Charles University
Faculty of Science

Study programme: Bioinformatics
Branch of study: Bioinformatics

Jan Hamalčík

The influence of non-domain regions' composition on the activity of multi-domain protein kinases

Vliv kompozice mimodoménových regionů na aktivitu vícedoménových protein kinás

Bachelor's thesis

Supervisor: doc. RNDr. Jiří Vondrášek, CSc.

Prague 2020

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

V Markarydu, Švédsko, 12.8.2020 Jan Hamalčík

Dedication.

I thank my supervisor doc. RNDr. Jiří Vondrášek, CSc., and especially my colleague David for their guidance, insights, and invaluable helpfulness throughout the research and writing process. The LaTeX template was provided by Mirek, who is thereby responsible for the overall delicate appearance of this thesis.

I am grateful to the owners of Smålandet Markaryds Älgsafari for generously providing me with a pleasant writing environment, including unlimited food, coffee, and beer supply.

A special thanks goes to my dearest Olga. You have been very patient and caring, while I was busy writing, your affectionate messages made me smile every time, and they encouraged me to go on and finish the thesis in this emotionally tough time, when I steadily miss you. I love you.

Last, but not least, I have to praise my parents for their everlasting support, and for their idea of me applying to the Bioinformatics study program.

Abstract

Kinases are among the most studied proteins, as they are important in many cellular processes. Today's research shows that their activity is not only dependent on the composition of the domains, but on the non-domain regions as well. This thesis tried to comprehend the influence of the linkers' composition on the function of multi-domain protein kinases in general. This was done by clustering human two-domain protein kinases with one protein kinase domain by the averaged physicochemical attributes of their inter-domain regions. The uniqueness of Gene Ontology terms and Enzyme Commission numbers within these clusters of proteins with different architectures was then investigated. However, due to multiple missteps, no such influence has been witnessed.

Keywords: protein domains, multi-domain proteins, protein function, non-domain regions, protein kinases

Abstrakt

Kinásy jsou jedny z nejvíce studovaných proteinů, neboť hrají důležitou roli v mnohých buněčných procesech. Aktuální výzkum prokazuje, že jejich aktivita není závislá pouze na komposici domén, nýbrž taktéž na nedoménových regionech. Tato bakalářská práce si dala za cíl obecně porozumět vlivu složení linkerů na funkci vícedoménových protein kinás. Dvoudoménové lidské protein kinásy s jedinou kinásovou doménou byly zklastrovány dle zprůměrovaných fysikálně chemických vlastností mezidoménových oblastí, načež se zkoumala unikátnost Gene Ontology a Enzyme Commission pojmů v rámci nalezených klastrů složených z proteinů různých architektur. Ovšem vzhledem k víceru nedostatkům, žádný takový vliv nebyl pozorován.

Klíčová slova: proteinové domény, multidoménové proteiny, proteinová funkce, nedoménové regiony, protein kinásy

Contents

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
ATP	adenosine triphosphate
EC	Enzyme Commission
GO	Gene Ontology
GTP	guanosine triphosphate
HMM	hidden Markov model
MSA	multiple sequence alignment
pI	isoelectric point
PK	protein kinase
UMAP	Uniform Manifold Approximation and Projection
URP.dat	`uniprot_reference_proteomes.dat`

Throughout this thesis, the standard one and three letter codes for the L-amino acids will be used.

Chapter 1

Introduction

1.1 Proteins and domains

Proteins are amino acid residue chains, polypeptides, that serve a variety of functions within living cells, including structural support and movement, interactions with cell's environment, and biochemical reaction catalysis [1]. To function properly, proteins have to fold into their native conformation, and as demonstrated by Anfinsen et al. [2] on bovine pancreatic ribonuclease, the information for correct folding is contained in the amino acid sequence itself.

Different sequences fold into different three-dimensional (3D) conformations. Certain local regions form secondary structures, such as α-helices and β-sheets, and these locally ordered regions associate to form the whole protein, or in the case of some larger proteins, to form folding units [3]. Levitt and Warshel [3] and Goldberg [4] define *protein domains* as folding units that would be stable if we would cleave them from the rest of the protein molecule. Nevertheless, today's definition requires that these subassemblies also conceivably function in isolation, and members of the same *domain family* tend to possess an ancient evolutionary relationship and often a similar function [5].

Many bioinformatical tools are available to perform domain identification within a protein sequence. The *SCOP* database organizes proteins of known three-dimensional structures according to their evolutionary and structural relationship [6]. It classifies non-redundant protein domains and defines them on two main levels of SCOP classification, family and superfamily. The designation of proteins in SCOP has been constructed mainly manually [7].

Other tools tend to implement semimanual approaches, often featuring profile hidden Markov models (HMMs), a powerful probabilistic method describing the sequence conservation in a family [8, 9]. The *Pfam* database [10] is such an example. Each protein family in Pfam consists of a seed alignment forming the basis to build an HMM-based profile [11] by engaging the HMMER software [12, 13]. Another protein structure classification database utilizing HMMs to scan protein sequences against it is called *CATH* [14]. CATH clusters proteins on four main levels, class (C), architecture (A), topology (T), and homologous superfamily (H) [15]. CATH, Pfam, and many more, are integrated in a general resource for protein families, domains, and func-

tional sites, called InterPro [16]. The authors aim to create a non-redundant characterization, and the software package InterProScan provides an interface to functionally classify novel nucleotide or protein sequences [17]. By uniting the member databases, InterPro exploits their individual strengths, thus significantly contributing in the troublesome effort of automatic annotation [18].

1.2 Protein kinase family

An example of a protein family is the *protein kinase* (PK) family, labeled as `PF00069` is the Pfam database. PKs are enzymes transfering a phosphate group from a phosphate donor, usually adenosine triphosphate (ATP), onto an acceptor amino acid in the substrate protein. PK domain autophosphorylation is possible as well. PKs can be classified based on the acceptor amino acid specificity; the phosphorylated residues are typically serine, threonine, tyrosine, or histidine [19].

The PK family is large and diverse [19, 20], and, generally, PKs are involved in most of the signal transduction in eukaryotic cells. They control a variety of cellular processes, among others metabolism, cell cycle progression, differentiation, transcription, cell movement, communication, and apoptosis [21–27]. As PKs function as a responsive regulatory system, it is not surprising that their turnover rate is rather fast, with a half-life averaging less than an hour [28]. Misproper PK activity can lead to unfavorable cell transformation and cancer [29, 30], and therefore, they are a fascinating subject of study in molecular biology and bioinformatics.

The 3D fold of PK domains is conserved [31] with common structural features specific to the PK family [32]. The enzyme structure is bilobal. The smaller lobe on the N-terminus is composed of a five-stranded antiparallel β-sheet and one α-helix [33] called αC, which takes part in the stabilization of the active conformation [34, 35]. The broadly helical larger lobe associated with the C-terminus adds up to have seven α-helices and one antiparallel β-sheet placed on the surface of the cleft between the two lobes, where the catalytic site is located [32, 33, 36]. Twelve major subdomain elements can be further recognized throughout the PK family [20, 37, 38].

Multiple sequence alignments (MSAs) of domains from the PK family uncover many highly conserved residues and motifs, some of which are strictly required for the PK activity. The nucleotide binding consensus sequence Gly-X-Gly-X-X-Gly is found in subdomain I and contains the first, second, and third essential glycine residue. Any side chain at the glycines' positions would interfere with the incoming ATP (or guanosine triphosphate (GTP)). The whole motif turns sharply around the nucleotide, with the first essential glycine being in contact with ribose, while the second provides space for the pyrophosphate group [20, 41]. The β-phosphate moiety's oxygens are hydrogen bonded with the backbone amides of the residues around the third fundamental glycine, and the side chains surrounding this motif contribute to the hydrophobic pocket for the adenine ring of ATP [38]. GTP can be utilized as well, however, ATP will always have a lower Michaelis constant K_{m} [42]. Regardless of which nucleotide triphosphate is used, the phosphate

