dimension encodes one token for each residue of the original sequence plus additional special tokens which are appended during preprocessing, while the embedding dimension encodes information about each token. The number of special tokens and the size of the embedding dimension will vary depending on the specific LM used. To enable direct comparisons between embeddings, we derive

representation functions to summarize the information encoded within the variably-sized embedding vectors into fixed-sized representation vectors. This is conceptually similar to pooling operations, typically used to condense information within convolutional architectures. Each representation function is applied along the token dimension of the embedding, defined as a function of the special tokens or sequence tokens. We explored 8 pretrained protein LMs and 9 different representation functions. After sampling all compatible pairs of protein LM and representation function, we generated 56 unique sets of representation vectors for each sequence dataset.

$$Representations = RepresentationFunction(Embeddings)$$

We explored a variety of distance metrics to calculate an all-vs-all distance matrix from the representation vectors: Euclidean distance, cosine distance, Manhattan distance, geodesic distance, and TS-SS [32]. We sampled each unique combination of representation vectors and distance metrics to generate 280 unique distance matrices for each sequence dataset.

$$DistanceMatrix = DistanceMetric(Representations)$$

Trees were calculated from distance matrices using the neighbor-joining algorithm [71].

$$EmbeddingTree = NeighborJoining(DistanceMatrix)$$

4.4.3 Evaluating branch support

The statistical confidence of a given bipartition of a tree can be evaluated using a VAE. We trained a VAE on a fixed set of representation vectors to resample replicate representation vectors. By learning a smooth latent state representation from the input data, the VAE becomes capable of regenerating the input data using the reparameterization trick which allows

backpropagation through a random node. This unique property of VAE enabled us to resample the original input with any desired number of replicates. To accurately model the underlying space of the protein representations, the VAE is trained on optimizing a combination of Mean Square Error (MSE), Kullback-Leibler divergence (KLD), and TS-SS Error (TSE). We applied the cosine annealing [78] to control for the weight of the KLD loss term.

$$Loss = \alpha \cdot MSE + \beta \cdot KLD + \gamma \cdot TSE$$

$$where\ \beta = Cosine(mod(Iteration - 1,\ MaxIteration)/\ MaxIteration)$$

We trained a separate VAE for each unique dataset of representations. VAEs were trained for 20,000 epochs with early stopping patience of 1000 epochs. Resampled representation vectors generated from the final model were used to build 500 replicate trees. Branch support values were assigned to the original tree using the replicate trees. We refer to this procedure and confidence metric as VAE Implemented Branch Support Estimation (VIBE).

4.4.5 Calculating sequence projections

To understand how a representation vector (generated from a given representation function) encodes an embedding, we calculated the cosine distance between the representation vector and each sequence token of the embedding. The resulting sequence projection vector has the same size as the protein sequence corresponding to the embedding. Sequence projections were further standardized to facilitate comparisons between sequences.

$$SequenceProjection = Standardization(CosineDistance(Representation, Embedding))$$

Bibliography

[1] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman, Pfam: The protein families database in 2021, Nucleic Acids Research. 49 (2021) D412–D419. https://doi.org/10.1093/nar/gkaa913.

[2] J.A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics. 23 (2007) 1875–1882. https://doi.org/10.1093/bioinformatics/btm270.

[3] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E. 87 (2013) 012707. https://doi.org/10.1103/PhysRevE.87.012707.

[4] Z. Yang, B. Rannala, Molecular phylogenetics: principles and practice, Nat Rev Genet. 13 (2012) 303–314. https://doi.org/10.1038/nrg3186.

[5] B. Rost, Twilight zone of protein sequence alignments, Protein Engineering, Design and Selection. 12 (1999) 85–94. https://doi.org/10.1093/protein/12.2.85.

[6] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research. 25 (1997) 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

[7] L.S. Johnson, S.R. Eddy, E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure, BMC Bioinformatics. 11 (2010) 431. https://doi.org/10.1186/1471-2105-11-431.

[8] M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nat Biotechnol. 35 (2017) 1026–1028. https://doi.org/10.1038/nbt.3988.

[9] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics. 28 (2012) 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

[10] R. Giancarlo, S.E. Rombo, F. Utro, Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies, Briefings in Bioinformatics. 15 (2014) 390–406. https://doi.org/10.1093/bib/bbt088.

[11] A. Zielezinski, S. Vinga, J. Almeida, W.M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biology. 18 (2017) 186. https://doi.org/10.1186/s13059-017-1319-7.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference

on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, 2017: pp. 6000–6010.

[13] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, PNAS. 118 (2021). https://doi.org/10.1073/pnas.2016239118.

[14] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing, IEEE Transactions on Pattern Analysis and Machine Intelligence. (2021) 1–1. https://doi.org/10.1109/TPAMI.2021.3095381.

[15] UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204-212. https://doi.org/10.1093/nar/gku989.

[16] C. Dallago, K. Schütze, M. Heinzinger, T. Olenyi, M. Littmann, A.X. Lu, K.K. Yang, S. Min, S. Yoon, J.T. Morton, B. Rost, Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets, Current Protocols. 1 (2021) e113. https://doi.org/10.1002/cpz1.113.

[17] M. Littmann, N. Bordin, M. Heinzinger, K. Schütze, C. Dallago, C. Orengo, B. Rost, Clustering FunFams using sequence embeddings improves EC purity, Bioinformatics. 37 (2021) 3449–3455. https://doi.org/10.1093/bioinformatics/btab371.

[18] M. Littmann, M. Heinzinger, C. Dallago, T. Olenyi, B. Rost, Embeddings from deep learning transfer GO annotations beyond homology, Sci Rep. 11 (2021) 1160. https://doi.org/10.1038/s41598-020-80786-0.

[19] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, A. Rives, Transformer protein language models are unsupervised structure learners, (2020) 2020.12.15.422761. https://doi.org/10.1101/2020.12.15.422761.

[20] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, A.G. Murzin, SCOP2 prototype: a new approach to protein structure mining, Nucleic Acids Research. 42 (2014) D310–D314. https://doi.org/10.1093/nar/gkt1242.

[21] C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins., The EMBO Journal. 5 (1986) 823–826. https://doi.org/10.1002/j.1460-2075.1986.tb04288.x.

[22] H. Hark Gan, R.A. Perlow, S. Roy, J. Ko, M. Wu, J. Huang, S. Yan, A. Nicoletta, J. Vafai, D. Sun, L. Wang, J.E. Noah, S. Pasquali, T. Schlick, Analysis of Protein Sequence/Structure Similarity Relationships, Biophysical Journal. 83 (2002) 2781–2791. https://doi.org/10.1016/S0006-3495(02)75287-9.

[23] M. Anisimova, M. Gil, J.-F. Dufayard, C. Dessimoz, O. Gascuel, Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes, Systematic Biology. 60 (2011) 685–699. https://doi.org/10.1093/sysbio/syr041.

[24] M.J. Sackin, "Good" and "Bad" Phenograms, Systematic Biology. 21 (1972) 225–226. https://doi.org/10.1093/sysbio/21.2.225.

[25] J. Sukumaran, M.T. Holder, DendroPy: a Python library for phylogenetic computing, Bioinformatics. 26 (2010) 1569–1571. https://doi.org/10.1093/bioinformatics/btq228.

[26] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics. 20 (1987) 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

[27] L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection, Journal of Open Source Software. 3 (2018) 861. https://doi.org/10.21105/joss.00861.

[28] A. Narayan, B. Berger, H. Cho, Assessing single-cell transcriptomic variability through density-preserving data visualization, Nat Biotechnol. 39 (2021) 765–774. https://doi.org/10.1038/s41587-020-00801-7.

[29] J. Huang, D. Tang, W. Zhong, S. Lu, L. Shou, M. Gong, D. Jiang, N. Duan, WhiteningBERT: An Easy Unsupervised Sentence Embedding Approach, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021: pp. 238–244. https://doi.org/10.18653/v1/2021.findings-emnlp.23.

[30] M. Steinbach, L. Ertöz, V. Kumar, The Challenges of Clustering High Dimensional Data, in: L.T. Wille (Ed.), New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition, Springer, Berlin, Heidelberg, 2004: pp. 273–309. https://doi.org/10.1007/978-3-662-08968-2_16.

[31] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, ArXiv:1312.6114 [Cs, Stat]. (2014). http://arxiv.org/abs/1312.6114 (accessed February 1, 2022).

[32] A. Heidarian, M.J. Dinneen, A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering, in: 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), 2016: pp. 142–151. https://doi.org/10.1109/BigDataService.2016.14.

[33] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The Protein Kinase Complement of the Human Genome, Science. 298 (2002) 1912–1934. https://doi.org/10.1126/science.1075762.

[34] N. Kannan, A.F. Neuwald, Did Protein Kinase Regulatory Mechanisms Evolve Through Elaboration of a Simple Structural Component?, Journal of Molecular Biology. 351 (2005) 956–972. https://doi.org/10.1016/j.jmb.2005.06.057.

[35] C.J. Leonard, L. Aravind, E.V. Koonin, Novel Families of Putative Protein Kinases in Bacteria and Archaea: Evolution of the "Eukaryotic" Protein Kinase Superfamily, Genome Res. 8 (1998) 1038–1047. https://doi.org/10.1101/gr.8.10.1038.

[36] M. Wilkinson, Majority-rule reduced consensus trees and their use in bootstrapping., Molecular Biology and Evolution. 13 (1996) 437–444. https://doi.org/10.1093/oxfordjournals.molbev.a025604.

[37] W. Yeung, Z. Ruan, N. Kannan, Emerging roles of the αC-β4 loop in protein kinase structure, function, evolution, and disease, IUBMB Life. 72 (2020) 1189–1202. https://doi.org/10.1002/iub.2253.

[38] L.J. Fulcher, G.P. Sapkota, Functions and regulation of the serine/threonine protein kinase CK1 family: moving beyond promiscuity, Biochemical Journal. 477 (2020) 4603–4621. https://doi.org/10.1042/BCJ20200506.

[39] D. Baum, Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups, Scitable by Nature Education. (2008). http://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956 (accessed February 1, 2022).

[40] S.K. Hanks, T. Hunter, The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification1, The FASEB Journal. 9 (1995) 576–596. https://doi.org/10.1096/fasebj.9.8.7768349.

[41] N. Kannan, A.F. Neuwald, Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha, Protein Sci. 13 (2004) 2059–2077. https://doi.org/10.1110/ps.04637904.

[42] W. Yeung, A. Kwon, R. Taujale, C. Bunn, A. Venkat, N. Kannan, Evolution of Functional Diversity in the Holozoan Tyrosine Kinome, Molecular Biology and Evolution. 38 (2021) 5625–5639. https://doi.org/10.1093/molbev/msab272.

[43] M.J. Chen, J.E. Dixon, G. Manning, Genomics and evolution of protein phosphatases, Science Signaling. (2017). https://doi.org/10.1126/scisignal.aag1796.

[44] M.Y. Galperin, E.V. Koonin, A. Bairoch, A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases, Protein Science. 7 (1998) 1829–1835. https://doi.org/10.1002/pro.5560070819.

[45] E. Kuznetsova, B. Nocek, G. Brown, K.S. Makarova, R. Flick, Y.I. Wolf, A. Khusnutdinova, E. Evdokimova, K. Jin, K. Tan, A.D. Hanson, G. Hasnain, R. Zallot, V. de Crécy-Lagard, M. Babu, A. Savchenko, A. Joachimiak, A.M. Edwards, E.V. Koonin, A.F. Yakunin, Functional Diversity of Haloacid Dehalogenase Superfamily Phosphatases from

Saccharomyces cerevisiae: BIOCHEMICAL, STRUCTURAL, AND EVOLUTIONARY INSIGHTS*, Journal of Biological Chemistry. 290 (2015) 18678–18698. https://doi.org/10.1074/jbc.M115.657916.

[46] A. Alonso, J. Sasin, N. Bottini, I. Friedberg, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon, T. Mustelin, Protein Tyrosine Phosphatases in the Human Genome, Cell. 117 (2004) 699–711. https://doi.org/10.1016/j.cell.2004.05.018.

[47] Y. Shi, Serine/Threonine Phosphatases: Mechanism through Structure, Cell. 139 (2009) 468–484. https://doi.org/10.1016/j.cell.2009.10.006.

[48] G. Swarup, S. Cohen, D.L. Garbers, Selective dephosphorylation of proteins containing phosphotyrosine by alkaline phosphatases., Journal of Biological Chemistry. 256 (1981) 8197–8201. https://doi.org/10.1016/S0021-9258(18)43408-4.

[49] A. Chakrabartty, R.A. Stinson, Properties of membrane-bound and solubilized forms of alkaline phosphatase from human liver, Biochimica et Biophysica Acta (BBA) - General Subjects. 839 (1985) 174–180. https://doi.org/10.1016/0304-4165(85)90034-0.

