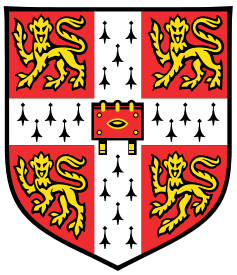
The evolution of protein kinase specificity



David Bradley

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College

September 2018

I would like to dedicate this thesis to my family

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words as defined by the Biology Degree Committee.

David Bradley
September 2018

Acknowledgements

There are several people who have helped me during the course of my PhD.

First and foremost I would like to thank my supervisor, Pedro Beltrao. I first arrived in the group as a humble Master's student on rotation. Pedro then took me on for the PhD and since then has supported me throughout my studies. Under his tutelage I have developed both as a scientist and as a person, and for this I am grateful.

Second I would like to thank Romain Studer. Romain supervised me for my rotation project at EMBL-EBI and introduced me to the wonderful world of protein evolution. Much of what I learnt under his supervision has proven invaluable for the PhD.

I would like to thank also the enormously talented postdocs of the Beltrao group past and present: Inigo Barrio, Bede Busby, Marco Galardini, Haruna Imamura, Brandon Invergo, Danish Memon, David Ochoa, Rahuman Sherriff, Romain Studer, and Cris Viéitez. I have been in awe of them since the very start and still am to this day. As clichéd as it sounds, I feel that I am a better person having known them.

To my PhD comrades also – Alistair Dunham, Claudia Hernandez, Abel Sousa, Marta Strumillo, and Omar Wagih – a PhD is never easy but I consider myself very lucky to have been able to study in the company of such brilliant peers.

I am grateful also to Thomas Lemberger and Maria Polychronidou for their excellent supervision during my internship at EMBO. I thank Pedro again for suggesting the internship, and also Areeb Jawed, Bede, and Cris for keeping me company upon my sojourn in Deutschland. I want to thank Cris Viéitez especially during this period for her perseverance with the very tricky kinase activity assays.

There are many others who have enriched my time at Cambridge. In particular I have in mind my fellow predocs at the EBI, the numerous interns and visitors of the Beltrao group, and my former students at Downing College. You are too many to name in person but I would like to thank you all.

Finally, I owe a debt of gratitude to my family and most of all to my mother and father. This thesis is dedicated to them.

Abstract

Protein phosphorylation represents one of the most important post-translational modifications (PTMs) for cell signalling, and is is catalysed by a group of enzymes called protein kinases. Through this activity they serve as key regulators of almost all cellular processes. This is achieved at any time by a network of different kinases that are transiently active. The fidelity of cell systems control therefore requires that each kinase targets only a restricted set of substrates. This specificity is achieved partly by contextual factors that separate kinases spatially and temporally, but also by sequence features that are encoded in the kinase domain itself.

For this thesis I focus on elements of kinase specificity that are encoded in the the active site of the enzyme. During these investigations I have tried to address three main questions: 1) How is specificity for residues surrounding the phosphorylation site determined in the kinase? 2) How did these specificities evolve? and 3) To what extent does kinase evolution correlate with the evolution of its substrates?

First, I developed a sequence-based method for the automated detection of kinase specificity determining residues (SDRs). The putative determinants were then rationalised using available structural data, and in two specific cases were validated experimentally. I also used mutation data from The Cancer Genome Atlas (TCGA) to demonstrate that kinase SDRs are often targeted during cancer.

Second, a global analysis of SDR evolution was performed for kinases following gene duplication and speciation, revealing that SDRs often diverge between paralogues but not between orthologues. This global analysis is followed by a detailed case study of G-protein coupled receptor kinase (GRKs) evolution using ancestral sequence reconstructions.

Third, I inferred global substrate preferences in a taxonomically broad range of species using phosphoproteome data. I then related the evolution of substrate motif sequences to that of their cognate effector kinases where possible. The results strongly suggest that many of the motifs emerged in a universal eukaryotic ancestor.

I finish by summarising the major findings of this doctoral research, which to my knowledge represents the most comprehensive analysis to date of protein kinase specificity and its evolution.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