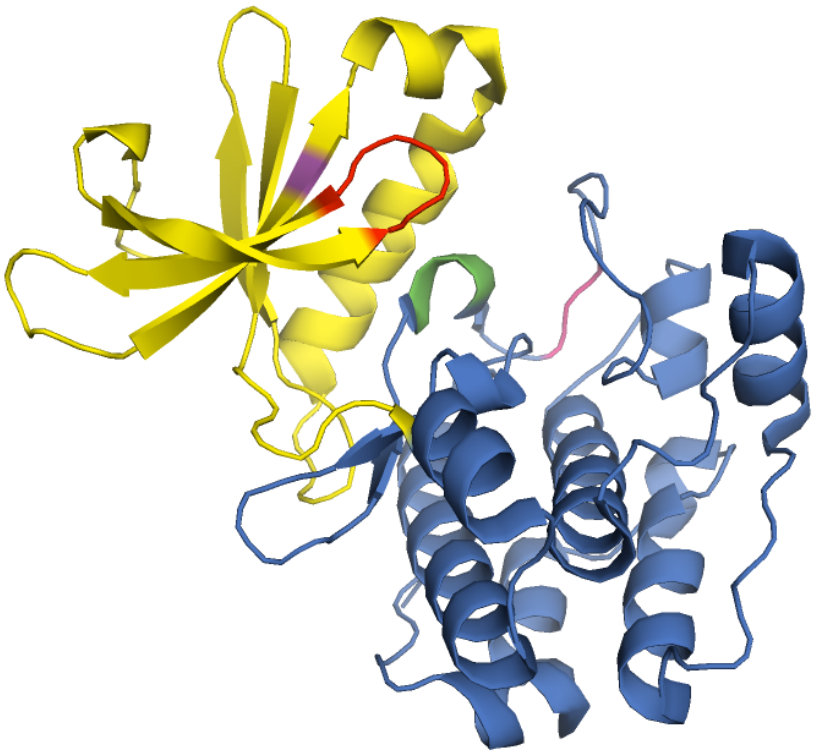

Figure 1.1 The bilobal structure of the PK domain from the Aurora A PK. Structure downloaded from the PDBe, entry code 4dee [39], visualized with the PyMOL Molecular Graphics System [40]. Yellow: The small lobe. Blue: The large lobe. Red: Gly-X-Gly-X-X-Gly motif. Magenta: The invariant lysine. Green: DFG motif. Pink: APE motif.

donor is bound as a complex together with a divalent cation, with Mn^{2+} usually preferred over Mg^{2+} and others [43–45].

An invariant lysine in subdomain II corresponding to protein kinase A-Cα Lys72 is thought to be the best characterized catalytic domain residue [20]. It forms a salt bridge with the carboxyl group of the practically invariant glutamate in subdomain III, which stabilizes the interactions between the lysine and the α- and β-phosphates of ATP [38]. As Kamps and Sefton [46] showed in their work on the $p60^{src}$ tyrosine PK, all substitutions of the lysine including arginine and histidine, result in a loss of the PK activity.

A highly conserved Asp-Phe-Gly (DFG) triad can be found in subdomain VII. It is a part of the so-called activation loop and helps orient the γ-phosphate of the ATP for transfer [38]. The activation loop is centrally located and provides a platform for the peptide substrate by interacting with the αC-helix [34, 35]. Autophosphorylation of certain residues within the activation loop, three tyrosines in case of the insulin receptor tyrosine PK, results in an active PK state; an unphosphorylated loop traverses through the cleft between the two lobes of the PK domain, thus rendering both the protein substrate and the ATP binding sites inaccessible [47]. The recognition of peptide substrates in the active state is then mediated by the Ala-Pro-Glu (APE) motif located in subdomain VIII [38].

Many other hydrophobic residues that do not form any primary sequence motif, or any particular secondary structure, characterize active PKs, as they either stabilize the whole domain, or regulate its overall activity [48, 49]. Especially one residue called the "gatekeeper" lying in the hinge region between the N and C lobes of the PK domain has recently been of great interest in PK inhibitor development. This gatekeeper residue guards a small hydrophobic cavity neighboring the ATP-binding site [50] and it anchors PK inhibitors bound in the ATP pocket [36, 51]. The particular type of the gatekeeper residue is specific to each PK domain and the size and shape of the side chain determine the cavity's druggability. Moreover, drug design is strongly affected by the adjacent DFG motif's conformation as well [52].

Still, not all members of the PK family possess the PK enzymatic activity [53–55]. Such seemingly inactive domains may have noncatalytic functions, or do not require the conserved catalytic residues mentioned above and use a modified catalytic mechanism instead [56]. However, as about two thirds of prokaryote proteins and around eighty percent of eukaryote proteins are multi-domain proteins [57, 58], the PK domains can also modulate other catalytic regions.

1.3 Multi-domain proteins

With a limited number of domain families present [59, 60], creating new functions may be somewhat difficult. Nature overcomes this obstacle by combining old building blocks instead of inventing new ones [61]. Müller, MacCallum, and Sternberg [62] estimated that 98% of the domains in the human proteome are duplicates; duplication is in fact one of the main sources for creation of new genes [63]. Consequently, it is not surprising that the major molecular mechanism leading to multi-domain proteins and novel combi-

nation is non-homologous recombination, which is sometimes referred to as "domain shuffling" [64].

Domains are not merged aimlessly when creating multi-domain proteins. The domain combinations seen in nature can be discriminated from a random model of domain combination, as shown by Apic, Huber, and Teichmann [65]. Putting together specific superfamilies results in more specific functions for individual molecules, and proteins with the same domain arrangement tend to be evolutionarily and functionally related [64, 66, 67]. The sequential order of protein families identified within a protein is called *domain architecture*.

The set of all architectures seems to be limited as well. A pattern is observed, where most domains tend to have only one or few combination partners in the context of multi-domain proteins. Moreover, if a specific domain is found on the N-terminus of a particular multi-domain protein, other members of the same family are usually to be found on the N-terminus as well, and vice versa. The orientation and type of neighborhood varies only in a few domain families, most of which are large and versatile [61, 68].

1.3.1 Non-domain regions

Domains within the same protein are connected by non-domain amino acid stretches. These can be long or short, and often possess a disordered structure. Considering the limited number of domains and architectures in nature, these *linkers* introduce new possibilities for structural assemblies, and may regulate, or sometimes even take part in the proteins' activity and stability [69]. Therefore, depending on the functionality of the specific domain, its associated linker generally requires a certain amino acid sequence [70]. It has been shown by Jakubec et al. [71] that residues in pairs of domains coevolve, and responses to mutations in residue pairs are also observed in non-domain regions [72], thus ensuring a suitable environment for the whole molecule.

Linkers often serve as rigid spacers between two domains. These molecular rulers are mostly α-helical, and they prevent unfavorable interactions between the neighboring modules [73, 74]. Mutations in such non-domain regions are not expected to affect the function of a protein in any way [75]; in different circumstances, however, alterations in the linker region can have an effect on the stability, proteolytic resistance, or solubility of single-chain proteins [76]. In particular, proline is the most common residue in linkers, and careful selection of residues around prolines is of utmost importance. Its stiff nature helps prevent ominous contacts of linkers and domains, as it can not participate comfortably in any standard secondary structure conformation due to its inability to hydrogen bond to other moieties [73, 74]. Other amino acids typical for linkers are glutamine and other polar and charged amino acids. Contrarily, residues with hydrophobic or aromatic side chains are more common in domains [77].

Nevertheless, many linkers are also soft and intrinsically disordered. The non-domain region of human immunoglobulin G1 exemplifies such a flexible amino acid sequence [78]. These often promote fundamental catalytic events in the overall function of proteins, as observed in the packaging of the tomato bushy stunt virus protein [79]. Since soft linkers connecting domains are not

merely flexible, but also allosterically regulated, it is no wonder that they are capable of facilitating protein folding and conformational changes of the whole molecule. The amino acid sequences of linkers, particularly of the residues in contact between linkers and adjoint domains, encode conformational states through which signals travel [73, 80]. Examples of such signals may be the phosphorylation of a distant residue or ATP binding. This can be illustrated by an otherwise flexible linker in Src PKs, which clamps SH2 and SH3 domains upon C-terminal tyrosine phosphorylation [81]. On the other hand, ATP binding causes the immobilization of neck linkers of kinesins, which subsequently extend towards the plus end of a microtubule, thus giving kinesins their forward drive [82–84].

Changes in both linker length and composition can alter folding kinetics, function, and stability of proteins [76, 85], as demonstated in many studies. A dramatic increase of the isoelectric point (pI) from 5.86 to 9.81 was observed after a deletion of 40 residues from a linker in mycobacteriophage D29 endolysin [86]. Conformations of protein domains in cellulosomes are primarily influenced by the length of the linkers, as the stiffness of the linkers is inversely proportional to the linker length [87]. In the kinesins mentioned above, the effectivity of kinesin runs is determined by the length of the neck linker [88]. Domain functions are predominantly affected by short linkers, notably those that are buried as well. Contrarily, long linkers permit 3D arrangements of the domains similar to those accessible in multi-domain constructs with the domains swapped, hence not having so much impact on the overall protein activity [67].