[50] G.L. Holliday, E. Akiva, E.C. Meng, S.D. Brown, S. Calhoun, U. Pieper, A. Sali, S.J. Booker, P.C. Babbitt, Atlas of the Radical SAM Superfamily: Divergent Evolution of Function Using a “Plug and Play” Domain, in: V. Bandarian (Ed.), Methods in Enzymology, Academic Press, 2018: pp. 1–71. https://doi.org/10.1016/bs.mie.2018.06.004.

[51] J.B. Broderick, B.R. Duffus, K.S. Duschene, E.M. Shepard, Radical S-Adenosylmethionine Enzymes, Chem. Rev. 114 (2014) 4229–4317. https://doi.org/10.1021/cr4004709.

[52] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[53] J.W. LaMattina, D.B. Nix, W.N. Lanzilotta, Radical new paradigm for heme degradation in Escherichia coli O157:H7, PNAS. 113 (2016) 12138–12143. https://doi.org/10.1073/pnas.1603209113.

[54] W. Ding, Y. Li, J. Zhao, X. Ji, T. Mo, H. Qianzhu, T. Tu, Z. Deng, Y. Yu, F. Chen, Q. Zhang, The Catalytic Mechanism of the Class C Radical S-Adenosylmethionine Methyltransferase NosN, Angewandte Chemie International Edition. 56 (2017) 3857–3861. https://doi.org/10.1002/anie.201609948.

[55] C. Paraskevopoulou, S.A. Fairhurst, D.J. Lowe, P. Brick, S. Onesti, The Elongator subunit Elp3 contains a Fe4S4 cluster and binds S-adenosylmethionine, Molecular Microbiology. 59 (2006) 795–806. https://doi.org/10.1111/j.1365-2958.2005.04989.x.

[56] A.S. Byer, E.M. Shepard, J.W. Peters, J.B. Broderick, Radical S-Adenosyl-l-methionine Chemistry in the Synthesis of Hydrogenase and Nitrogenase Metal Cofactors*, Journal of Biological Chemistry. 290 (2015) 3987–3994. https://doi.org/10.1074/jbc.R114.578161.

[57] M.K. Fenwick, D. Su, M. Dong, H. Lin, S.E. Ealick, Structural Basis of the Substrate Selectivity of Viperin, Biochemistry. 59 (2020) 652–662. https://doi.org/10.1021/acs.biochem.9b00741.

[58] A. Bernheim, A. Millman, G. Ofir, G. Meitav, C. Avraham, H. Shomar, M.M. Rosenberg, N. Tal, S. Melamed, G. Amitai, R. Sorek, Prokaryotic viperins produce diverse antiviral molecules, Nature. 589 (2021) 120–124. https://doi.org/10.1038/s41586-020-2762-2.

[59] D.P. Dowling, N.A. Bruender, A.P. Young, R.M. McCarty, V. Bandarian, C.L. Drennan, Radical SAM enzyme QueE defines a new minimal core fold and metal-dependent mechanism, Nat Chem Biol. 10 (2014) 106–112. https://doi.org/10.1038/nchembio.1426.

[60] O. Berteau, A. Benjdia, DNA Repair by the Radical SAM Enzyme Spore Photoproduct Lyase: From Biochemistry to Structural Investigations, Photochemistry and Photobiology. 93 (2017) 67–77. https://doi.org/10.1111/php.12702.

[61] D.P. Dowling, J.L. Vey, A.K. Croft, C.L. Drennan, Structural diversity in the AdoMet radical enzyme superfamily, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1824 (2012) 1178–1195. https://doi.org/10.1016/j.bbapap.2012.04.006.

[62] T.A.J. Grell, A.P. Young, C.L. Drennan, V. Bandarian, Biochemical and Structural Characterization of a Schiff Base in the Radical-Mediated Biosynthesis of 4-Demethylwyosine by TYW1, J. Am. Chem. Soc. 140 (2018) 6842–6852. https://doi.org/10.1021/jacs.8b01493.

[63] G. Abrusán, J.A. Marsh, Alpha Helices Are More Robust to Mutations than Beta Strands, PLOS Computational Biology. 12 (2016) e1005242. https://doi.org/10.1371/journal.pcbi.1005242.

[64] K. Illergård, D.H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence—A study of structural response in protein cores, Proteins: Structure, Function, and Bioinformatics. 77 (2009) 499–508. https://doi.org/10.1002/prot.22458.

[65] J.D. Thompson, B. Linard, O. Lecompte, O. Poch, A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives, PLOS ONE. 6 (2011) e18093. https://doi.org/10.1371/journal.pone.0018093.

[66] T.D. Schneider, R.M. Stephens, Sequence logos: a new way to display consensus sequences, Nucleic Acids Research. 18 (1990) 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

[67] H.J. Atkinson, J.H. Morris, T.E. Ferrin, P.C. Babbitt, Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies, PLOS ONE. 4 (2009) e4345. https://doi.org/10.1371/journal.pone.0004345.

[68] L. van der Maaten, G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research. 9 (2008).

[69] J. Felsenstein, Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters, Systematic Biology. 22 (1973) 240–249. https://doi.org/10.1093/sysbio/22.3.240.

[70] J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics. 17 (2001) 754–755. https://doi.org/10.1093/bioinformatics/17.8.754.

[71] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., Molecular Biology and Evolution. 4 (1987) 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

[72] B. Efron, E. Halloran, S. Holmes, Bootstrap confidence levels for phylogenetic trees, PNAS. 93 (1996) 13429–13429. https://doi.org/10.1073/pnas.93.23.13429.

[73] J. Bergsten, A review of long-branch attraction, Cladistics. 21 (2005) 163–193. https://doi.org/10.1111/j.1096-0031.2005.00059.x.

[74] T. Gao, X. Yao, D. Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021: pp. 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552.

[75] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019: pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

[76] L.-C. Huang, R. Taujale, N. Gravel, A. Venkat, W. Yeung, D.P. Byrne, P.A. Eyers, N. Kannan, KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases, BMC Bioinformatics. 22 (2021) 446. https://doi.org/10.1186/s12859-021-04358-3.

[77] W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers. 22 (1983) 2577–2637. https://doi.org/10.1002/bip.360221211.

[78] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, L. Carin, Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019: pp. 240–250. https://doi.org/10.18653/v1/N19-1021.

Chapter 5

DISCUSSION AND CONCLUDING REMARKS

5.1 Achievement of goals

I have developed and applied novel computational approaches towards the study of diverse enzymes, with a focus on the protein kinase superfamily. In order to address the major research question described in Chapter 1, I have developed an array-based library for manipulating and traversing large multiple sequence alignments which is available on GitHub:

https://github.com/waylandy/HelperBunny.

Using the framework established by my library, I conducted an in-depth study of sequence variations of the protein kinase αC-β4 loop region (described in Chapter 2) [1] as well as an evolutionary characterization of diverse holozoan tyrosine kinases (described in Chapter 3) [2]. I also developed a new class of methods for protein sequence analysis using embeddings

learned from Transformer protein language models (described in Chapter 4) which were applied towards studying protein kinases, phosphatase enzymes, and radical S-Adenosyl-L-Methionine (SAM) enzymes.

5.1.1 Evolutionary history of protein kinase enzymes

My work has shed further light upon the complex evolutionary history of the protein kinase superfamily. While protein kinases are known to be distantly related to small molecule and lipid kinases [3], their exact relationship has been difficult to determine due to the difficulties in reliably aligning highly divergent sequences. Circumventing this issue, I developed an alternative quantitative approach for alignment-independent phylogenetic inference and applied my method towards inferring a new phylogenetic tree encompassing the human kinase fold enzyme superfamily. This new phylogenetic tree confidently places the CK1 group as the most basal protein kinase clade, while the placement of other major protein kinase groups largely remains in agreement with the currently accepted phylogeny of the human protein kinases [4]. Using traditional alignment-based methods, I have also inferred an updated phylogenetic tree for the tyrosine kinase group, which generalizes beyond the widely-used human-centric model by accounting for diverse holozoan organisms. Furthermore, our evolutionary model is orthogonally validated by conserved intron positions/phases as well as a variety of family-specific structure-functional features.

5.1.2 Sequence variations across protein kinase enzymes

Leveraging large structurally validated sequence alignments, I cataloged diverse sequence variation across the protein kinase superfamily. Through comparative sequence analyses, I described protein kinase group-specific sequence variations within the αC-β4 loop

region. Most protein kinases conserve the HxN motif, except for the AGC group, which conserves the HPF motif at the equivalent position, and the CK1 group, which lacks any conserved motif at the equivalent position. Using the BPPS algorithm [5], I also defined ~50 distinct families of tyrosine kinases — each characterized by family-specific sequence motifs which reflect family-specific structure-function. While many of these family-specific motifs map to known regulatory sites such as the Src-like inactive conformation [6] (SrcM-specific) or the molecular brake [7] (FPVR-specific), many of these motifs have not been characterized. Thus, these family-specific motifs serve as excellent targets for future studies.

5.1.3 Other contributions

Beyond the main contributions described in this dissertation, I have also contributed to a number of other published research endeavors: cataloging orthologs for each human protein kinase [8], cataloging protein domain organizations for major pseudokinase families across the protein kinome [9], cataloging variations of the AGC tail TOR interacting motif across the major clades of eukaryotes [10], cataloging sequence variations associated with inter-domain interactions in Tec family kinases [11], inferring the evolution of cyclin and modelling CDK1-cyclin binding interfaces in ciliates [12], inferring the evolution of fold A glycosyltransferase enzymes [13], building an online database for browsing family-specific motifs across fold A glycosyltransferase enzymes [14], analyzing the folding landscape of diverse glycosyltransferase enzymes [15], predicting kinase drug responses [16], modeling conformational landscape and dynamics for MLKL family pseudokinases [17,18], docking and modeling drug binding dynamics in Tribbles 2 pseudokinase [19], and modeling the effects of phosphorylation on the dynamics of HIPK2 kinase [20].

5.2 Future directions

Leveraging large multiple sequence alignments, I have developed and applied new approaches for analyzing protein kinase evolution. Although my work in this dissertation specifically applies these methods towards in-depth analyses of the αC-β4 loop region and the tyrosine kinase group, these methodologies could also be generalized towards analyzing other conserved regions of the kinase domain or other kinase groups/families. Future studies may also be aided by recent advances in artificial intelligence-guided protein structural prediction [21,22]. The incorporation of high-quality protein models may help further unravel the complex relationships between protein kinase sequence, structure, and function.

Born from the intersection of bioinformatics and representation learning, embedding-based protein sequence analyses represent a brand-new class of methods with lots of room for further development. The representation of protein sequences as numerical matrices, rather than discrete strings, presents a wide range of new possibilities for unsupervised analysis. Alignment-independent evolutionary inference is just one of many potential applications for the embedding-based analytical framework.

5.2.1 An updated evolutionary model for the protein kinase superfamily

Our current understanding of protein kinase evolution is still largely based on kinome studies of various model organisms published around two decades ago [4,23]. In the post-genomic era, a rapidly increasing wealth of sequence information and new computational methods continuously provides better resources for a more holistic characterization. This dissertation takes advantage of these advances towards producing an updated evolutionary model for tyrosine kinase evolution. However, these methods could be re-applied towards

characterizing the remaining protein kinase groups using a similar alignment-based approach. The curation of structurally validated sequence alignments for each of the major groups will be accelerated by the newfound availability of high-quality structural models from the AlphaFold model [21] and database [24]. In conjunction with experimentally solved structures in the PDB, AlphaFold models would also be helpful for predicting the structural context of kinase family-specific sequence motifs or insertion segments.

To provide an example, my studies identified that the αC-β4 loop is typically eight residues long. However, a sizable subset of protein kinases sequences exhibits an extended αC-β4 loop segment. While I was able to analyze kinase families which had experimentally determined structures, the AlphaFold database would have facilitated a broader analysis for identifying unique extended αC-β4 loop conformations across the full protein kinome.

To provide another example, my studies of the tyrosine kinase group identified family-specific sequence motifs which, based on available crystal structures, seemed to be involved with stabilizing family-specific kinase inactive conformations. I hypothesized that tyrosine kinases which conserve these motifs are also likely to adopt their associated inactive conformations. For instance, I predicted that the Lmr family kinases would be capable of adopting the IRK-like inactive conformation, which was later confirmed in a recent crystallographic study [25]. Although AlphaFold models should not be used as a replacement for experimentally determined protein structures, future studies might benefit from utilizing the AlphaFold models to speculate upon potential regulatory interactions mediated by family-specific motifs.

In my final study, I proposed that embedding-based methods provide the means to infer a unified model that describes the evolutionary relationships between all kinase-fold enzymes. My

embedding tree of the human kinase could be expanded to encompass diverse kinases across all domains of life. Personally, I believe that the evolutionary characterization of divergent protein families such as SelO are of particular interest; SelO adopts the bilobal kinase fold but shares minimal sequence similarity to other known kinase fold enzymes and catalyzes AMPylation rather than phosphorylation [26]. Although the human ortholog was included in the human embedding tree, it was placed as a long branch in the outgroup which suggests that SelO is very unlike any of the other kinase fold enzymes in humans. Beyond humans, there are many taxa-specific kinase families [27,28], the inclusion of which should provide additional evolutionary context. By taking advantage of the sheer volume of available protein sequence information, I believe that it is possible to characterize the evolution of kinase fold enzymes across the entire tree of life using a combination of alignment and embedding-based approaches.