3DID database of Three-dimensional Interacting Domains

AC Ancestral conservation

AGC Protein kinase **A**, **G**, **C**

AP Affinity propagation

APK Atypical protein kinase

ASCH Activation segment C-terminal helix

ASR Ancestral Sequence Reconstruction

ATM/ATR Ataxia telangiectasia mutated/ and Rad53 related

ATP Adenosine triphosphate

AUC Area under the ROC curve

BADASP Burst after duplication and ancestral sequence prediction

BARK Beta adrenergic receptor kinase

BLAST Basic Local Alignment Search Tool

CAMK Calcium/calmodulin-dependent protein kinase

cAMP cyclic-Adenosine monophosphate

CEASAR Connecting Enzymes And Substrates at Amino acid Resolution

CHARMM Chemistry at Harvard Macromolecular Mechanics

CK1/2 Casein kinase 1/2

CDK Cyclin dependent kinase

CMGC CDK, MAPK, GSK3, CK2

DAG Diacylglycerol

DSK Dual specificity kinase

ELK ePK-like kinase

EM Energy minimisation

ePK eukaryotic protein kinase

GPCR G protein-coupled receptor

GRK G protein-coupled Receptor Kinase

GSK Glycogen synthase kinase

GUI Graphical user interface

HGT Horizontal gene transfer

HMM Hidden Markov model

HPRD Human protein reference database

IMAC Immobilised metal affinity chromatography

LC-MS Liquid chromatography mass spectrometry

LOOCV Leave-one-out cross-validation

MAFFT Multiple sequence alignment based on fast Fourier transform

MAPK Mitogen-activated protein kinase

MD Molecular dynamics

ML Maximum likelihood

MS Mass spectrometry

MSA Multiple sequence alignment

OPL Orientated peptide library

PCA Principal components analysis

PDB Protein data bank

PDK Phosphoinositide-dependent kinase

Phospho-ELM Phospho- eukaryotic linear motif

PK Protein kinase

PKA Protein kinase A

PKC Protein kinase C

PLK Polo-like kinase

PPM Position probability matrix

ProtCID Protein common interface database

PSPL Positional scanning peptide library

PSSM Position-specific scoring matrix

PTM Post-translational modification

PWM Position weight matrix

RAxML Randomized Axelerated Maximum Likelihood

RC Recent conservation

REST-API Representational state transfer applicatioan programming interface

ROC Receiver-operator characteristic

RSK Ribosomal S6 kinase

SDR Specificity determining residue

SERIOHL-KILR Serine-orientated human library-kinase library reactions

SH2 Src Homology 2

SIFTS Structure Integration with Function, Taxonomy, and Sequence

SOM Self-Organising Maps

SPEER Specificity prediction using amino acids' Properties, Entropy, and Evolution Rate

SRPK Serine-arginine (S-R) protein kinase

STRING Search Tool for the Retrieval of Interacting Genes/proteins

TCGA The Cancer Genome Atlas

TK Tyrosine kinase

TKL Tyrosine kinase-like

Chapter 1

Introduction

1.1 The protein kinase superfamily

1.1.1 Protein kinase function

Protein kinases are enzymes that catalyse the phosphorylation of other proteins (Figure 1.1).

Protein phosphorylation was discovered in the early 20th century and later realised (in the 1950s) to play a role in the activation of enzymes such as glycogen phopsphorylase (Cohen, 2002; Fischer and Krebs, 1955; Krebs, 1983; Krebs and Fischer, 1956; Levene and Alsberg, 1906; Lipmann and Levene, 1932). Now it is known to modulate protein function in many other ways. Specifically, it can stabilise or destabilise the target, promote or inhibit protein-protein interactions, and also direct the target towards specific subcellular localisations (Beltrao et al., 2013; Holt, 2012; Pawson and Scott, 2005).

Phosphorylase kinase and cAMP-dependent protein kinase of the glycogenolytic pathway were the first to be characterised biochemically (Fischer and Krebs, 1955; Krebs and Fischer, 1956; Taylor and Kornev, 2011; Walsh et al., 1968). Subsequent research demonstrated the role of kinases in many other metabolic pathways and cellular processes. The discovery in particular of kinases sensitive to second messengers such as Ca^{2+} and diacylglycerol (DAG) implied a pivotal role for kinases in cellular signal transduction (Cohen, 2002; Kishimoto et al., 1980; Pawson and Scott, 2005). This role is perhaps best represented by MAPK (mitogen-activated protein kinase) pathways – tri-partite kinase cascades in which the upstream kinase (MAPKKK) activates a target (MAPKK) that in turn activates another downstream kinase (MAPK). The discovery of MAPK cascades in the 1990s therefore highlighted the ability of kinases to serve as both catalysts and substrates for phosphorylation, thus enabling the formation of kinase-substrate networks capable of signal amplification (Cohen, 2002; Pearson et al., 2001; Roskoski, 2015).

It is now clear that protein kinases are required for almost all cellular pathways, including fundamental processes such as apoptosis and cell cycle control. Their ubiquity is best explained by the properties of the PO_4^{3-} group itself, in that it is charged, reactive, and labile (Hunter, 2012). Its addition to proteins can also be catalytically reversed by complementary phosphatase enzymes. Phosphate modification is therefore a highly effective label for the regulation of dynamic processes.

The central importance of protein kinases to cell biology is underlined by the fact that their dysregulation is responsible for several different diseases. Such 'kinasopathies' include achondroplasia, Parkinson's disease, and many different cancers (Izarzugaza et al., 2012; Lahiry et al., 2010; Stenberg et al., 1999). This is one of the reasons why kinases have long been the subject of intensive research efforts. In 2005 for example it was reported that they account for around 30 percent of all spending on drug development (Knight and Shokat, 2005).