Other authors present examples where the amino acid composition is more relevant for the protein function compared to length. Klement et al. [89] concluded this after exploring the functionality of cytotoxic engineered antibody fragments. In addition, Ikebe et al. [90] demonstrated the importance of selected linker residues in smooth muscle myosins, in which the actin-translocating activity was terminated upon deletion or substitution of these amino acids. As stated previously, a conserved linker sequence is crucial for the kinesins' microtubule-based motility as well [91, 92].

Yet, it remains to be uncovered how the composition of the non-domain regions affects the function of multi-domain PKs in general. For instance, the function of PKs is regulated not only by the common structural elements of the PK domain, but also by the linkers [93]. These non-core segments do not show any significant sequence similarity, and the composition is highly variable [32]. This thesis aims to expose the relationship between the linkers' composition and PK activity and specificity. This will be done by analyzing sequences of two-domain proteins containing exactly one PK domain, clustering the inter-domain regions by computationally acquired physicochemical properties averaged over the linker sequences, and embedding unique *Gene Ontology* (GO) [94, 95] terms and *Enzyme Commission* (EC) [96, 97] numbers into the identified linker groups.

Chapter 2

Methods

Protein sequences and MSAs of Pfam domain sequences were obtained from The European Bioinformatics Institute's FTP[1]. The archived files `Pfam-A.full` and `uniprot_reference_proteomes.dat` (URP.dat) corresponding to Pfam's release 32.0 were downloaded; the MSAs consist of sequences from the UniProt reference proteomes version 2018_04. Protein family annotation software pfamannot[2] has then been implemented in C++17 to parse these files and to extract architectures of all proteins containing at least one `PF00069` domain, starting and ending positions of all domains present within such proteins, primary sequences of the whole molecules, and the organisms from which they originate. A total of 330,302 such proteins were identified, of which 127,697 were multi-domain proteins.

To reduce the evolutionary noise, only multi-domain PKs from humans were selected for further analysis. A total of 542 multi-domain PKs were identified. The eukaryotic protein subcellular location predictor DeepLoc-1.0 [98] was then applied to these proteins[3], and the molecules with their subcellular location predicted to be the cytoplasm or nucleus were chosen, as the amino acid composition, thus also the physicochemical properties of membrane proteins differ significantly from their soluble counterparts [99–101]. In addition, membrane proteins have their conformational degrees of freedom reduced due to their placement in the lipid bilayer, and are therefore not suitable for the purpose of this thesis. Two-domain proteins with a single PK domain were then selected from the emerging 244 cytosolic or nuclear human multi-domain PKs, yielding the final dataset containing a total of 117 protein molecules.

Two sets of physicochemical properties were then acquired for the inter-domain regions in the final dataset. The first set, which was in the spotlight of this study, consists of quadruples composed of the logarithm of the linker's length, the linker's pI, the percentage of charged amino acids in the linker, and the linker's GRAVY index [102]. The pI and the GRAVY index roughly describe the linker's electric charge and its hydrophobicity, respectively. These values were calculated with the ExPASy's ProtParam

[1] `ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/`

[2] `https://github.com/hamalcij/pfamannot`

[3] Sequences spanning more than 6,000 residues were discarded due to DeepLoc's length limitation.

tool [103], and their distributions across the whole dataset were computed. In the second set, each linker was described using the 553 physicochemical and biochemical properties of amino acids from the AAindex database, version 9.2 [104–108]. A dimensionality reduction technique Uniform Manifold Approximation and Projection (UMAP) [109] was then applied to various normalized subsets of both sets, thus producing two-dimensional (2D) representations of the feature spaces aminable to a visual analysis resulting in, for example, the recognition of cluster structure in the data.

The EC and GO terms associated with proteins in the URP.dat file were parsed. The former describes the enzyme specificity [96], the latter terms biological processes, cellular components, and molecular functions of genes and their products [94, 95]. The GO is a weakly hierarchical vocabulary [94], and the hierarchy level of GO terms in the URP.dat file is not standardized. Therefore, for each protein, parents of its terms were recursively obtained and merged from the `go-basic.obo` file[4], release 2020-06-01, thus enabling cluster analysis on different GO levels. GO terms labeled as obsolete were excluded. The EC and GO terms were then embedded into the UMAP projections and the presence of unique terms within the identified clusters was examined. The same process was executed on the density clusters (see below) as well. The results were visualized using the Python graphics package Matplotlib [110], version 1.5.1.

[4] `http://current.geneontology.org/ontology/go-basic.obo`

Chapter 3

Results

3.1 The first set of physicochemical attributes

In this section, results for the first set of physicochemical properties described above will be presented. These include the length, pI, GRAVY index, and percentage of charged amino acids within the inter-domain regions of human two-domain PKs containing one PK domain.

3.1.1 Feature distributions (densities)

The density analysis gave interesting insights into the distribution of physicochemical properties within the studied dataset. The linker length density peaks around 75 residues. A small portion of extremely long inter-domain regions spanning 500 to 800 residues is also present. However, it is questionable, whether these "linkers" might not represent a yet unrecognized domain, and thus not being suitable for the definition of a linker [111].

A nontrivial outcome has been observed while studying the density of linkers' pI, visualized in figure 3.1. Most non-domain regions in human two-domain PKs are either very acidic, or, on the other hand, very basic. Interestingly, the region of a neutral pH contains almost no proteins, forming an empty gap around pH 8. This allows for dividing the studied PKs into two groups based on their linker's acidity.

Au contraire, neither the percentage of charged amino acids within the inter-domain regions nor the linkers' GRAVY index form distinct clusters. Their densities show only single local maxima. All linkers from the final dataset are hydrophilic; the fraction of charged residues averages around 0.25.

3.1.2 UMAP dimensionality reduction

Three clusters corresponding to three distinct linker types were identified visually in the UMAP projection of the first set of the linkers' physicochemical attributes. Throughout this thesis, the linker types will be referred to based on the clusters' positioning in figure 3.2: "L-linkers" for the lower left cluster, "M-linkers" for the middle cluster, and "R-linkers" for the right cluster. L-linkers are best characterized by their extreme length of more than

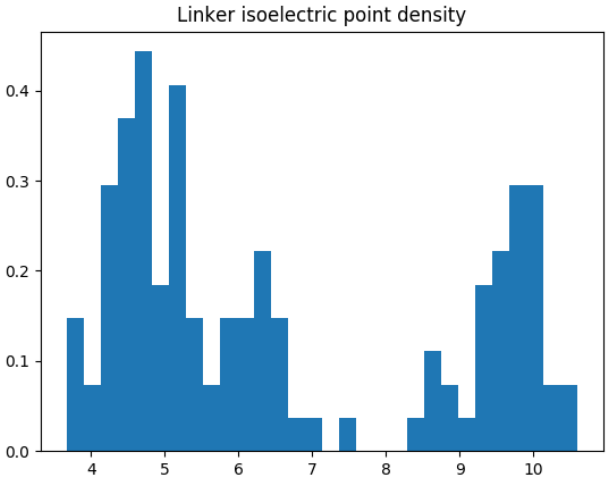

Figure 3.1 The pI density of inter-domain regions from human two-domain PKs in the studied dataset. Two groups of proteins are evident, one with acidic linkers, one with basic linkers.

600 residues, R-linkers by their extremely basic pI, and M-linkers by their extremely acidic pI.

By embedding the architectures of the proteins from the studied dataset into the UMAP representation it became visible that many architectures feature more than one linker type, as seen in figure 3.3. Especially, the architecture `PF00069_PF00780` stands out, being endowed with all three linker classes. On the other hand, most architectures equipped with linkers from only one category are represented by a very small subset of proteins from the set of human two-domain PKs.

3.1.3 Gene Ontology terms

All considered proteins had at least two GO terms assigned, one of them being always `GO:0005524; F:ATP binding`. The second most common GO term was `GO:0004674; F:protein serine/threonine kinase activity`, and the third most frequent GO term was `GO:0004672; F:protein kinase activity`.