5.2.2 Hybrid alignment-embedding approaches

While my dissertation specifically highlights strategies for using embedding vectors to facilitate alignment-independent sequence analysis, there are also potential strategies for integrating both alignment and embedding-based methodologies. Alignment-based strategies excel at modeling short evolutionary distances, while embedding-based strategies excel at modeling long evolutionary distances. By combining these complementary strengths, the development of hybrid alignment-embedding techniques would facilitate the creation of broader and more extensive phylogenetic models — capable of inferring evolutionary relationships between distantly-related protein superfamilies while also capturing the granular relationships of proteins within the superfamily.

I propose a potential strategy for inferring a phylogenetic tree using a hybrid alignment-embedding approach. Given a dataset of divergent protein sequences that likely share a common evolutionary origin, start by building an embedding-based tree. Assuming that these sequences are too divergent to reliably produce a single multiple sequence alignment, identify major clades which can be reliably aligned and create separate alignments corresponding to different regions of the embedding tree. Infer alignment-based trees from each of these separate alignments, then replace the corresponding branches of the embedding tree with each alignment-based tree. Evolutionary relationships near the base of the tree will be inferred by embedding-based methods, while relationships near the tips of the tree will be inferred by alignment-based methods. In summary, the weakness of embedding trees can be resolved by iteratively refining the topology using traditional alignment-based evolutionary inference.

Another potential strategy for conducting a hybrid alignment-embedding approach for evolutionary inference is to use utilize a Transformer that encodes multiple sequence alignments rather than just sequences. Transformers are a general neural network architecture which can be trained to create embedding vectors for a wide variety of inputs such as English text [29], amino acid sequences [30,31], nucleotide sequences [32], and multiple sequence alignments [33]. Alignment embeddings vectors can be used in a very similar fashion as sequence embedding vectors. Given a large protein sequence dataset, one could cluster the sequences into distinct families, then produce individual multiple sequence alignments for each family. Upon converting each alignment into an embedding vector, one could create an embedding tree where each tip represents an entire protein family rather than a single sequence — conceptually to HMM-based phylogenetic trees [34].

Hybrid alignment-embedding methods also have diverse applications beyond evolutionary inference. Sequence embedding vectors can also be used as a heuristic for aligning divergent protein sequences. I provide an example using sequence projection vectors, calculated as a function of the sequence embedding. Given a pair of sequence projection vectors which correspond to two distantly related protein sequences, projection vectors could be aligned using the dynamic time warp algorithm [35]. Mapping the equivalent sequence alignment from a time warp is a potential strategy for identifying homologous sequence regions. Sequence embeddings may also be used as a heuristic for rapidly identifying homologous segments across a set of sequences by Fourier transform [36], which may have potential applications in multiple sequence alignments and remote homolog detection.

5.2.3 Representing concepts using sets of sequence embeddings

The most significant advantage of sequence embedding vectors is their ability to be modified by numerical matrix operations. Within this dissertation, I utilized this important property to facilitate alignment-independent sequence comparisons and infer fast/slow evolving sites. Another potential application is to combine multiple sequence embedding vectors to represent larger concepts such as an entire organism. This idea could be expanded towards developing an embedding-based method for inferring species evolution.

Gene trees infer the evolution of a single genetic locus, which may not necessarily agree with the evolution of the species [37]. One method of estimating a species tree is to calculate a consensus tree from a collection of gene trees [38]. This method can also be applied for embedding trees. Alternatively, one could represent a specific organism as a function of its constituent genes by either pooling or concatenating a set of representation vectors to create a

new fixed-size array. One could infer a species tree by representing each species using a fixed-size array generated from a predefined set of well-conserved genes.

This method would also be useful for modeling disease states using data from cancer sequencing studies. A patient sample could be represented as a function of its proteome, which may include a mix of mutant and wild-type sequences. Furthermore, gene expression data may also be incorporated. The representation of specific disease states as fixed-size arrays would be useful for downstream manifold clustering analyses as well as potential input features for the supervised prediction of cancer-related targets.

5.2.4 Embedding-based network analyses

Embedding vectors can be applied to many data analysis methods beyond what was explicitly shown in this dissertation. For example, embedding-based methods are also applicable in network-based approaches for protein sequence analysis which are applicable for analyzing larger datasets [39]. All-vs-all comparisons of embedding distance can be used to build networks for clustering large sequence datasets by defining edges based on a minimum cutoff value. Embedding distance matrices may also be amenable for inferring phylogenetic networks using the median-joining algorithm [40].

5.2.5 Alignment-independent ancestral sequence reconstruction

My dissertation describes one method for generating embedding trees using the neighbor-joining algorithm [41]; however there are plenty of alternative strategies which may offer unique advantages. Under an alignment-based framework, maximum likelihood methods can be used to infer a phylogenetic tree [42] and infer ancestral sequences which appear inside the tree [43]. As

a major limitation, alignment-based ancestral sequence reconstruction is unable to infer insertion regions because they do not appear within the alignment. I believe that utilizing a maximum likelihood approach alongside an embedding-based framework may facilitate a means for alignment-independent ancestral sequence reconstruction, capable of modeling insertions. This would be a significant undertaking that would require the development of several prerequisite methods.

One key advancement would be the development of a Transformer protein language model which produces fixed-size embedding vectors. Current state-of-the-art Transformer models produce variable-sized embedding vectors which cannot be directly compared. Thus, representation functions were designed to resolve this issue by irreversibly summarizing the information within a variable-sized embedding vector into a fixed-size representation vector — conceptually similar to pooling operations [44]. However, the requirement of representation vectors precludes a direct strategy for regenerating the original sequence using the Transformer's decoder. Recent research has suggested potential strategies for producing fixed-size embeddings which would facilitate direct comparisons [45,46]. The development of fixed-size embedding vectors would drastically simplify the process of analyzing and generating embedding trees by removing the need for testing different representation functions, thus reducing the number of hyperparameters. Most importantly, embedding vectors can be reversibly converted into a corresponding protein sequence using a Transformer protein language model.

Another key advancement would be the development of a likelihood-based approach for inferring embedding-based trees. Given that a protein sequence can be meaningfully represented as a fixed-size embedding or representation vector, this fixed-size vector symbolizes the coordinates for a specific point within a theoretical high-dimensional sequence space. My work

in this dissertation has shown that sequences that share higher sequence identity also tend to be located closer together within this high-dimensional space. Thus, the accumulation of random sequence variations can be modeled as a Wiener process across this high-dimensional sequence space. Further expanding upon this model, gene duplication events can be modeled as making a copy of a point, then simulating two independent Wiener processes starting from the point of duplication. This iterative process of diffusion and duplications can be represented by a diffusion tree [47]. Applying these principles towards evolutionary inference, it should be possible to calculate the likelihood of a diffusion tree, given a set of fixed-size vectors which represent present-day sequences located at each tip and a rate of diffusion which corresponds to an overall evolutionary rate. By sampling diverse diffusion tree topologies [48], this function can be used to search for a maximum likelihood tree. Once a reasonable topology has been identified, it could be used to simulate the diffusion process and predict the coordinates of each extant sequences within the theoretical sequence space. If fixed-sized embedding vectors (and not representations) were used to model the sequence space, these coordinates could be converted back into an amino acid sequence using the decoder of its respective Transformer.

Extending beyond applications in ancestral sequence reconstruction, the development of fixed-size embedding vectors would also have useful applications in generative models such variational autoencoders [49], generative adversarial networks [50], and diffusion models [51,52]. These models typically utilize fixed-size inputs. If a generative model was trained on the embeddings of enzymes with specific properties, then any new embeddings produced by the model could be decodes back to a corresponding variable-length protein sequence. Overall, the reversible encoding of variable-length protein sequences within a fixed-length vector would be a

foundational advancement towards an unprecedented level of flexibility in generating synthetic amino acid sequences for protein engineering.

5.2.6 Heuristics for predicting functional residues

Conserved residues, determined from multiple sequence alignments, are an excellent heuristic for identifying structure-functionally important residues for further experimental characterization. However, this technique is incapable of quantifying differences between closely related sequences because their conservation statistics would be derived from the same alignment. Furthermore, unaligned regions are not quantified. My work suggests that sequence projections, calculated from embedding vectors, are highly correlated with sequence conservation. Unlike traditional alignment-based metrics which rely on first-order statistics, sequence projections can be calculated at the single-sequence resolution and account for the full sequence context. Consequently, sequence projections provide a more high-resolution heuristic for identifying the differences that distinguish sequences within the same family. Used in combination with alignment-based methods, these techniques provide useful methods for prioritizing the experimental characterization of functionally-relevant residues.

Bibliography

[1] W. Yeung, Z. Ruan, N. Kannan, Emerging roles of the αC-β4 loop in protein kinase structure, function, evolution, and disease, IUBMB Life. 72 (2020) 1189–1202. https://doi.org/10.1002/iub.2253.

[2] W. Yeung, A. Kwon, R. Taujale, C. Bunn, A. Venkat, N. Kannan, Evolution of Functional Diversity in the Holozoan Tyrosine Kinome, Molecular Biology and Evolution. 38 (2021) 5625–5639. https://doi.org/10.1093/molbev/msab272.

[3] N. Kannan, A.F. Neuwald, Did Protein Kinase Regulatory Mechanisms Evolve Through Elaboration of a Simple Structural Component?, Journal of Molecular Biology. 351 (2005) 956–972. https://doi.org/10.1016/j.jmb.2005.06.057.

[4] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, S. Sudarsanam, The Protein Kinase Complement of the Human Genome, Science. 298 (2002) 1912–1934. https://doi.org/10.1126/science.1075762.

[5] A.F. Neuwald, A Bayesian Sampler for Optimization of Protein Domain Hierarchies, Journal of Computational Biology. 21 (2014) 269–286. https://doi.org/10.1089/cmb.2013.0099.

[6] N.H. Shah, J.F. Amacher, L.M. Nocka, J. Kuriyan, The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases, Critical Reviews in Biochemistry and Molecular Biology. 53 (2018) 535–563. https://doi.org/10.1080/10409238.2018.1495173.

[7] H. Chen, J. Ma, W. Li, A.V. Eliseenkova, C. Xu, T.A. Neubert, W.T. Miller, M. Mohammadi, A Molecular Brake in the Kinase Hinge Region Regulates the Activity of Receptor Tyrosine Kinases, Molecular Cell. 27 (2007) 717–730. https://doi.org/10.1016/j.molcel.2007.06.028.

[8] L.-C. Huang, R. Taujale, N. Gravel, A. Venkat, W. Yeung, D.P. Byrne, P.A. Eyers, N. Kannan, KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases, BMC Bioinformatics. 22 (2021) 446. https://doi.org/10.1186/s12859-021-04358-3.

[9] A. Kwon, S. Scott, R. Taujale, W. Yeung, K.J. Kochut, P.A. Eyers, N. Kannan, Tracing the origin and evolution of pseudokinases across the tree of life, Sci. Signal. 12 (2019). https://doi.org/10.1126/scisignal.aav3810.

[10] T.R. Baffi, G. Lordén, J.M. Wozniak, A. Feichtner, W. Yeung, A.P. Kornev, C.C. King, J.C. Del Rio, A.J. Limaye, J. Bogomolovas, C.M. Gould, J. Chen, E.J. Kennedy, N. Kannan, D.J. Gonzalez, E. Stefan, S.S. Taylor, A.C. Newton, mTORC2 controls the activity of PKC and Akt by phosphorylating a conserved TOR interaction motif, Science Signaling. 14 (2021) eabe4509. https://doi.org/10.1126/scisignal.abe4509.

[11] N. Amatya, T.E. Wales, A. Kwon, W. Yeung, R.E. Joseph, D.B. Fulton, N. Kannan, J.R. Engen, A.H. Andreotti, Lipid-targeting pleckstrin homology domain turns its autoinhibitory face toward the TEC kinases, PNAS. 116 (2019) 21539–21544. https://doi.org/10.1073/pnas.1907566116.

[12] Y.-Y. Jiang, W. Maier, U.N. Chukka, M. Choromanski, C. Lee, E. Joachimiak, D. Wloga, W. Yeung, N. Kannan, J. Frankel, J. Gaertig, Mutual antagonism between Hippo signaling and cyclin E drives intracellular pattern formation, Journal of Cell Biology. 219 (2020) e202002077. https://doi.org/10.1083/jcb.202002077.

[13] R. Taujale, A. Venkat, L.-C. Huang, Z. Zhou, W. Yeung, K.M. Rasheed, S. Li, A.S. Edison, K.W. Moremen, N. Kannan, Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases, ELife. 9 (2020) e54532. https://doi.org/10.7554/eLife.54532.

[14] R. Taujale, S. Soleymani, A. Priyadarshi, A. Venkat, W. Yeung, K.J. Kochut, N. Kannan, GTXplorer: A portal to navigate and visualize the evolutionary information encoded in fold A glycosyltransferases, Glycobiology. 31 (2021) 1472–1477. https://doi.org/10.1093/glycob/cwab082.