However, the search for unique GO terms within both the UMAP representation and the pI clustering of the non-domain regions on different GO hierarchical levels gave no satisfactory results, albeit discovering many GO terms exclusive for each examined linker group. Unfortunately, the numbers of proteins carrying these terms were too low to permit a robust analysis. No single GO terms were identified which would be associated exclusively with all members of individual clusters.

There were only two GO terms standing out regarding their proportion. `GO:0000287; F:magnesium ion binding` found on hierarchy level 6 is exclusive to the acidic linker group, as well as to the M-linkers. It was encountered in 12 proteins covering 8 different architectures, including those having the PK domain on the N-terminus, as well as those carrying the PK do-

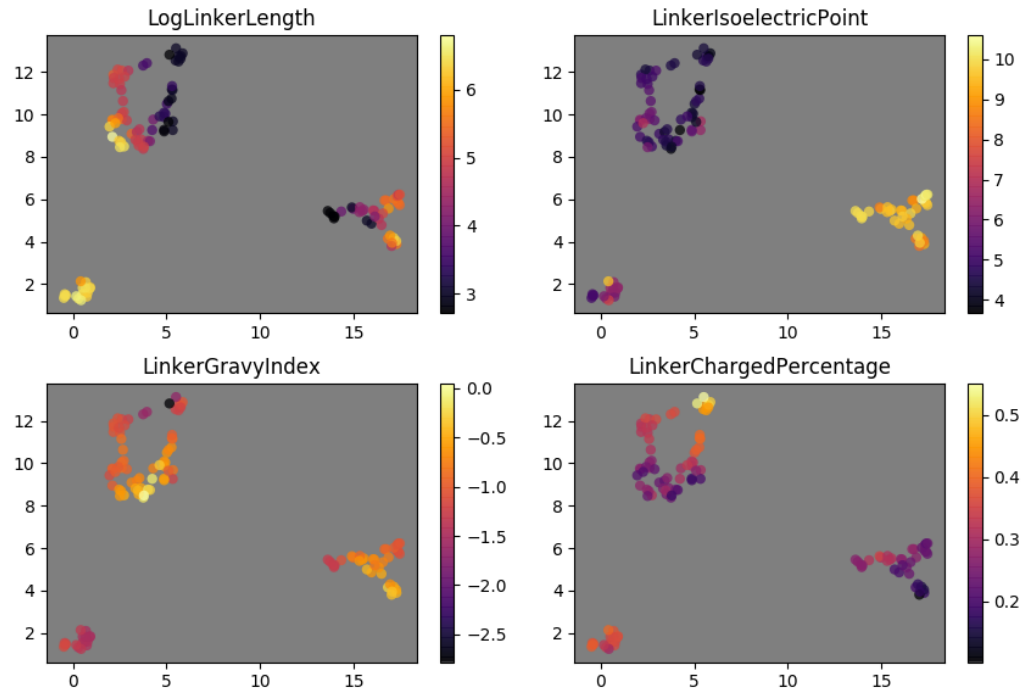

Figure 3.2 The UMAP dimensionality reduction on the normalized linker attributes: logarithm of the linker's length, its pI, its GRAVY index, and a percentage of charged amino acids within the inter-domain regions of human two-domain PKs. Each subfigure presents the distribution of the values of the respective property.

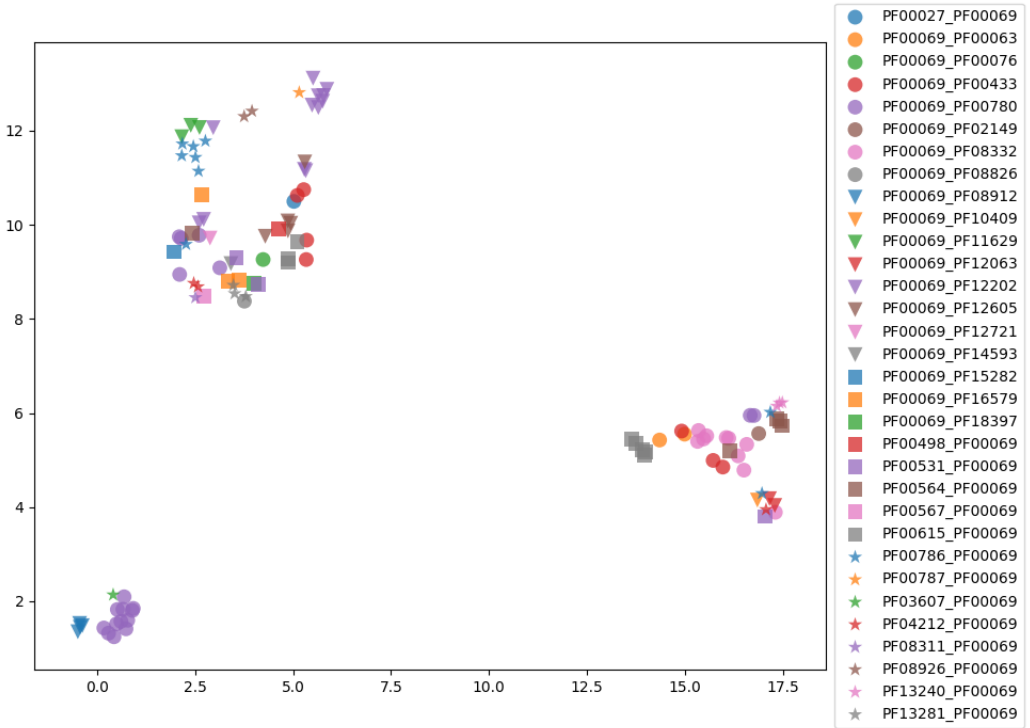

Figure 3.3 Architectures of the studied molecules embedded into the UMAP projection of the four physicochemical attributes of the inter-domain regions.

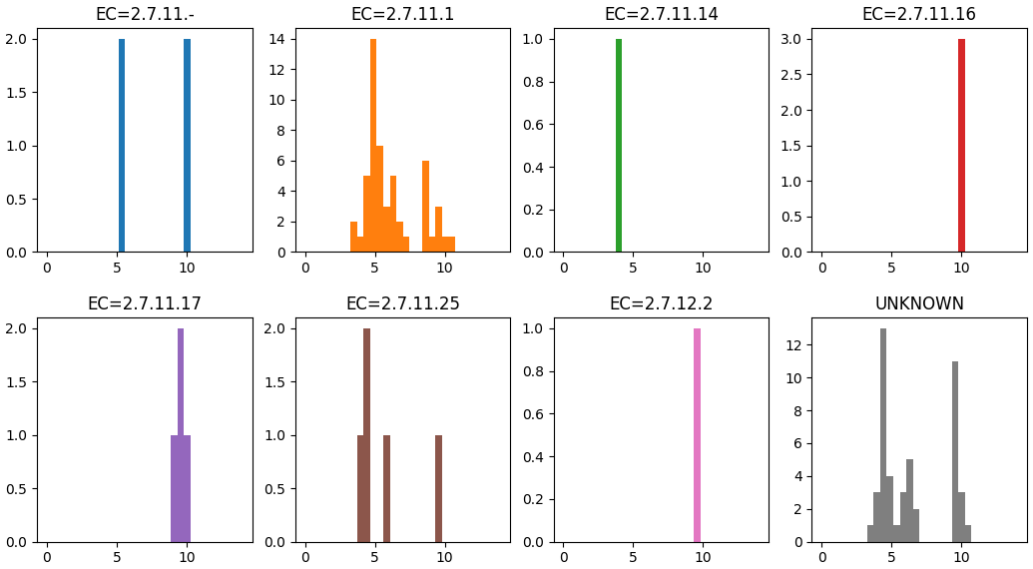

Figure 3.4 Occurrences of EC numbers in human two-domain PKs. The x-axis represents the pI of the molecules' linkers.

main on the C-terminus. The second vocable is `GO:0005516; F:calmodulin binding`, appearing on the hierarchy level 4. This term is limited to the basic linker group and to the R-linkers. In this case, all 10 proteins accredited with this GO term possess the same architecture.

3.1.4 Enzyme Commission numbers

Out of the 117 human two-domain PKs, 70 molecules have an EC number assigned in the URP.dat file. For the remaining 47, the EC term will be reffered to as "`UNKNOWN`". 3 kinases had more than one EC number specified; some of these extra terms are outside of the `2.7.11 protein-serine/threonine kinases` group. Only 14 molecules had, however, EC classification other than `2.7.11.-`, `2.7.11.1 non-specific serine/threonine protein kinase`, or `UNKNOWN`.