[15] R. Taujale, Z. Zhou, W. Yeung, K.W. Moremen, S. Li, N. Kannan, Mapping the glycosyltransferase fold landscape using interpretable deep learning, Nat Commun. 12 (2021) 5656. https://doi.org/10.1038/s41467-021-25975-9.

[16] L.-C. Huang, W. Yeung, Y. Wang, H. Cheng, A. Venkat, S. Li, P. Ma, K. Rasheed, N. Kannan, Quantitative Structure–Mutation–Activity Relationship Tests (QSMART) model for protein kinase inhibitor response prediction, BMC Bioinformatics. 21 (2020) 520. https://doi.org/10.1186/s12859-020-03842-6.

[17] K.A. Davies, C. Fitzgibbon, S.N. Young, S.E. Garnish, W. Yeung, D. Coursier, R.W. Birkinshaw, J.J. Sandow, W.I.L. Lehmann, L.-Y. Liang, I.S. Lucet, J.D. Chalmers, W.M. Patrick, N. Kannan, E.J. Petrie, P.E. Czabotar, J.M. Murphy, Distinct pseudokinase domain conformations underlie divergent activation mechanisms among vertebrate MLKL orthologues, Nat Commun. 11 (2020) 3060. https://doi.org/10.1038/s41467-020-16823-3.

[18] S.E. Garnish, Y. Meng, A. Koide, J.J. Sandow, E. Denbaum, A.V. Jacobsen, W. Yeung, A.L. Samson, C.R. Horne, C. Fitzgibbon, S.N. Young, P.P.C. Smith, A.I. Webb, E.J. Petrie, J.M. Hildebrand, N. Kannan, P.E. Czabotar, S. Koide, J.M. Murphy, Conformational interconversion of MLKL and disengagement from RIPK3 precede cell death by necroptosis, Nat Commun. 12 (2021) 2211. https://doi.org/10.1038/s41467-021-22400-z.

[19] D.M. Foulkes, D.P. Byrne, W. Yeung, S. Shrestha, F.P. Bailey, S. Ferries, C.E. Eyers, K. Keeshan, C. Wells, D.H. Drewry, W.J. Zuercher, N. Kannan, P.A. Eyers, Covalent inhibitors of EGFR family protein kinases induce degradation of human Tribbles 2 (TRIB2) pseudokinase in cancer cells, Science Signaling. 11 (2018) eaat7951. https://doi.org/10.1126/scisignal.aat7951.

[20] C. Agnew, L. Liu, S. Liu, W. Xu, L. You, W. Yeung, N. Kannan, D. Jablons, N. Jura, The crystal structure of the protein kinase HIPK2 reveals a unique architecture of its CMGC-insert region, Journal of Biological Chemistry. 294 (2019) 13545–13559. https://doi.org/10.1074/jbc.RA119.009725.

[21] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[22] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, J. Wang, Q. Cong, L.N. Kinch, R.D. Schaeffer, C. Millán, H. Park, C. Adams, C.R. Glassman, A. DeGiovanni, J.H. Pereira, A.V. Rodrigues, A.A. van Dijk, A.C. Ebrecht, D.J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M.K. Rathinaswamy, U. Dalwadi, C.K. Yip, J.E. Burke, K.C. Garcia, N.V. Grishin, P.D. Adams, R.J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network, Science. 373 (2021) 871–876. https://doi.org/10.1126/science.abj8754.

[23] G. Manning, G.D. Plowman, T. Hunter, S. Sudarsanam, Evolution of protein kinase signaling from yeast to man, Trends in Biochemical Sciences. 27 (2002) 514–520. https://doi.org/10.1016/S0968-0004(02)02179-5.

[24] K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G.J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S.A.A. Kohl, A. Potapenko, A.J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A.W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome, Nature. 596 (2021) 590–596. https://doi.org/10.1038/s41586-021-03828-1.

[25] A. Ditsiou, C. Cilibrasi, N. Simigdala, A. Papakyriakou, L. Milton-Harris, V. Vella, J.E. Nettleship, J.H. Lo, S. Soni, G. Smbatyan, P. Ntavelou, T. Gagliano, M.C. Iachini, S. Khurshid, T. Simon, L. Zhou, S. Hassell-Hart, P. Carter, L.H. Pearl, R.L. Owen, R.J. Owens, S.M. Roe, N.E. Chayen, H.-J. Lenz, J. Spencer, C. Prodromou, A. Klinakis, J. Stebbing, G. Giamas, The structure-function relationship of oncogenic LMTK3, Science Advances. 6 (2020) eabc3099. https://doi.org/10.1126/sciadv.abc3099.

[26] A. Sreelatha, S.S. Yee, V.A. Lopez, B.C. Park, L.N. Kinch, S. Pilch, K.A. Servage, J. Zhang, J. Jiou, M. Karasiewicz-Urbańska, M. Łobocka, N.V. Grishin, K. Orth, R. Kucharczyk, K. Pawłowski, D.R. Tomchick, V.S. Tagliabracci, Protein AMPylation by an Evolutionarily Conserved Pseudokinase, Cell. 175 (2018) 809-821.e19. https://doi.org/10.1016/j.cell.2018.08.046.

[27] M. Gradowski, B. Baranowski, K. Pawłowski, The expanding world of protein kinase-like families in bacteria: forty families and counting, Biochemical Society Transactions. 48 (2020) 1337–1352. https://doi.org/10.1042/BST20190712.

[28] N. Kannan, S.S. Taylor, Y. Zhai, J.C. Venter, G. Manning, Structural and Functional Diversity of the Microbial Kinome, PLOS Biology. 5 (2007) e17. https://doi.org/10.1371/journal.pbio.0050017.

[29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019: pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[30] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, PNAS. 118 (2021). https://doi.org/10.1073/pnas.2016239118.

[31] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, B. Rost, ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing, IEEE Transactions on Pattern Analysis and Machine Intelligence. (2021) 1–1. https://doi.org/10.1109/TPAMI.2021.3095381.

[32] Y. Ji, Z. Zhou, H. Liu, R.V. Davuluri, DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, Bioinformatics. 37 (2021) 2112–2120. https://doi.org/10.1093/bioinformatics/btab083.

[33] R.M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, MSA Transformer, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021: pp. 8844–8856. https://proceedings.mlr.press/v139/rao21a.html (accessed March 4, 2022).

[34] L. Huo, H. Zhang, X. Huo, Y. Yang, X. Li, Y. Yin, pHMM-tree: phylogeny of profile hidden Markov models, Bioinformatics. 33 (2017) 1093–1095. https://doi.org/10.1093/bioinformatics/btw779.

[35] S. Salvador, P. Chan, Toward accurate dynamic time warping in linear time and space, Intell. Data Anal. 11 (2007) 561–580.

[36] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Research. 30 (2002) 3059–3066. https://doi.org/10.1093/nar/gkf436.

[37] P. Pamilo, M. Nei, Relationships between gene trees and species trees., Molecular Biology and Evolution. 5 (1988) 568–583. https://doi.org/10.1093/oxfordjournals.molbev.a040517.

[38] J.H. Degnan, M. DeGiorgio, D. Bryant, N.A. Rosenberg, Properties of Consensus Methods for Inferring Species Trees from Gene Trees, Systematic Biology. 58 (2009) 35–54. https://doi.org/10.1093/sysbio/syp008.

[39] H.J. Atkinson, J.H. Morris, T.E. Ferrin, P.C. Babbitt, Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies, PLOS ONE. 4 (2009) e4345. https://doi.org/10.1371/journal.pone.0004345.

[40] H.J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies., Molecular Biology and Evolution. 16 (1999) 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036.

[41] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., Molecular Biology and Evolution. 4 (1987) 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

[42] J. Felsenstein, Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters, Systematic Biology. 22 (1973) 240–249. https://doi.org/10.1093/sysbio/22.3.240.

[43] Z. Yang, S. Kumar, M. Nei, A new method of inference of ancestral nucleotide and amino acid sequences, Genetics. 141 (1995) 1641–1650. https://doi.org/10.1093/genetics/141.4.1641.

[44] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM. 60 (2017) 84–90. https://doi.org/10.1145/3065386.

[45] T. Gao, X. Yao, D. Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021: pp. 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552.

[46] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019: pp. 3982–3992. https://doi.org/10.18653/v1/D19-1410.

[47] D.A. Knowles, Z. Ghahramani, Pitman-Yor Diffusion Trees, UAI. (2011) 9.

[48] K. Takahashi, M. Nei, Efficiencies of Fast Algorithms of Phylogenetic Inference Under the Criteria of Maximum Parsimony, Minimum Evolution, and Maximum Likelihood When a Large Number of Sequences Are Used, Molecular Biology and Evolution. 17 (2000) 1251–1258. https://doi.org/10.1093/oxfordjournals.molbev.a026408.

[49] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, ArXiv:1312.6114 [Cs, Stat]. (2014). http://arxiv.org/abs/1312.6114 (accessed February 1, 2022).

[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2014. https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (accessed March 5, 2022).

[51] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep Unsupervised Learning using Nonequilibrium Thermodynamics, in: Proceedings of the 32nd International Conference on Machine Learning, PMLR, 2015: pp. 2256–2265. https://proceedings.mlr.press/v37/sohl-dickstein15.html (accessed March 5, 2022).

[52] P. Dhariwal, A. Nichol, Diffusion Models Beat GANs on Image Synthesis, ArXiv:2105.05233 [Cs, Stat]. (2021). http://arxiv.org/abs/2105.05233 (accessed March 5, 2022).

APPENDIX A

ARRAY-BASED COMPARATIVE SEQUENCE ANALYSIS

Introduction

This appendix explains the rationale and advantages of a NumPy array-based framework for comparative sequence analysis. Many of the ideas described here are implemented in the HelperBunny library, a collection of object-oriented code for the low-level navigation and analysis of multiple sequence alignments. The library is freely available at:

https://github.com/waylandy/HelperBunny

For example applications, the HelperBunny library was used extensively for analyzing a large alignment of tyrosine kinase sequences across holozoan organisms [1]. See the associated computational notebook (analysis sequence.ipynb) in the following repository:

https://github.com/esbgkannan/holozoan_tk_evolution

Sequence alignment formats

There are many file formats for storing multiple sequence alignments, each with unique advantages and disadvantages. Among these, I found A2M to be one of the most versatile formats, capable of representing both aligned and unaligned positions while also showing the full sequence context. Furthermore, the A2M format obeys all FASTA formatting conventions and can be parsed as such. According to A2M conventions, a sequence header line is designated by an initial greater-than sign, while the corresponding aligned sequence is placed on the next line without line breaks. On the sequence line, aligned residue positions are written in uppercase or dashes, while unaligned residues are written in lowercase. To provide the full sequence context, unaligned N or C-terminal regions may be written as lowercase strings which flank the aligned segment. Periods may also be added to move aligned columns to equivalent positions, but this is not advisable because it greatly inflates the file size.

Functionally, the A2M format provides a space-efficient strategy for storing both an alignment and its full-length sequence context within the same file. The alignment can be extracted by only parsing dashes and uppercase characters, while the original sequence can be extracted by not parsing the dashes. This format is particularly useful for mapping alignment positions to sequence positions, which is a common task for many workflows. If the full-length sequence is required for downstream analyses, an A2M alignment would potentially remove the need to map between two separate files containing each the sequence alignment and the full-length sequence. An example application would be determining if the conservation of a sequence motif within the aligned region is correlated with the presence of a specific protein domain outside of the aligned region.

Implementing an A2M alignment array object

Alignments can be represented as a two-dimensional array where the horizontal axis represents each sequence while the vertical axis represents each aligned position. This framework can be expanded to accommodate A2M alignments by redefining the vertical axis to contain alternating insertions and positions. Each element of a position column is one exactly character long, either a dash or a single uppercase character. Each element of an insertion column is a variable number of characters long, containing all lowercase characters in between the two flanking alignment positions. Under this framework, odd columns will be insertions, while even columns will be positions. Alternatively, empty insertion columns may be deleted from the array; however, this would require the index of insertion and position columns to be dynamically determined.

The alignment array framework can be implemented in Python using NumPy array objects [2,3]. Furthermore, the datatype should be specified as object, not string. Each element of a NumPy array must be the same size. The element size is determined by the largest element in the array. While position elements will always be one character, insertion elements can be a variable number of characters long. The presence of a single long insertion can lead to an enormous array which cannot be stored in a reasonable amount of memory. By specifying the datatype as an object, each element of the array will instead be a fixed-size reference to a string stored elsewhere in memory.

In addition to the alignment, the sequence headers should also be stored as a one-dimensional NumPy array with the datatype specified as an object. The size of the sequence header array should match the size of the vertical axis in the alignment array.