Figure 3.4 shows the distribution of the linkers' pI for each EC number. All 5 specialized EC terms are vastly underrepresented within the studied dataset, yet, for example, the `2.7.11.25 mitogen-activated protein kinase kinase kinases` are present in both acidic and basic linker clusters. Contrarily, basic inter-domain regions seem to be specific to both `2.7.11.16 G-protein coupled receptor kinases` and `2.7.11.17` Ca^{2+}`/calmodulin-dependent protein kinases`; nevertheless, they are not to be found in more than one architecture.

The inadequate representation of the EC numbers in the final dataset was also the reason behind the unsatisfactory results of embedding the terms into the UMAP representation. On one hand, the numbers `2.7.11.14`, `2.7.11.16`, and `2.7.11.17` are specific to one of the three defined linker groups; on the other hand, they are not to be found in more than one architecture, as already mentioned above. In conclusion, the first set of physicochemical attributes did not provide any functionally relevant results.

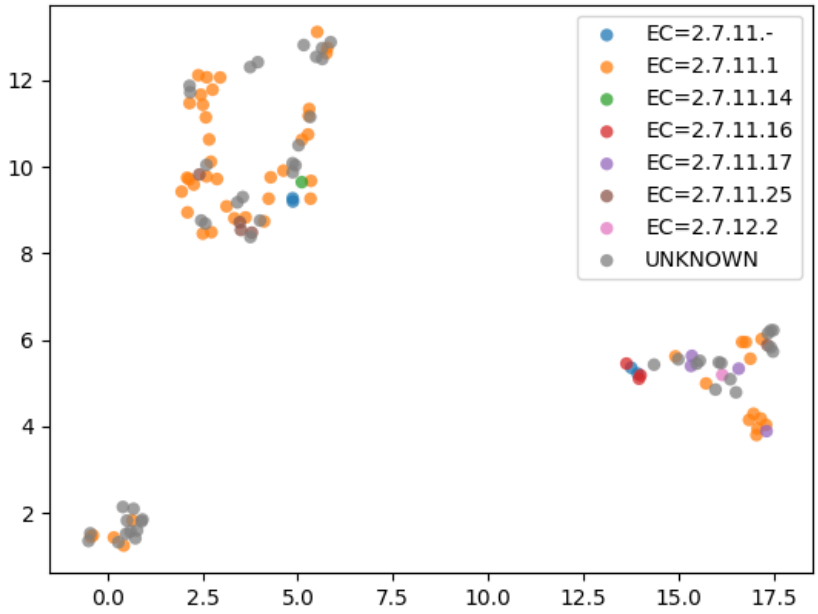

Figure 3.5 Embedding of the EC numbers into the UMAP representation.

3.2 The second set of physicochemical attributes

The two-dimensional UMAP representation of the 553 physicochemical data from the AAindex database did not reflect the clustering of the 4 physicochemical attributes described above. As seen in figure 3.6, one small cluster of inter-domain regions from human two-domain PKs emerges on the right side of the visualization; however, it contains only proteins with the same architecture `PF00069_PF12202`. In the large cloud on the left side, generally, molecules with the same architecture create local bundles and they only seldom mix with proteins with other architectures. Due to the lack of clearly distinguishable clusters, no GO or EC embedding was performed.

In conclusion, no obvious influence of the inter-domain regions' composition on the activity or specificity of human two-domain PKs with exactly one PK domain was observed, as no clustering of the linkers' physicochemical properties resulted in the colocalization of like GO or EC terms associated with proteins with different architectures.

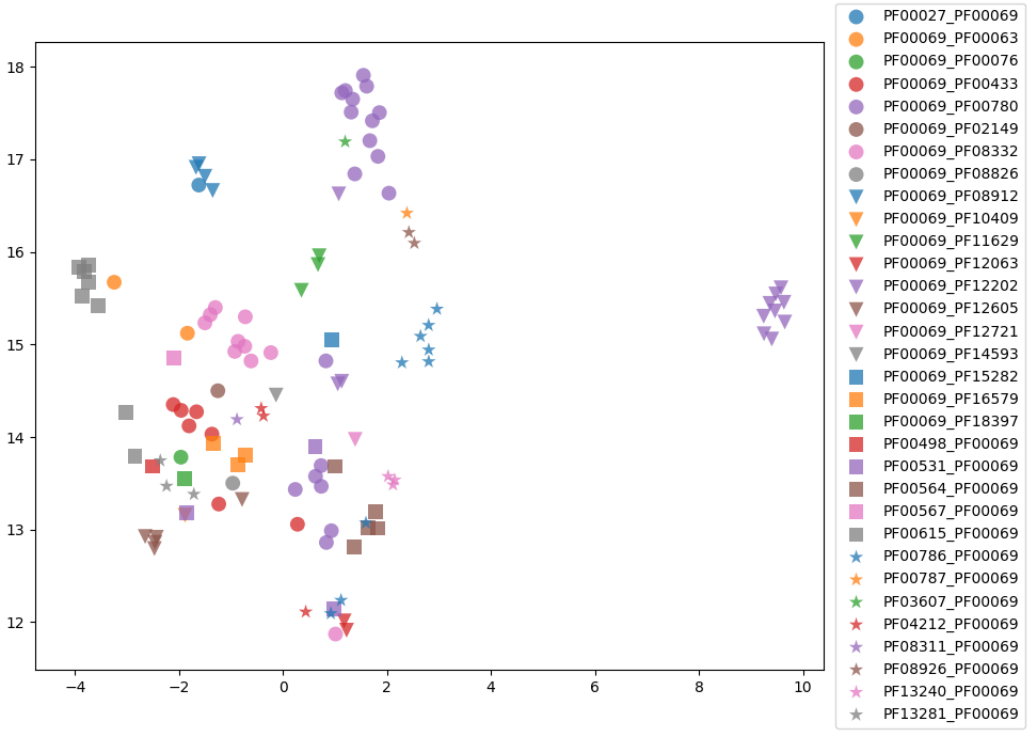

Figure 3.6 Architectures of the two-domain PKs with a single PK domain embedded into the UMAP representation of 553 physicochemical properties from the AAindex database.

Chapter 4

Discussion

The length and amino acid composition of non-domain regions can be crucial for the regulation of multi-domain proteins' activity, including PKs [93, 112]. However, to our knowledge, the influence of the linkers' composition on the overall protein function has not been described *in general* yet. This thesis tried to address this issue by selecting a dataset of evolutionarily related multi-domain proteins containing a PK domain, clustering these molecules based on the physicochemical attributes of their inter-domain regions, and by identifying GO terms and EC numbers specific to the detected clusters of proteins with various architectures.

Even though it was possible to divide proteins from the studied dataset into three groups based on the clusters visible in the UMAP representation of the 4-dimensional space of the linkers' normalized physicochemical characters, no frequent functional annotation terms could characterize the defined clusters. There may be several reasons for the lack of success of the designed method. For example, there may really not be any overall influence of the inter-domain regions' composition on the function of two-domain PKs. However, based on the literature search presented above, this proposition seems rather improbable [70–72, 76, 79, 80, 82, 84–93, 113–116].

The narrow dataset may be a significant problem. In a sample of large size, more information is available; on the other hand, false presumptions may be concluded from a small sample generating the observations [117, 118]. In the studied dataset, 32 different architectures were present within the 117 molecules. Proteins with the same architecture are frequently related, both evolutionarily and functionally [64, 66, 67], and the linkers may introduce new structural features [69], thus enhancing the functionality. With 3.65625 proteins per architecture on average, no robust conclusion can be made on how the composition of the linkers affects the behavior of these molecules.

Furthermore, both functional annotation services, GO and EC, have drawbacks regarding their completeness and homogeneity. For instance, Gaudet and Dessimoz [119] claim that the GO is biased and unevenly incomplete. The GO database is a reflection of primary literature [120], therefore, less studied areas are inadequately represented in the GO. On the other hand, more comprehensively covered parts of the GO are not flawless either, as they can provide contradictory information which can be caused, for example, by differences in experimental conditions of comparable research [121, 122].

Within the studied 117 human two-domain PKs, a total of 1,982 GO annotations were fetched from the UPR.dat file, peaking at 69 terms assigned to a single protein, namely the AAPK1_HUMAN PK. These include even the GO terms from the same hierarchy. For example, it is possible to reach GO:0008610; lipid biosynthetic process by recursively applying the transitive relationship "is_a" on GO:0045542; positive regulation of cholesterol biosynthetic process, which are both explicitly mentioned in the uniprot_reference_proteomes.dat file. Furthermore, the same protein is classified as 2.7.11.1 non-specific serine/threonine protein kinase; however, at the same time, has the specific EC numbers 2.7.11.26, 2.7.11.27, and 2.7.11.31 assigned to it as well.

To be assigned an EC number, there must be a direct experimental evidence that the proposed enzyme actually catalyses the claimed reaction [123]. 60 proteins from the studied dataset were described as 2.7.11.1 non-specific serine/threonine protein kinase, meaning that these PKs are either non-specific, or their specificity has not been analyzed to date[1]. The strict nature of the EC, as well as the excessiveness of the GO combined with insufficient amount of research could be overcome by measuring similarities of the GO terms [124, 125] within the protein clusters instead of enforcing their absolute exclusion.

Furthermore, it may be desirable to employ a more stable method than UMAP to produce a significant outcome. The low-dimensional representation generated by UMAP is dependent on the choice of several hyperparameters[2]. Besides, UMAP is a stochastic algorithm, and exact reproduction of the results is only possible by fixing a random seed state[3]. The observed linker types are therefore determined by UMAP parametrization.

The elucidation of the effect of the linkers' composition on protein activity could improve our ability to predict the function of multi-domain proteins, but further research is needed to disclose the presented problem. Throughout this thesis, only rough and averaged characteristics of the whole linkers were considered. In reality, only a couple of particular residues may have an influence on the protein activity, the rest of them may be unimportant. The herein proposed method is not capable of resolving such residues.

[1] https://www.brenda-enzymes.org/enzyme.php?ecno=2.7.11.1

[2] https://umap-learn.readthedocs.io/en/latest/parameters.html

[3] https://umap-learn.readthedocs.io/en/latest/reproducibility.html

Chapter 5

Conclusion

The results of this thesis did not provide clear evidence of the effect of the inter-domain regions' composition on the function, activity, or specificity of human two-domain PKs. The proteins could be grouped according to their linker's pI, and three linker types were defined from the UMAP representation of the quadruples of the averaged physicochemical attributes of the linkers. However, no GO term or EC number could characterize the identified protein groups. A different low-dimensional representation of the linkers' properties was obtained when a larger set of physicochemical attributes was taken into account.

The presented method relied on the informativeness of the functional annotation services. Nevertheless, the EC and the GO were found to be too general and too precise, respectively. To reduce the detail of the GO, analysis on different GO hierarchy levels was put through, yet, the identification of unique terms within the clusters has been equally unsuccessful. As the complete proteome of human two-domain PKs with a single PK domain was examined, the correlation between the average linker region's properties and the PK activity can be hereby excluded.

Bibliography

[1] Bruce Alberts. "Molecular biology of the cell". In: (2018).

[2] Christian B Anfinsen et al. "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain". In: *Proceedings of the National Academy of Sciences of the United States of America* 47.9 (1961), p. 1309.

[3] Michael Levitt and Arieh Warshel. "Computer simulation of protein folding". In: *Nature* 253.5494 (1975), pp. 694–698.

[4] Michel E Goldberg. "Tertiary structure of Escherichia coli β-D-galactosidase". In: *Journal of Molecular Biology* 46.3 (1969), pp. 441–446.

[5] Chris P Ponting and Robert R Russell. "The natural history of protein domains". In: *Annual review of biophysics and biomolecular structure* 31.1 (2002), pp. 45–71.

[6] Alexey G Murzin et al. "SCOP: a structural classification of proteins database for the investigation of sequences and structures". In: *Journal of molecular biology* 247.4 (1995), pp. 536–540.

[7] Antonina Andreeva et al. "The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures". In: *Nucleic acids research* 48.D1 (2020), pp. D376–D382.

[8] Anders Krogh et al. "Hidden Markov models in computational biology. Applications to protein modeling". In: *Journal of molecular biology* 235.5 (1994), pp. 1501–1531.

[9] Sean R Eddy. "Hidden markov models". In: *Current opinion in structural biology* 6.3 (1996), pp. 361–365.

[10] Erik LL Sonnhammer, Sean R Eddy, and Richard Durbin. "Pfam: a comprehensive database of protein domain families based on seed alignments". In: *Proteins: Structure, Function, and Bioinformatics* 28.3 (1997), pp. 405–420.

[11] Sara El-Gebali et al. "The Pfam protein families database in 2019". In: *Nucleic acids research* 47.D1 (2019), pp. D427–D432.

[12] Robert D Finn et al. "The Pfam protein families database". In: *Nucleic acids research* 38.suppl_1 (2010), pp. D211–D222.

[13] Robert D Finn, Jody Clements, and Sean R Eddy. "HMMER web server: interactive sequence similarity searching". In: *Nucleic acids research* 39.suppl_2 (2011), W29–W37.

[14] Natalie L Dawson et al. “CATH: an expanded resource to predict protein function through structure and sequence”. In: *Nucleic acids research* 45.D1 (2017), pp. D289–D295.

[15] Christine A Orengo et al. “CATH–a hierarchic classification of protein domain structures”. In: *Structure* 5.8 (1997), pp. 1093–1109.

[16] Robert D Finn et al. “InterPro in 2017—beyond protein family and domain annotations”. In: *Nucleic acids research* 45.D1 (2017), pp. D190–D199.

[17] Evgeni M Zdobnov and Rolf Apweiler. “InterProScan–an integration platform for the signature-recognition methods in InterPro”. In: *Bioinformatics* 17.9 (2001), pp. 847–848.

[18] Rolf Apweiler et al. “InterPro—an integrated documentation resource for protein families, domains and functional sites”. In: *Bioinformatics* 16.12 (2000), pp. 1145–1150.

[19] Tony Hunter. “[1] Protein kinase classification”. In: *Methods in enzymology*. Vol. 200. Elsevier, 1991, pp. 3–37.

[20] Steven K Hanks, Anne Marie Quinn, and Tony Hunter. “The protein kinase family: conserved features and deduced phylogeny of the catalytic domains”. In: *Science* 241.4861 (1988), pp. 42–52.

[21] Bruce E Kemp et al. *AMP-activated protein kinase, super metabolic regulator*. 2003.

[22] Shuhei Matsuoka, Mingxia Huang, and Stephen J Elledge. “Linkage of ATM to cell cycle regulation by the Chk2 protein kinase”. In: *Science* 282.5395 (1998), pp. 1893–1897.

[23] Gary L Johnson and Richard R Vaillancourt. “Sequential protein kinase reactions controlling cell growth and differentiation”. In: *Current opinion in cell biology* 6.2 (1994), pp. 230–238.

[24] Linda Vermeulen et al. “Transcriptional activation of the NF-κB p65 subunit by mitogen-and stress-activated protein kinase-1 (MSK1)”. In: *The EMBO journal* 22.6 (2003), pp. 1313–1324.

[25] Philip Chen, Kiran Gupta, and Alan Wells. “Cell movement elicited by epidermal growth factor receptor requires kinase and autophosphorylation but is separable from mitogenesis”. In: *The Journal of cell biology* 124.4 (1994), pp. 547–555.

[26] Bonnie J Warn-Cramer et al. “Regulation of connexin-43 gap junctional intercellular communication by mitogen-activated protein kinase”. In: *Journal of Biological Chemistry* 273.15 (1998), pp. 9188–9196.

[27] Timothy G Cross et al. “Serine/threonine protein kinases and apoptosis”. In: *Experimental cell research* 256.1 (2000), pp. 34–41.

[28] Tony Hunter. “Phosphotyrosine—a new protein modification”. In: *Trends in biochemical sciences* 7.7 (1982), pp. 246–249.

[29] Jussi Koivunen, Vesa Aaltonen, and Juha Peltonen. "Protein kinase C (PKC) family in cancer progression". In: *Cancer letters* 235.1 (2006), pp. 1–10.

[30] Antonio Caretta and Carla Mucignat-Caretta. "Protein kinase a in cancer". In: *Cancers* 3.1 (2011), pp. 913–926.

[31] Peter Man-Un Ung, Rayees Rahman, and Avner Schlessinger. "Redefining the protein kinase conformational space with machine learning". In: *Cell chemical biology* 25.7 (2018), pp. 916–924.

[32] Susan Serota Taylor and Elzbieta Radzio-Andzelm. "Three protein kinase structures define a common motif". In: *Structure* 2.5 (1994), pp. 345–355.