Array indexing

One of the major advantages of the NumPy array object is the advanced indexing capabilities. NumPy arrays can be indexed along a given dimension using either Boolean arrays or (lists of) integers. If using a Boolean array, the size of the array must correspond with the size of the dimension being queried. These indexing routines can be used to retrieve a subset of alignment positions/insertions, retrieve a subset of sequences, or both at the same time. For instance, one could index the columns of the alignment array to remove all insertion positions. One could also index the rows of the alignment array using a function of the sequence header array.

Array operations

NumPy arrays also implement vectorized Python operators. The most pertinent operators for alignment arrays are “equals”, “not equals”, and “not”. To provide an example, if we evaluate if an alignment array is equal to the dash character, this will return a Boolean array of the same size indicating all positions of the original array that contain a gap. Utilizing built-in NumPy array operations, one can take the mean across the horizontal axis, which will yield the percent gap at each position of the alignment. In NumPy, a Boolean array can also be inverted using a tilde which represents “not”.

The alignment array framework facilitates many complex analyses through the creative usage of array operations and indexing. For instance, one may index the columns of the array of retrieve alignment position, then utilize a series of vectorized array operations to calculate the amino acid frequencies at each position. By calculating the Shannon entropy of each position, we quantify the conservation of each alignment position. To provide another example, one may

utilize array operations to calculate the percentage gaps at each position, the percentage insertions between each position, and the median length of non-zero size insertions between each position. A comparative plot of these three variables shows the prevalence of insertions and deletions throughout the alignment.

Advanced querying with Boolean algebra

Complex queries for indexing alignment arrays can be created using Boolean algebra. As previously mentioned, Boolean arrays can be used for indexing. Expanding upon this idea, new Boolean arrays can be defined as a function of other Boolean arrays, which can be used to describe a more complex set of conditions. For example, if I had two Boolean arrays, one which indicates whether position 10 is glycine and another which indicates whether position 20 is glycine, the product of these two arrays would yield a new Boolean array which indicates sequences which conserve a glycine at both positions 10 and position 20, while the summation of these two arrays would yield a new Boolean array which indicates sequences which conserve a glycine at either position 10 or 20 or both. Boolean addition is synonymous with “or”, while Boolean multiplication is synonymous with “and”. For more complex queries, these functions may also be used in combination with the “not” operator.

The ability to quickly build complex queries through Boolean algebra represents a major advantage of array-based sequence analysis. To take advantage of these capabilities, sequence features should be represented as separate arrays which correspond to either axis of the alignment array. This maintains mutual compatibility between all arrays, as these operations cannot be performed on arrays of different sizes.

Integration with other methods/data

Many existing tools can be adapted into the array-based framework for downstream analyses. Specific strategies for implementation will be discussed in the following subsections.

Sequence filtering

CD-HIT is a popular program for filtering and clustering large protein datasets [4]. The standard method for using CD-HIT is to provide an input sequence dataset and a desired percent sequence similarity cutoff for filtering. The program will output a smaller dataset containing a subset of the original input, which can be used for downstream analyses. To adapt this towards an array-based framework, the filtered subset can be represented as a Boolean array relative to the original sequence dataset. To implement this, I define a function that takes an alignment array and a sequence similarity cutoff for filtering. From the alignment array, I write a temporary FASTA file where the headers are replaced with the index of each sequence. Running CD-HIT on this temporary file will yield a filtered file that contains the indices of sequences that made it past the filter. This list of indices can be converted to a Boolean array by initializing an empty array with the same size as the number of sequences then setting the corresponding indices to True. The function finishes by deleting the temporary files then returning the Boolean array. Indexing the alignment array with this output will yield the filtered set.

For more complex analyses, it is possible to generate multiple Boolean arrays corresponding to different filtering cutoffs. Evaluating the sum of each array will yield the number of sequences that made it past the filter while evaluating the mean of each array will yield the percentage of sequences that made it past the filter. This is a space-efficient solution

that provides an easy way to manage and test the effects of applying different cutoffs throughout sequence analysis.

Protein domain analysis

The conserved domain database (CDD) is a popular tool for identifying protein domains within full-length protein sequences [5]. A FASTA containing unaligned full-length protein sequences can be extracted from the A2M alignment then passed into the CDD server. Results will return as a tab-separated file where each line describes a protein domain that was detected in a given sequence. Parsing this file, it is possible to generate a list of headers that contain each unique domain. These can be converted into Boolean arrays by cross-referencing the sequence header array. TMHMM is another popular tool that identifies transmembrane helices within full-length protein sequences [6]. This also returns a delimited file where each line predicts a transmembrane helix on a given sequence which can be converted into a Boolean array in a similar fashion as CDD. To provide an example application, one may utilize a series of array operations to derive the frequency of each domain or combination of domains within the dataset.

Taxonomic analysis

For large sequence databases such as nr [7] and UniProt [8,9], sequence headers will include the taxon (species) name or taxon number of each sequence source. These identifiers can be used to retrieve the full taxonomic classification for each organism from NCBI taxdump [10]. Upon downloading taxdump, a list of parent-child node relationships will be listed by taxon number under “nodes.dmp”, while a list of each taxon number and corresponding taxon name and rank is provided under “names.dmp”. This information can be efficiently stored as a Huffman tree [11] using NetworkX for fast querying [12]. Given an array of species, one could

use this framework to generate a Boolean array of whether each species is classified under a given taxon such as “Chordata”. From an array of species, one could also query a taxonomic rank, such as phylum, in order to generate a new array that would list the classification of each species at that particular rank. To provide an example application, it would be possible to derive a taxa conservation table for a given set of sequences which may be further stratified by other variables such as protein family or the presence/absence of motifs.

Direct coupling analysis

Implemented by CCMpred [13], direct coupling analysis quantifies second-order relationships between columns of a sequence alignment based on the Potts model. Although there are many potential applications for this model, it is typically used to predict pairwise residue contacts [14]. To reframe this into a NumPy array-based framework, I define an object that, given an alignment array, will run CCMpred and store the model parameters for downstream analyses. From the alignment array, I write a temporary PSICOV file that only contains the alignment positions of each sequence with the headers. Running CCMpred on the temporary PSICOV file with the “-r” option will export the full Potts model in raw text, which can be parsed into the fields and coupling matrices. These parameters are stored as attributes of the object for easy access in downstream analyses. Finally, the temporary files can be deleted. Storing the model parameters as NumPy arrays provides an easy way to interface with the alignment array as well as perform array operations for downstream analyses.

The two parameters of the model are the fields, a two-dimensional array where the first axis corresponds to each alignment position while the second axis corresponds to each amino acid, and the couplings, a four-dimensional array where the first two axes correspond to each

pair of alignment positions while the last two axes correspond to each possible pair of amino acid. Pairwise residue contacts can be predicted by calculating the Frobenius norm of the last two axes of the coupling matrix [14], then applying average product correction [15]. As another application, it is also possible to calculate the statistical free energy of a given protein sequence by summing the associated values for each residue and residue pair, provided by the field and coupling matrices, respectively [16]. Using this framework, multiple Potts models may also be generated, calculated from different subsets of the alignment, which may yield different predicted contacts or statistical energies.

Sequence constraint analysis

Bayesian Partitioning with Pattern Selection (BPPS) is a program and hierarchical clustering algorithm for sequence alignments that identifies groups of sequences that share a conserved set of sequence motifs, where each motif is associated with a weight which quantifies the degree to which the motif is specific to the cluster [17]. Although BPPS provides many outputs, virtually all clustering results can be derived from the LPR file, which lists each sequence cluster alongside their corresponding weighted motifs. The HPT file, which explicitly lists the hierarchical clustering scheme, may also be useful; however, this information can typically be inferred from the LPR file. I adopt this into the alignment array-based framework by defining an object which, given an LPR file, will parse and store each cluster and their weighted motifs. The degree to which a given aligned sequence fits into a cluster can be calculated by the weighted score of all cluster-specific motifs which are conserved by the sequence divided by the total weighted score of all motifs for that cluster. The resulting LPR object may be used to score a corresponding alignment — for each sequence in the alignment, calculate the degree to which each fit into each cluster. This results in a two-dimensional array with axes corresponding to the

number of sequences in the alignment and the number of clusters defined by the LPR object. An optimal cutoff to be classified into a cluster is a score of 0.7 or higher.

Scientific Python stack

Built upon the NumPy arrays, the scientific Python stack provides a broad set of tools for downstream tasks such as machine learning, data analysis, and plotting [18]. The establishment of a NumPy array-based framework for representing multiple sequence alignments facilitates seamless integration into the greater Python ecosystem. Built to accommodate a wide range of fields, the adoption of this framework will support interdisciplinary research efforts and promote new methods and discoveries.

Bibliography

[1] W. Yeung, A. Kwon, R. Taujale, C. Bunn, A. Venkat, N. Kannan, Evolution of Functional Diversity in the Holozoan Tyrosine Kinome, Molecular Biology and Evolution. 38 (2021) 5625–5639. https://doi.org/10.1093/molbev/msab272.

[2] S. van der Walt, S.C. Colbert, G. Varoquaux, The NumPy Array: A Structure for Efficient Numerical Computation, Computing in Science Engineering. 13 (2011) 22–30. https://doi.org/10.1109/MCSE.2011.37.

[3] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature. 585 (2020) 357–362. https://doi.org/10.1038/s41586-020-2649-2.

[4] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics. 28 (2012) 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

[5] A. Marchler-Bauer, S. Lu, J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, J.D. Jackson, Z. Ke, C.J. Lanczycki, F. Lu, G.H. Marchler, M. Mullokandov, M.V. Omelchenko, C.L. Robertson, J.S. Song, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S.H. Bryant, CDD: a Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Research. 39 (2011) D225–D229. https://doi.org/10.1093/nar/gkq1189.

[6] A. Krogh, B. Larsson, G. von Heijne, E.L.L. Sonnhammer, Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, Journal of Molecular Biology. 305 (2001) 567–580. https://doi.org/10.1006/jmbi.2000.4315.

[7] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Research. 35 (2007) D61–D65. https://doi.org/10.1093/nar/gkl842.

[8] UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204-212. https://doi.org/10.1093/nar/gku989.

[9] The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research. 49 (2021) D480–D489. https://doi.org/10.1093/nar/gkaa1100.

[10] C.L. Schoch, S. Ciufo, M. Domrachev, C.L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J.P. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, NCBI Taxonomy: a comprehensive update on

curation, resources and tools, Database. 2020 (2020) baaa062. https://doi.org/10.1093/database/baaa062.

[11] D.A. Huffman, A Method for the Construction of Minimum-Redundancy Codes, Proceedings of the IRE. 40 (1952) 1098–1101. https://doi.org/10.1109/JRPROC.1952.273898.

[12] A. Hagberg, P. Swart, D. Chult, Exploring Network Structure, Dynamics, and Function Using NetworkX, in: 2008.

[13] S. Seemayer, M. Gruber, J. Söding, CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations, Bioinformatics. 30 (2014) 3128–3130. https://doi.org/10.1093/bioinformatics/btu500.

[14] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, Phys. Rev. E. 87 (2013) 012707. https://doi.org/10.1103/PhysRevE.87.012707.

[15] S.D. Dunn, L.M. Wahl, G.B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, Bioinformatics. 24 (2008) 333–340. https://doi.org/10.1093/bioinformatics/btm604.

[16] R.M. Levy, A. Haldane, W.F. Flynn, Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness, Current Opinion in Structural Biology. 43 (2017) 55–62. https://doi.org/10.1016/j.sbi.2016.11.004.

[17] A.F. Neuwald, A Bayesian Sampler for Optimization of Protein Domain Hierarchies, Journal of Computational Biology. 21 (2014) 269–286. https://doi.org/10.1089/cmb.2013.0099.

[18] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods. 17 (2020) 261–272. https://doi.org/10.1038/s41592-019-0686-2.

APPENDIX B

INVESTIGATING EVOLUTIONARY CONSTRAINTS IN THE TYROSINE KINOME

Introduction

We previously defined a sequence constraint-guided phylogeny of the holozoan tyrosine kinome which depicts the gain or loss of slow-evolving sites, indicative of functionalization events in evolutionary history [1]. We characterized the evolution of 48 distinct clusters of tyrosine kinases, each defined by a shared set of slow-evolving sites (i.e. evolutionary constraints). These sites constrain the evolution of distinct families and are often associated with protein structure-function [2]. The study of family-specific evolutionary constraints provides a valuable way to identify family-specific modes of kinase structure function.