[33] Daniel R Knighton et al. "Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase". In: *Science* 253.5018 (1991), pp. 407–414.

[34] Henrik Möbitz. "The ABC of protein kinase conformations Proteins and proteomics". In: (2015).

[35] Morgan Huse and John Kuriyan. "The conformational plasticity of protein kinases". In: *Cell* 109.3 (2002), pp. 275–282.

[36] Mohammad Azam et al. "Activation of tyrosine kinases by mutation of the gatekeeper threonine". In: *Nature structural & molecular biology* 15.10 (2008), p. 1109.

[37] Steven K Hanks and Anne Marie Quinn. "[2] Protein kinase catalytic domain sequence database: Identification of conserved features of primary structure and classification of family members". In: *Methods in enzymology*. Vol. 200. Elsevier, 1991, pp. 38–62.

[38] Steven K Hanks and Tony Hunter. "The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification 1". In: *The FASEB journal* 9.8 (1995), pp. 576–596.

[39] Harshani R Lawrence et al. "Development of o-chlorophenyl substituted pyrimidines as exceptionally potent aurora kinase inhibitors". In: *Journal of medicinal chemistry* 55.17 (2012), pp. 7392–7416.

[40] LLC Schrödinger. "The PyMOL Molecular Graphics System, Version 1.8". Nov. 2015.

[41] Rik K Wierenga and Wim GJ Hol. "Predicted nucleotide-binding properties of p21 protein and its cancer-associated variant". In: *Nature* 302.5911 (1983), pp. 842–844.

[42] Tony Hunter and Jonathan A Cooper. "Protein-tyrosine kinases". In: *Annual review of biochemistry* 54.1 (1985), pp. 897–930.

[43] Owen N Witte, Asim Dasgupta, and David Baltimore. "Abelson murine leukaemia virus protein is phosphorylated in vitro to form phosphotyrosine". In: *Nature* 283.5750 (1980), pp. 826–831.

[44] ND Richert et al. "Characterization of an immune complex kinase in immunoprecipitates of avian sarcoma virus-transformed fibroblasts." In: *Journal of Virology* 31.3 (1979), pp. 696–706.

[45] TW Wong and Allan R Goldberg. "Purification and characterization of the major species of tyrosine protein kinase in rat liver." In: *Journal of Biological Chemistry* 259.13 (1984), pp. 8505–8512.

[46] MP Kamps and BARTHOLOMEW M Sefton. "Neither arginine nor histidine can carry out the function of lysine-295 in the ATP-binding site of p60src." In: *Molecular and cellular biology* 6.3 (1986), pp. 751–757.

[47] Stevan R Hubbard. "Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog". In: *The EMBO journal* 16.18 (1997), pp. 5572–5581.

[48] Alexandr P Kornev et al. "Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism". In: *Proceedings of the National Academy of Sciences* 103.47 (2006), pp. 17783–17788.

[49] Alexandr P Kornev and Susan S Taylor. "Defining the conserved internal architecture of a protein kinase". In: *Biochimica Et Biophysica Acta (BBA)-Proteins and Proteomics* 1804.3 (2010), pp. 440–444.

[50] Martin EM Noble, Jane A Endicott, and Louise N Johnson. "Protein kinase inhibitors: insights into drug design from structure". In: *Science* 303.5665 (2004), pp. 1800–1805.

[51] Liang Tong et al. "A highly specific inhibitor of human p38 MAP kinase binds in the ATP pocket". In: *Nature structural biology* 4.4 (1997), pp. 311–316.

[52] Fabio Zuccotto et al. "Through the "gatekeeper door": exploiting the active kinase conformation". In: *Journal of medicinal chemistry* 53.7 (2010), pp. 2681–2694.

[53] Christos G Zervas and Nicholas H Brown. "Integrin adhesion: when is a kinase a kinase?" In: *Current biology* 12.10 (2002), R350–R351.

[54] Deborah K Morrison. "KSR: a MAPK scaffold of the Ras pathway?" In: *Journal of cell science* 114.9 (2001), pp. 1609–1612.

[55] Michael Kroiher, Michael A Miller, and Robert E Steele. "Deceiving appearances: signaling by "dead" and "fractured" receptor protein-tyrosine kinases". In: *Bioessays* 23.1 (2001), pp. 69–76.

[56] Gerard Manning et al. "The protein kinase complement of the human genome". In: *Science* 298.5600 (2002), pp. 1912–1934.

[57] Sarah A Teichmann, Jong Park, and Cyrus Chothia. "Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements". In: *Proceedings of the National Academy of Sciences* 95.25 (1998), pp. 14658–14663.

[58] Mark Gerstein. "How representative are the known structures of the proteins in a complete genome? A comprehensive structural census". In: *Folding and Design* 3.6 (1998), pp. 497–512.

[59] Cyrus Chothia. "One thousand families for the molecular biologist". In: *Nature* 357.6379 (1992), pp. 543–544.

[60] Yuri I Wolf, Nick V Grishin, and Eugene V Koonin. "Estimating the number of protein folds and families from complete genome data". In: *Journal of molecular biology* 299.4 (2000), pp. 897–905.

[61] Gordana Apic, Julian Gough, and Sarah A Teichmann. "An insight into domain combinations". In: *Bioinformatics* 17.suppl_1 (2001), S83–S89.

[62] Arne Müller, Robert M MacCallum, and Michael JE Sternberg. "Structural characterization of the human proteome". In: *Genome research* 12.11 (2002), pp. 1625–1641.

[63] Michael Lynch and John S Conery. "The evolutionary fate and consequences of duplicate genes". In: *science* 290.5494 (2000), pp. 1151–1155.

[64] Christine Vogel et al. "Structure, function and evolution of multidomain proteins". In: *Current opinion in structural biology* 14.2 (2004), pp. 208–216.

[65] Gordana Apic, Wolfgang Huber, and Sarah A Teichmann. "Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination". In: *Journal of structural and functional genomics* 4.2-3 (2003), pp. 67–78.

[66] Hedi Hegyi and M Gerstein. "Annotation transfer for genomics: measuring functional divergence in multi-domain proteins". In: *Genome research* 11.10 (2001), pp. 1632–1640.

[67] Matthew Bashton and Cyrus Chothia. "The geometry of domain combination in proteins". In: *Journal of molecular biology* 315.4 (2002), pp. 927–939.

[68] Gordana Apic, Julian Gough, and Sarah A Teichmann. "Domain combinations in archaeal, eubacterial and eukaryotic proteomes". In: *Journal of molecular biology* 310.2 (2001), pp. 311–325.

[69] Elena Papaleo et al. "The role of protein loops and linkers in conformational dynamics and allostery". In: *Chemical reviews* 116.11 (2016), pp. 6391–6423.

[70] Rajesh S Gokhale and Chaitan Khosla. "Role of linkers in communication between protein modules". In: *Current opinion in chemical biology* 4.1 (2000), pp. 22–27.

[71] David Jakubec et al. "Widespread evolutionary crosstalk among protein domains in the context of multi-domain proteins". In: *Plos one* 13.8 (2018), e0203085.

[72] Robert G Smock et al. "An interdomain sector mediating allostery in Hsp70 molecular chaperones". In: *Molecular systems biology* 6.1 (2010), p. 414.

[73] Richard A George and Jaap Heringa. "An analysis of protein domain linkers: their classification and role in protein folding". In: *Protein Engineering, Design and Selection* 15.11 (2002), pp. 871–879.

[74] Willy Wriggers, Sugoto Chakravarty, and Patricia A Jennings. “Control of protein functional dynamics by peptide linkers”. In: *Peptide Science: Original Research on Biomolecules* 80.6 (2005), pp. 736–746.

[75] CD Bottema et al. “Missense mutations and evolutionary conservation of amino acids: evidence that many of the amino acids in factor IX function as" spacer" elements.” In: *American journal of human genetics* 49.4 (1991), p. 820.

[76] Clifford R Robinson and Robert T Sauer. “Optimizing the stability of single-chain proteins by linker length and composition mutagenesis”. In: *Proceedings of the National Academy of Sciences* 95.11 (1998), pp. 5929–5934.

[77] Daniel Brüne, Miguel Andrade-Navarro, and Pablo Mier. “Proteome-wide comparison between the amino acid composition of domains and linkers”. In: *BMC research notes* 11.1 (2018), p. 117.