To encourage further studies and foster new hypotheses, we created an online resource using the KinView framework [3]. This allows users to browse the unique sequence constraints

of each tyrosine family/subgroup through an intuitive, user-friendly interface. A temporary website for this resource is available at:

https://uga-tyro-96ffcedb509f.netlify.app

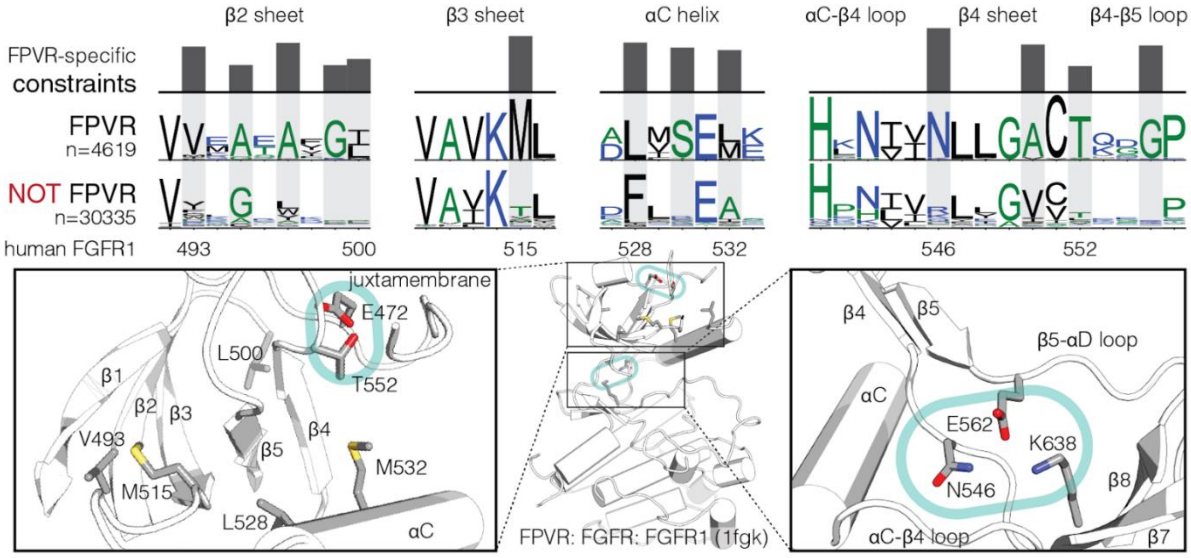
Here, we demonstrate how family-specific constraints facilitate the research of novel tyrosine kinase structure function by examining three major tyrosine kinase subgroups which comprise nearly half of the tyrosine kinome: FPVR, IRKL, and SrcM. Searching the literature, we note that strong FPVR-specific and IRKL-specific constraints have been experimentally characterized to some extent in the past; however, to the best of our knowledge, strong SrcM-specific evolutionary constraints lack any experimental characterization.

Evolutionary constraints describe three distinct subgroup-specific modes of kinase regulation

We analyze the structural contexts of evolutionary constraints specific to each of the three major tyrosine kinase subgroups: FPVR, IRKL, and SrcM. Based on previously published works, we find evidence of direct experimental characterization for a small handful of important regulatory residues which we have previously determined to be evolutionary constraints. Several evolutionary constraints residues have also been indirectly characterized by large scale proteomics studies. Overall, the majority of evolutionary constraints lack any characterization whatsoever which leaves much room for future exploration.

FPVR-specific constraints describe the molecular brake and juxtamembrane docking site

The single strongest FPVR-specific constraint is an asparagine in the αC-β4 loop ($FGFR1^{N546}$). Previous studies have shown that this αC-β4 loop asparagine is a member of the molecular brake triad which stabilizes an autoinhibited kinase conformation [4]. Our data does not define the remaining two residues of the molecular brake triad as FPVR-specific constraints because they are conserved throughout all tyrosine kinases. Our database shows that the molecular brake asparagine is conserved by the all FPVR families (except for Ret) as well as the closely-related Tie family, which forms a sister clade to the FPVR kinases [1].



Evolutionary constraints of the FPVR subgroup

Comparative sequence logos and structural mappings of subgroup-specific motifs for the FPVR subgroup of tyrosine kinases. Sequence logos for the strongest evolutionary constraints are shown on top, with comparative sequence logos for sequences outside each subgroup provided below. Evolutionary constraints are highlighted in gray, with the height of the histogram reflecting the degree of divergence at that position between the subgroup and sequences outside the subgroup. Notable regulatory interactions formed by strong evolutionary constraints are circled in turquoise on a representative structure of FPVR (FGFR1) [5].

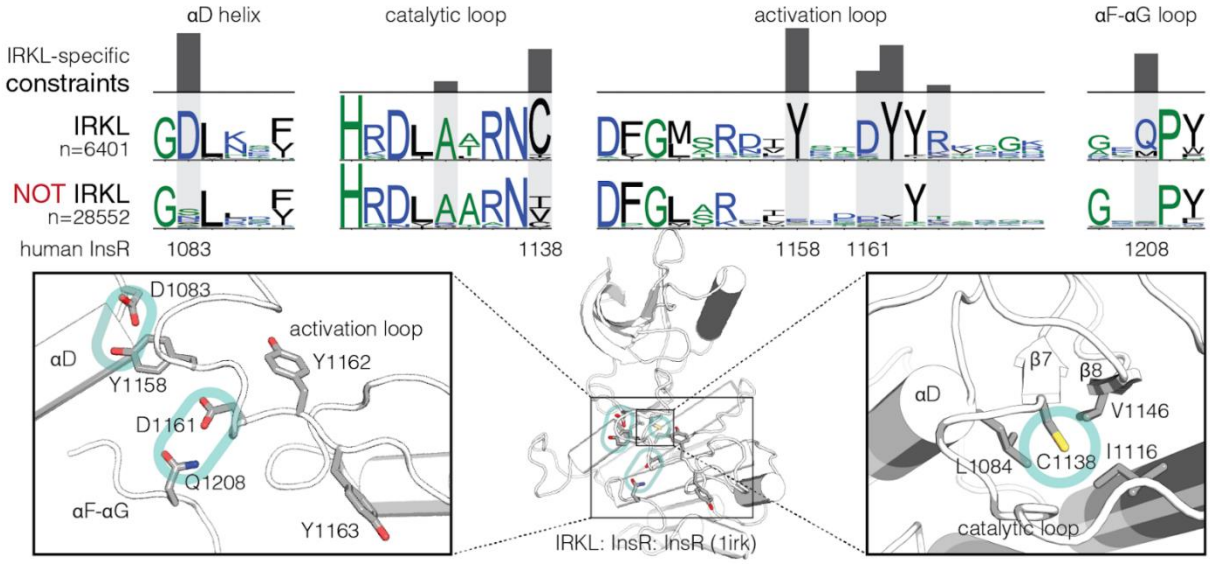
The majority of strong FPVR-specific constraints form a large hydrophobic patch in the N-lobe of the kinase domain, coinciding with part of the juxtamembrane docking site. Located between the kinase domain and transmembrane helix, the juxtamembrane segment is an important regulatory component of the receptor tyrosine kinases. For the majority of receptor tyrosine kinases, the juxtamembrane segment stabilizes unique autoinhibitory interactions which vary depending on the family [6]. In addition to these hydrophobic residues, we also identify a conserved hydrogen bond formed between a FPVR-specific, β4-β5 loop threonine ($FGFR1^{T552}$) and the juxtamembrane WEx motif. We hypothesize that these FPVR-specific constraints describe an FPVR-specific mode of juxtamembrane docking.

IRKL-specific constraints describe the IRK-like inactive conformation and a redox regulatory site

The strongest IRKL-specific constraint is an activation loop tyrosine ($InsR^{Y1158}$). This residue is part of the phosphorylatable YxxDYY activation loop motif which has been previously shown to regulate catalytic activity in insulin receptor family kinases [7]. The YxxDYY motif is conserved in all IRKL families, suggesting that specific phosphorylation of this motif is a conserved regulatory feature throughout the IRKL subgroup [8].

Analyzing a representative crystal structure, we observe that the strongest IRKL-specific constraints stabilize a conserved autoinhibited activation loop conformation which occludes the ATP-binding pocket [7]. We refer to this as the IRK-like inactive conformation. The IRK-like inactive conformation is stabilized by four IRKL-specific constraint residues which forms two pairs of conserved hydrogen bonds. The first pair is $InsR^{Y1158}$ (activation loop, YxxDYY-Tyr1) and $InsR^{D1083}$ (αD helix). The second pair is $InsR^{D1161}$ (activation loop, YxxDYY-Asp) and

InsR^{Q1208} (αF-αG loop). The first pair is conserved across IRKL, while the second pair is conserved in all IRKL families except ALK and CCK4. Phylogenetic inference suggests that divergence of the ALK and CCK4 families are not related, occurring as two independent events [1]. Possibly a result of this divergence, structures of autoinhibited ALK are noticeably distinct from the IRK-like inactive conformation [9]. While CCK4 has not yet been crystallized, we predict that CCK4 may still be capable of adopting the IRK-like inactive conformation through alternative interactions. CCK4 is a catalytically inactive pseudokinase which indicates that it may be under different selective pressures than other IRKL kinases.



Evolutionary constraints of the IRKL subgroup

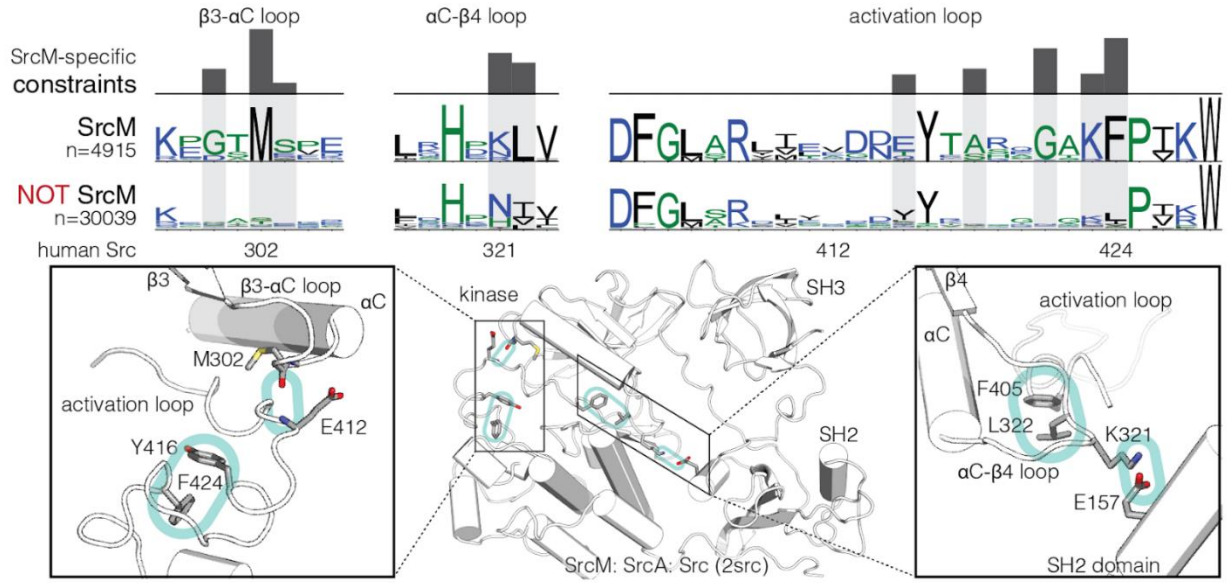
Comparative sequence logos and structural mappings of subgroup-specific motifs for the IRKL subgroup of tyrosine kinases. Sequence logos for the strongest evolutionary constraints are shown on top, with comparative sequence logos for sequences outside each subgroup provided below. Evolutionary constraints are highlighted in gray, with the height of the histogram reflecting the degree of divergence at that position between the subgroup and sequences outside the subgroup. Notable regulatory interactions formed by strong evolutionary constraints are circled in turquoise on a representative structure of IRKL (IRK) [10].

Another strong IRKL-specific constraint is a catalytic loop cysteine ($InsR^{C1138}$) buried behind the activation loop in the hydrophobic core. In a previous study, an alanine scan of all cysteine residues in human InsR kinase reveals that $InsR^{C1138A}$ exhibits significantly decreased autophosphorylation of the YxxDYY motif under oxidative stress [11]. The activity of other cysteine mutants was comparable to wild-type. The oxidized cysteine residue likely destabilizes the autoinhibited conformation by disrupting hydrophobic packing interactions. Furthermore, the Oximouse database [12] reports sulfenylation at the equivalent catalytic loop cysteine in six IRKL kinases across four IRKL families: CCK4 ($PTK7^{C928}$), InsR ($INSR^{C1155}$), Lmr ($LMTK2^{C270}$, $LMTK3^{C272}$), and Trk ($NTRK2^{C681}$, $NTRK3^{C685}$). Our data suggests that redox regulation via the catalytic loop cysteine may be a conserved regulatory mechanism throughout IRKL kinases, except for the Ror family which does not conserve the catalytic loop cysteine.

SrcM-specific constraints describe the Src-like inactive conformation and an SH2 binding site

The strongest SrcM-specific constraint is a methionine (Src^{M302}) in the β3-αC loop. To the best of our knowledge, the role of the SrcM-specific methionine has not been characterized in previous studies, nor is its role in SrcM regulation immediately obvious. A representative crystal structure reveals that a cluster of strong SrcM constraints localize near the Src-like inactive conformation: an autoinhibited activation loop conformation shared by many Src-related kinases [13,14]. We hypothesize that the SrcM-specific methionine may play a role in stabilizing the β3-αC loop to form an isolated β-bridge with the activation loop. We obverse an equivalent β-bridge in SrcA (2src) [13], SrcB (1qcf) [15], and Tec (4y93) [16], while Abl (2g1t) [17] exhibits similar β3-αC backbone interaction against activation loop side chains. This conserved interaction stabilizes the activation loop to form a 3^{10} helix which abuts the αC helix. Downstream, a cluster

of SrcM-specific constraints forms a SrcM-specific GxKF motif in the activation loop. Further stabilizing the Src-like inactive conformation, the GxKF-Phe (Src^{F424}) forms a T-shaped pi stacking interaction with a phosphorylatable activation loop tyrosine. We observe equivalent T-shaped pi stacking interactions in SrcA (2src) [13], SrcB (1qcf) [15], and Tec (4y93) [16].



Evolutionary constraints of the SrcM subgroup
Comparative sequence logos and structural mappings of subgroup-specific motifs for the SrcM subgroup of tyrosine kinases. Sequence logos for the strongest evolutionary constraints are shown on top, with comparative sequence logos for sequences outside each subgroup provided below. Evolutionary constraints are highlighted in gray, with the height of the histogram reflecting the degree of divergence at that position between the subgroup and sequences outside the subgroup. Notable regulatory interactions formed by strong evolutionary constraints are circled in turquoise on a representative structure of SrcM (Src) [13].

We also identify strong SrcM-specific constraints are also found in the αC-β4 loop. SrcM kinases diverge from other protein kinases in that αC-β4 loop in that they conserve a HxKL motif in place of the canonical HxNx motif [18]. The SrcM-specific HxKL motif forms interactions that bridge the activation loop to the SH3-SH2 binding interface. In the autoinhibited

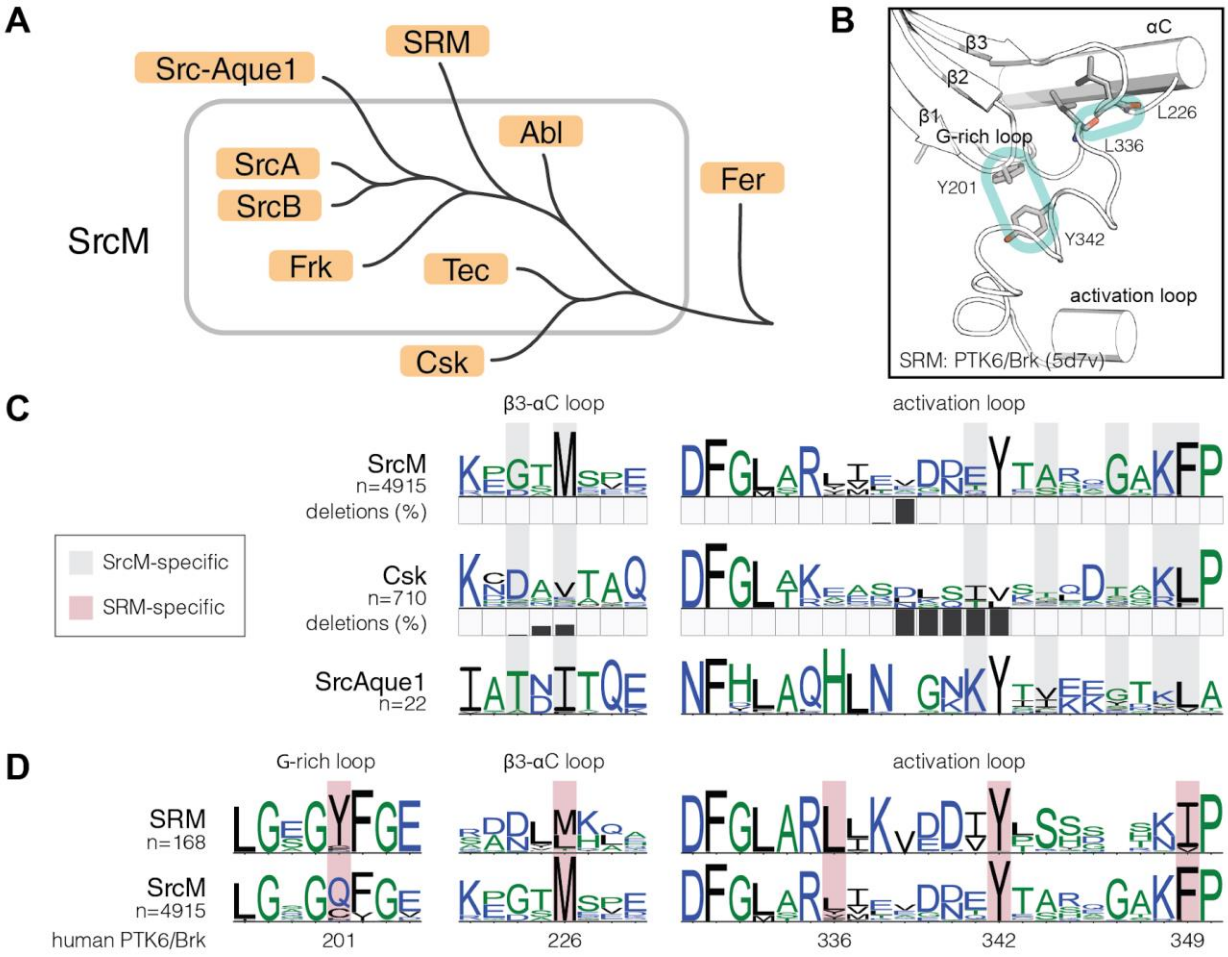
SH3-SH2-kinase domain assembly, a structure of Src depicts the HxKL-Lys (Src^{K321}) forming a salt bridge with Src^{E157} in the SH2 domain. The HxKL-Leu (Src^{L322}) forms hydrophobic contacts with the DFG-Phe in the regulatory spine: a transient structural motif whose assembly is correlated with kinase activation [19]. The HxKL motif is conserved in all SrcM families except Abl which retains a more canonical HxNL motif. This variation may be related to a mechanism of Abl-specific SH3-SH2-kinase assembly which uniquely requires N-terminal myristoylation [20].

Independent divergence of three SrcM-related families from a common SrcM ancestor

In the previous section, we observed that several tyrosine kinase families classified under one of the three major subgroups slightly diverge from subgroup-specific evolutionary constraints. For instance, Abl retains many evolutionary constraints surrounding the Src-like inactive conformation, but lacks SrcM-specific HxKL motif in the SH3-SH2-kinase interface. If this trend continues, the loss of too many subgroup-specific constraints could lead to the establishment of a new standalone family of kinases. In fact, a constraint-guided phylogeny of the tyrosine kinome indicates the divergence of three SrcM-related families from the SrcM group [1]. Likely descendents of a common SrcM ancestor, the Csk, SrcAque1, and SRM families have lost many SrcM-specific constraints throughout evolution.

The Csk family has acquired a large deletion in the kinase activation loop. This deletion likely prohibits Csk kinases from adopting the Src-like inactive conformation, thus losing the evolutionary pressure to maintain SrcM-specific constraints. Sure enough, comparative sequence logos indicate that Csk evolutionary constraints surrounding the Src-like inactive conformation such as the β3-αC methionine and the activation loop phenylalanine. We speculate that the

divergence of the Csk family represents a neofunctionalization event: Csk gained the unique ability to regulate other SrcM kinases via C-tail phosphorylate [14].



The SRM family diverges from the SrcM subgroup.
(A) We show the SrcM clade from our constraint-guided phylogeny of the tyrosine kinome [1]. The Csk, SrcAque1, and SRM families have lost SrcM-specific constraints throughout their evolution. (B) Two conserved interactions, circled in turquoise, help stabilize the Src-like inactive state in a structure of PTK6 (Brk) [21]. The upper circle shows a conserved β-bridge between the β3-αC loop and activation loop. The lower circle shows the β1-β2 loop tyrosine forming pi-stacking interactions with the activation loop tyrosine. (C) Comparative sequence logos depict differences between SrcM, Csk, and SrcAque. The grey highlight bars represent the positions of SrcM-specific constraints. Percent deletions per alignment position are shown under SrcM and Csk. (D) Comparative sequence logos depict differences between the SRM family and SrcM subgroup. The salmon-colored highlight bars represent the positions of SrcM-specific constraints. Residue numbers relative to human PTK6 (Brk) are provided below.

The SrcAque1 family [22] are putative pseudokinases which lack the DFG motif, a crucial component for catalyzing phosphotransfer [23]. The loss of catalytic activity abolishes the need to dynamically regulate an autoinhibited inactive conformation. As expected, comparative sequence logos indicate that Csk evolutionary constraints surrounding the Src-like inactive conformation such as the β3-αC methionine and the activation loop phenylalanine. Perhaps performing protein scaffolding functions, SrcAque1 may still be capable of adopting the Src-like inactive conformation. However, we predict that SrcAque1 lacks the ability (or need) to dynamically transition between multiple conformations. Divergence of the SrcAque1 family likely represents a nonfunctionalization event, losing the ability to catalyze phosphotransfer.

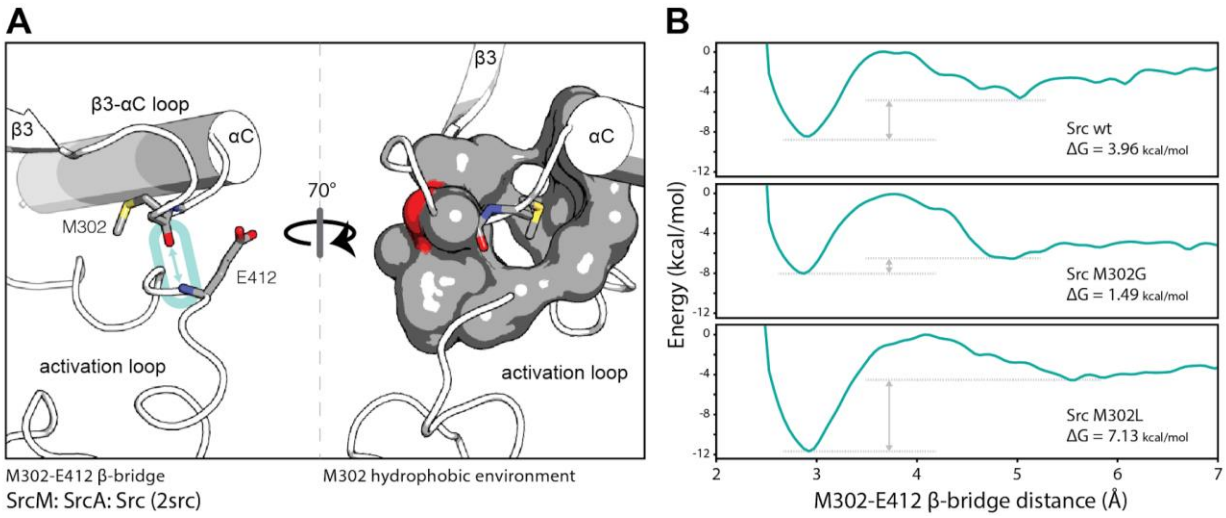
The SRM family exhibit similar structure-function to the SrcM kinases despite lacking many SrcM-specific constraints [14,21]. In fact, PTK6/Brk is capable of adopting the SrcM-specific inactive conformation despite lacking the β3-αC methionine and the activation loop phenylalanine. PTK6 conserves the isolated β-bridge between the β3-αC loop and activation loop despite lacking the β3-αC methionine. Additionally, the phosphorylatable activation loop tyrosine in PTK6 stabilizes a T-shaped pi stacking interaction with an SRM-specific tyrosine in the β1-β2 (G-rich) loop rather than the GxKF-Phe. While the SRM family may have descended from an ancestral SrcM kinase, SRM kinases lost the evolutionary pressure to maintain SrcM-specific constraints after evolving alternative means of stabilizing the Src-like inactive conformation using new SRM-specific constraints such as the G-rich loop tyrosine.

The SrcM-specific methionine provides a dynamic locking mechanism for maintaining the Src-like inactive state.

While many strong FPVR and IRKL-specific constraints have been experimentally characterized, we do not find direct experimental characterization for any strong SrcM-specific constraints. Thus, we sought to characterize the strongest SrcM-specific constraint and determine why evolution has conserved a methionine in the β3-αC loop in SrcM kinases. As previously mentioned, the SrcM-specific methionine (Src^{M302}) forms hydrophobic packing interactions within the kinase N-lobe. We believe that this packing interaction stabilizes the β3-αC backbone to form an isolated β-bridge against the activation loop which is observed in the Src-like inactive state. Outside of the SrcM subgroup, closely related families also conserve hydrophobic residues at the equivalent position, albeit conserving less entropic residues such as leucine (SRM), isoleucine (Src-Aque1), and valine (CSK). The methionine side likely confers a delicate balance of stability and instability for the β3-αC loop: stable enough to maintain the Src-like inactive state, but too stable to hinder transition to the active conformation.