[78] Peter M Colman et al. “Structure of the human antibody molecule Kol (immunoglobulin G1): an electron density map at 5 Å resolution”. In: *Journal of molecular biology* 100.3 (1976), pp. 257–278.

[79] FK Winkler et al. “Tomato bushy stunt virus at 5.5-Å resolution”. In: *Nature* 265.5594 (1977), pp. 509–513.

[80] Buyong Ma et al. “Dynamic allostery: linkers are not merely flexible”. In: *Structure* 19.7 (2011), pp. 907–917.

[81] Matthew A Young et al. “Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation”. In: *Cell* 105.1 (2001), pp. 115–126.

[82] Sarah Rice et al. “A structural change in the kinesin motor protein that drives motility”. In: *Nature* 402.6763 (1999), pp. 778–784.

[83] Steven S Rosenfeld, Geraldine M Jefferson, and Peter H King. “ATP reorients the neck linker of kinesin in two sequential steps”. In: *Journal of Biological Chemistry* 276.43 (2001), pp. 40167–40174.

[84] Ahmad S Khalil et al. “Kinesin’s cover-neck bundle folds forward to generate force”. In: *Proceedings of the National Academy of Sciences* 105.49 (2008), pp. 19247–19252.

[85] Hans C van Leeuwen et al. “Linker length and composition influence the flexibility of Oct-1 DNA binding”. In: *The EMBO journal* 16.8 (1997), pp. 2043–2053.

[86] Amol Arunrao Pohane, Neelam Devidas Patidar, and Vikas Jain. “Modulation of domain–domain interaction and protein function by a charged linker: A case study of mycobacteriophage D29 endolysin”. In: *FEBS letters* 589.6 (2015), pp. 695–701.

[87] Bartosz Różycki et al. “The length but not the sequence of peptide linker modules exerts the primary influence on the conformations of protein domains in cellulosome multi-enzyme complexes”. In: *Physical Chemistry Chemical Physics* 19.32 (2017), pp. 21414–21425.

[88] Shankar Shastry and William O Hancock. "Neck linker length determines the degree of processivity in kinesin-1 and kinesin-2 motors". In: *Current Biology* 20.10 (2010), pp. 939–943.

[89] Maximilian Klement et al. "Effect of linker flexibility and length on the functionality of a cytotoxic engineered antibody fragment". In: *Journal of Biotechnology* 199 (2015), pp. 90–97.

[90] Mitsuo Ikebe et al. "A hinge at the central helix of the regulatory light chain of myosin is critical for phosphorylation-dependent regulation of smooth muscle myosin motor activity". In: *Journal of Biological Chemistry* 273.28 (1998), pp. 17702–17707.

[91] Ryan B Case et al. "Role of the kinesin neck linker and catalytic core in microtubule-based motility". In: *Current Biology* 10.3 (2000), pp. 157–160.

[92] Venkatesh Hariharan and William O Hancock. "Insights into the mechanical properties of the kinesin neck linker domain from sequence analysis and molecular dynamics simulations". In: *Cellular and molecular bioengineering* 2.2 (2009), pp. 177–189.

[93] Gergő Gógl et al. "Disordered protein kinase regions in regulation of kinase domain cores". In: *Trends in biochemical sciences* 44.4 (2019), pp. 300–311.

[94] Michael Ashburner et al. "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), pp. 25–29.

[95] Gene Ontology Consortium. "The gene ontology resource: 20 years and still GOing strong". In: *Nucleic acids research* 47.D1 (2019), pp. D330–D338.

[96] Edwin C Webb et al. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Ed. 6. Academic Press, 1992.

[97] Lisa Jeske et al. "BRENDA in 2019: a European ELIXIR core data resource". In: *Nucleic acids research* 47.D1 (2019), pp. D542–D549.

[98] José Juan Almagro Armenteros et al. "DeepLoc: prediction of protein subcellular localization using deep learning". In: *Bioinformatics* 33.21 (2017), pp. 3387–3395.

[99] Roderick A Capaldi and Garret Vanderkooi. "The low polarity of many membrane proteins". In: *Proceedings of the National Academy of Sciences* 69.4 (1972), pp. 930–932.

[100] Gunnar von HEIJNE and Ylva GAVEL. "Topogenic signals in integral membrane proteins". In: *European Journal of Biochemistry* 174.4 (1988), pp. 671–678.

[101] Gabor E Tusnady and Istvan Simon. "Principles governing amino acid composition of integral membrane proteins: application to topology prediction". In: *Journal of molecular biology* 283.2 (1998), pp. 489–506.

[102] Jack Kyte and Russell F Doolittle. “A simple method for displaying the hydropathic character of a protein”. In: *Journal of molecular biology* 157.1 (1982), pp. 105–132.

[103] Elisabeth Gasteiger et al. “Protein identification and analysis tools on the ExPASy server”. In: *The proteomics protocols handbook*. Springer, 2005, pp. 571–607.

[104] Kenta Nakai, Akinori Kidera, and Minoru Kanehisa. “Cluster analysis of amino acid indices for prediction of protein structure and function”. In: *Protein Engineering, Design and Selection* 2.2 (1988), pp. 93–100.

[105] Kentaro Tomii and Minoru Kanehisa. “Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins”. In: *Protein Engineering, Design and Selection* 9.1 (1996), pp. 27–36.

[106] Shuichi Kawashima, Hiroyuki Ogata, and Minoru Kanehisa. “AAindex: amino acid index database”. In: *Nucleic acids research* 27.1 (1999), pp. 368–369.

[107] Shuichi Kawashima and Minoru Kanehisa. “AAindex: amino acid index database”. In: *Nucleic acids research* 28.1 (2000), pp. 374–374.

[108] Shuichi Kawashima et al. “AAindex: amino acid index database, progress report 2008”. In: *Nucleic acids research* 36.suppl_1 (2007), pp. D202–D205.

[109] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[110] John D Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing in science & engineering* 9.3 (2007), pp. 90–95.

[111] Teresa Milano et al. “Structural properties of the linkers connecting the N-and C-terminal domains in the MocR bacterial transcriptional regulators”. In: *Biochimie open* 3 (2016), pp. 8–18.

[112] Dominico Vigil et al. “Conformational differences among solution structures of the type Iα, IIα and IIβ protein kinase A regulatory subunit homodimers: role of the linker regions”. In: *Journal of molecular biology* 337.5 (2004), pp. 1183–1194.

[113] Miles A Pufall and Barbara J Graves. “Autoinhibitory domains: modular effectors of cellular regulation”. In: *Annual review of cell and developmental biology* 18.1 (2002), pp. 421–462.

[114] Jin Liu and Ruth Nussinov. “Molecular dynamics reveal the essential role of linker motions in the function of Cullin–RING E3 ligases”. In: *Journal of molecular biology* 396.5 (2010), pp. 1508–1523.

[115] Kedra Cyrus et al. “Impact of linker length on the activity of PROTACs”. In: *Molecular BioSystems* 7.2 (2011), pp. 359–364.

[116] Charalampos G Kalodimos. “NMR reveals novel mechanisms of protein activity regulation”. In: *Protein Science* 20.5 (2011), pp. 773–782.

[117] Jeffrey S Tanaka. "" How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables". In: *Child development* (1987), pp. 134–146.

[118] Jianping Hua et al. "Optimal number of features as a function of sample size for various classification rules". In: *Bioinformatics* 21.8 (2005), pp. 1509–1515.

[119] Pascale Gaudet and Christophe Dessimoz. "Gene ontology: pitfalls, biases, and remedies". In: *The Gene Ontology Handbook*. Humana Press, New York, NY, 2017, pp. 189–205.

[120] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[121] Claudia Hass et al. "The response regulator 2 mediates ethylene signalling and hormone signal integration in Arabidopsis". In: *The EMBO journal* 23.16 (2004), pp. 3290–3302.

[122] Michael G Mason et al. "Multiple type-B response regulators mediate cytokinin signal transduction in Arabidopsis". In: *The Plant Cell* 17.11 (2005), pp. 3007–3018.

[123] Andrew G McDonald and Keith F Tipton. "Fifty-five years of enzyme classification: advances and difficulties". In: *The FEBS journal* 281.2 (2014), pp. 583–592.

[124] Bo Li et al. "Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins". In: *arXiv preprint arXiv:1001.0958* (2010).

[125] Chenguang Zhao and Zheng Wang. "GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms". In: *Scientific reports* 8.1 (2018), pp. 1–10.