We quantify the effect of residue entropy on stabilizing the Src-like inactive conformation. Potential of mean force (PMF) calculations reveal that a leucine variant (Src^{M302L}) stabilizes the Src-like inactive conformation, while a glycine variant (Src^{M302G}) de-stabilizes the Src-like inactive conformation. The PMF (also known as free energy surface) was calculated as a function of M302-E412 β-bridge distance using the weighted histogram analysis method (WHAM) from a simulated ensemble of solvated, all-atom umbrella samples (see methods for more details). In wild type Src, results estimate that the free energy of breaking the β-bridge is +3.96 kcal/mol. Compared to wild type, Src^{M302L} (ΔG = 7.13 kcal/mol) stabilizes the β-bridge by 3.17 kcal/mol, while Src^{M302G} (ΔG = 1.49 kcal/mol) destabilizes the β-bridge by 2.47 kcal/mol.

Overall results suggest that Src^{M302G} destabilizes the Src-like inactive conformation by increasing β3-αC loop entropy, while Src^{M302L} stabilizes the Src-like inactive conformation by decreasing β3-αC loop entropy by facilitating more stable hydrophobic packing interactions.

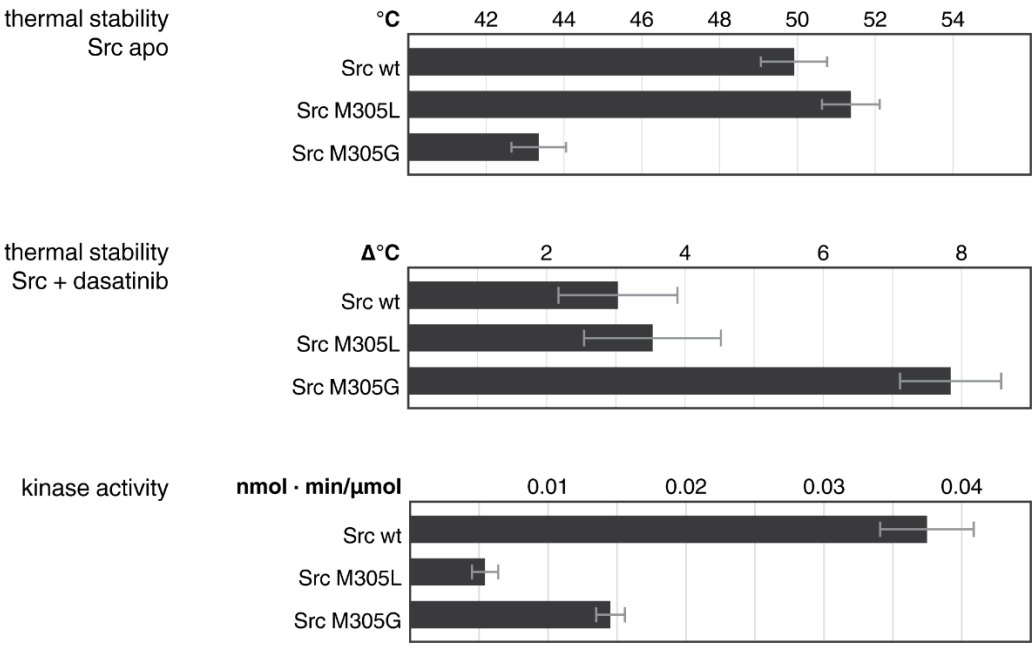


Frk stably adopts a Src-like conformation & Lmr1 stably adopts an IRK-like conformation.
(A) The SrcM-specific methionine is shown from two angles in Src kinase. The left depicted a conserved β-bridge formed between the β3-αC loop and activation loop. The right depicts the methionine side chain buried inside of a hydrophobic pocket. (B) The potential mean force of several wild type Src compared to several Src-specific methionine mutants (M302G and M302L). The free energy surface is provided as a function of M302-E412 β-bridge distance.

In vitro characterization of the SrcM-specific methionine

We further experimentally characterized the effects of Src^{M302L} and Src^{M302G} using full-length chicken Src which includes the SH3, SH2, and kinase domains. Point mutations were created using the New England BioLab Q5 Site-Directed Mutagenesis Kit and confirmed by Sanger sequencing. Protein was co-expressed with YopH tyrosine protein phosphatase in BL21 Rosetta2 DE3 cells, then purified by cobalt affinity, followed by anion exchange which yielded

the dephosphorylated protein.



Stability and activity assays for SH3-SH2-Src.
The thermal stability of Src kinase and methionine mutants was determined by differential scanning fluorimetry, while kinase activity assays were determined by ATP/NADH coupled assay.

Changes in global protein stability were quantified by differential scanning fluorimetry (DSF). In comparisons across apo Src, M302L stabilizes the protein by ~1°C compared to the wild type, while M302G destabilizes the protein by ~7°C. This appears to be in agreement with our PMF simulations in which M302L stabilizes the dephosphorylated inactive state, while M302G destabilizes the dephosphorylated inactive state. The addition of dasatinib increases the thermal stability of the wild type Src and M302L by ~3°C, while dramatically increasing the stability of M302G by ~8°C. Dasatinib is a Type I kinase inhibitor which binds the kinase active conformation. The addition of ligand is expected to stabilize a protein, however M302G appears significantly more stabilized by dasatinib compared to wild type and M302L. This suggests that the active conformation is more easily accessed by deposphorylated M302G. Kinase activity was

quantified by an ATP/NADH-coupled assay. Interestingly, both M302L and M302G are significantly less active than wild type Src with Src^{M302L} being the least active of the two mutants. While M302G seems to more easily access the kinase active conformation, this may not necessarily translate to increased kinase activity.

Bibliography

[1] W. Yeung, A. Kwon, R. Taujale, C. Bunn, A. Venkat, N. Kannan, Evolution of Functional Diversity in the Holozoan Tyrosine Kinome, Molecular Biology and Evolution. 38 (2021) 5625–5639. https://doi.org/10.1093/molbev/msab272.

[2] J.A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics. 23 (2007) 1875–1882. https://doi.org/10.1093/bioinformatics/btm270.

[3] D.I. McSkimming, S. Dastgheib, T.R. Baffi, D.P. Byrne, S. Ferries, S.T. Scott, A.C. Newton, C.E. Eyers, K.J. Kochut, P.A. Eyers, N. Kannan, KinView: a visual comparative sequence analysis tool for integrated kinome research, Mol. BioSyst. 12 (2016) 3651–3665. https://doi.org/10.1039/C6MB00466K.

[4] H. Chen, J. Ma, W. Li, A.V. Eliseenkova, C. Xu, T.A. Neubert, W.T. Miller, M. Mohammadi, A Molecular Brake in the Kinase Hinge Region Regulates the Activity of Receptor Tyrosine Kinases, Molecular Cell. 27 (2007) 717–730. https://doi.org/10.1016/j.molcel.2007.06.028.

[5] M. Mohammadi, J. Schlessinger, S.R. Hubbard, Structure of the FGF Receptor Tyrosine Kinase Domain Reveals a Novel Autoinhibitory Mechanism, Cell. 86 (1996) 577–587. https://doi.org/10.1016/S0092-8674(00)80131-2.

[6] M.A. Lemmon, J. Schlessinger, Cell Signaling by Receptor Tyrosine Kinases, Cell. 141 (2010) 1117–1134. https://doi.org/10.1016/j.cell.2010.06.011.

[7] S.C. Artim, J.M. Mendrola, M.A. Lemmon, Assessing the range of kinase autoinhibition mechanisms in the insulin receptor family, Biochemical Journal. 448 (2012) 213–220. https://doi.org/10.1042/BJ20121365.

[8] P.-L. Liu, L. Du, Y. Huang, S.-M. Gao, M. Yu, Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants, BMC Evolutionary Biology. 17 (2017) 47. https://doi.org/10.1186/s12862-017-0891-5.

[9] C.C. Lee, Y. Jia, N. Li, X. Sun, K. Ng, E. Ambing, M.-Y. Gao, S. Hua, C. Chen, S. Kim, P.-Y. Michellys, S.A. Lesley, J.L. Harris, G. Spraggon, Crystal structure of the ALK (anaplastic lymphoma kinase) catalytic domain, Biochemical Journal. 430 (2010) 425–437. https://doi.org/10.1042/BJ20100609.

[10] S.R. Hubbard, L. Wei, W.A. Hendrickson, Crystal structure of the tyrosine kinase domain of the human insulin receptor, Nature. 372 (1994) 746–754. https://doi.org/10.1038/372746a0.

[11] E. Schmid, A. Hotz-Wagenblatt, V. Hack, W. Dröge, Phosphorylation of the insulin receptor kinase by phosphocreatine in combination with hydrogen peroxide: the structural

basis of redox priming, The FASEB Journal. 13 (1999) 1491–1500. https://doi.org/10.1096/fasebj.13.12.1491.

[12] H. Xiao, M.P. Jedrychowski, D.K. Schweppe, E.L. Huttlin, Q. Yu, D.E. Heppner, J. Li, J. Long, E.L. Mills, J. Szpyt, Z. He, G. Du, R. Garrity, A. Reddy, L.P. Vaites, J.A. Paulo, T. Zhang, N.S. Gray, S.P. Gygi, E.T. Chouchani, A Quantitative Tissue-Specific Landscape of Protein Redox Regulation during Aging, Cell. 180 (2020) 968-983.e24. https://doi.org/10.1016/j.cell.2020.02.012.

[13] W. Xu, A. Doshi, M. Lei, M.J. Eck, S.C. Harrison, Crystal Structures of c-Src Reveal Features of Its Autoinhibitory Mechanism, Molecular Cell. 3 (1999) 629–638. https://doi.org/10.1016/S1097-2765(00)80356-1.

[14] N.H. Shah, J.F. Amacher, L.M. Nocka, J. Kuriyan, The Src module: an ancient scaffold in the evolution of cytoplasmic tyrosine kinases, Critical Reviews in Biochemistry and Molecular Biology. 53 (2018) 535–563. https://doi.org/10.1080/10409238.2018.1495173.

[15] T. Schindler, F. Sicheri, A. Pico, A. Gazit, A. Levitzki, J. Kuriyan, Crystal Structure of Hck in Complex with a Src Family—Selective Tyrosine Kinase Inhibitor, Molecular Cell. 3 (1999) 639–648. https://doi.org/10.1016/S1097-2765(00)80357-3.

[16] Q. Wang, E.M. Vogan, L.M. Nocka, C.E. Rosen, J.A. Zorn, S.C. Harrison, J. Kuriyan, Autoinhibition of Bruton's tyrosine kinase (Btk) and activation by soluble inositol hexakisphosphate, ELife. 4 (2015) e06074. https://doi.org/10.7554/eLife.06074.

[17] N.M. Levinson, O. Kuchment, K. Shen, M.A. Young, M. Koldobskiy, M. Karplus, P.A. Cole, J. Kuriyan, A Src-Like Inactive Conformation in the Abl Tyrosine Kinase Domain, PLOS Biology. 4 (2006) e144. https://doi.org/10.1371/journal.pbio.0040144.

[18] W. Yeung, Z. Ruan, N. Kannan, Emerging roles of the αC-β4 loop in protein kinase structure, function, evolution, and disease, IUBMB Life. 72 (2020) 1189–1202. https://doi.org/10.1002/iub.2253.

[19] S.S. Taylor, A.S. Shaw, N. Kannan, A.P. Kornev, Integration of signaling in the kinome: Architecture and regulation of the αC Helix, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1854 (2015) 1567–1574. https://doi.org/10.1016/j.bbapap.2015.04.007.

[20] O. Hantschel, B. Nagar, S. Guettler, J. Kretzschmar, K. Dorey, J. Kuriyan, G. Superti-Furga, A Myristoyl/Phosphotyrosine Switch Regulates c-Abl, Cell. 112 (2003) 845–857. https://doi.org/10.1016/S0092-8674(03)00191-0.

[21] M.K. Thakur, A. Kumar, S. Birudukota, S. Swaminathan, R. Tyagi, R. Gosu, Crystal structure of the kinase domain of human protein tyrosine kinase 6 (PTK6) at 2.33 Å resolution, Biochemical and Biophysical Research Communications. 478 (2016) 637–642. https://doi.org/10.1016/j.bbrc.2016.07.121.

[22] M. Srivastava, O. Simakov, J. Chapman, B. Fahey, M.E.A. Gauthier, T. Mitros, G.S. Richards, C. Conaco, M. Dacre, U. Hellsten, C. Larroux, N.H. Putnam, M. Stanke, M. Adamska, A. Darling, S.M. Degnan, T.H. Oakley, D.C. Plachetzki, Y. Zhai, M. Adamski, A. Calcino, S.F. Cummins, D.M. Goodstein, C. Harris, D.J. Jackson, S.P. Leys, S. Shu, B.J. Woodcroft, M. Vervoort, K.S. Kosik, G. Manning, B.M. Degnan, D.S. Rokhsar, The Amphimedon queenslandica genome and the evolution of animal complexity, Nature. 466 (2010) 720–726. https://doi.org/10.1038/nature09201.

[23] A. Kwon, S. Scott, R. Taujale, W. Yeung, K.J. Kochut, P.A. Eyers, N. Kannan, Tracing the origin and evolution of pseudokinases across the tree of life, Sci. Signal. 12 (2019). https://doi.org/10.1126/scisignal.aav3810.