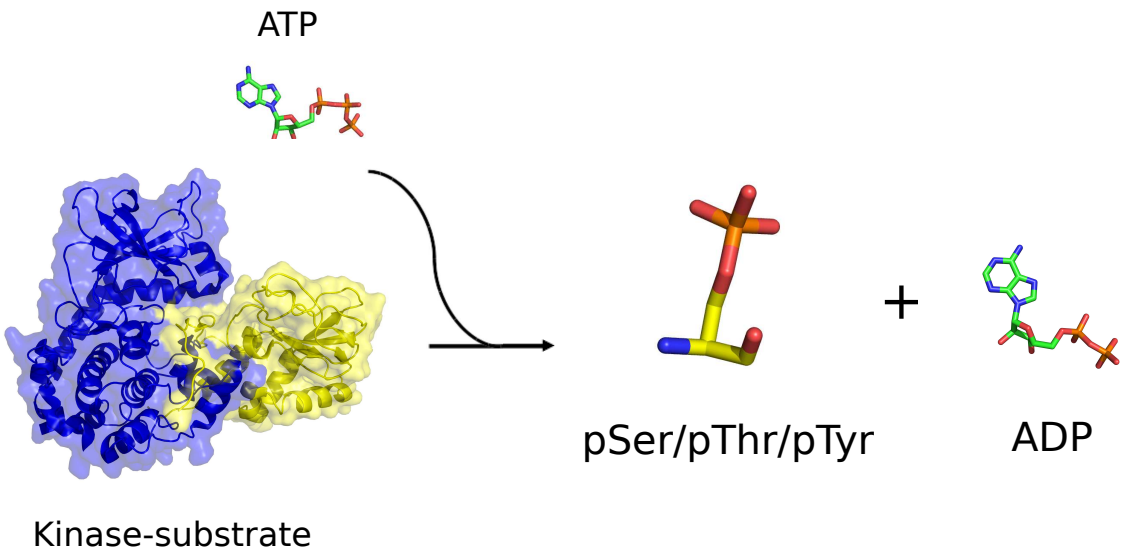


Fig. 1.1 Protein kinases catalyse the transfer of the ATP γ-phosphate to the phosphoacceptor serine, threonine, or tyrosine

1.1.2 Protein kinase sequence and structure

Note: protein kinase residues in this thesis are numbered according to their position in the Pfam protein kinase domain (PF00069). This is discussed further in the subsection below ('Numbering of protein kinase residues').

Most eukaryotic protein kinases contain at least one catalytic kinase domain approximately 260 amino acids in length (Figure 1.2). Many also contain additional protein domains

that are important for their function. In some cases these functions are linked directly to that of the kinase catalytic domain. Kinases of the Polo/PLK *Family* for example feature a 'Polo box' domain that will bind phospho-serine/threonines and thus recruit the whole molecule to previously phosphorylated substrates (Archambault and Glover, 2009; Park et al., 2010).

The kinase domain itself was first sequenced in 1981 by Edman degradation (Shoji et al., 1981). The subsequent sequencing of other kinase domains enabled the identification of highly conserved residues with presumed importance for kinase function. By 1995 there were sufficient sequences (~400) for the detailed analysis of kinase sequence variation by Hanks and Hunter (Hanks and Hunter, 1995), who divided the domain into 12 smaller subdomains (I-XII). At this time a kinase-peptide structural model had been generated for the protein kinase A catalytic subunit in complex with a pseudosubstrate inhibitor (Knighton et al., 1991). Examination of these structures therefore allowed putative roles to be assigned to the 12 subdomains and associated residues.

Kinase subdomains I, II, III, and IV comprise the smaller N-terminal lobe, which is involved primarily in the binding and orientation of adenosine triphosphate (ATP) (Hanks and Hunter, 1995; Roskoski, 2007). This lobe is composed mainly of β sheets but also contains the αC helix. The αC helix is partially exposed at the side of the substrate-binding cleft and is reorientated upon kinase activation to promote substrate binding (Beenstock et al., 2016; McSkimming et al., 2017). The C helix also features an invariant E (position 48) that interacts with the invariant K (position 30) of subdomain II. Both residues contribute to the stabilisation of the ATP α- and β- phosphates (Endicott et al., 2012; Kornev and Taylor, 2015). The two other invariant positions (8 and 10) of the N-terminal lobe form the glycine-rich repeat (**G**-x-**G**-x-x-G) that is important for ATP binding and catalysis (Hemmer et al., 1997).

The larger C-terminal lobe is mainly α helical and comprises subdomains VI to XII (Hanks and Hunter, 1995). The catalytic loop extends from the αE helix and contains the xR**D**xKxx**N** motif (Taylor and Kornev, 2011). The invariant D at position 123 aligns structurally with the substrate phosphoacceptor and serves as a proton acceptor from the substrate -OH group (Bossemeyer, 1995). This function is assisted by the K at position 125 that forms ionic interactions with the γ-phosphate of ATP (Knighton et al., 1991). The invariant N at position 128 also contributes to catalysis by stabilising D123 and chelating the secondary Mg^{2+} ion found in the nucleotide binding site (Hanks and Hunter, 1995; Johnson et al., 1996).

The invariant D of the '**D**FG' motif is present on a contiguous loop. This residue chelates the Mg^{2+} ion between the ATP β- and γ- phosphates. Active and inactive kinase structures can be distinguished by 'in' and 'out' conformations of the DFG motif, respectively, which describes whether or not the motif phenylalanine is buried in a hydrophobic pocket between

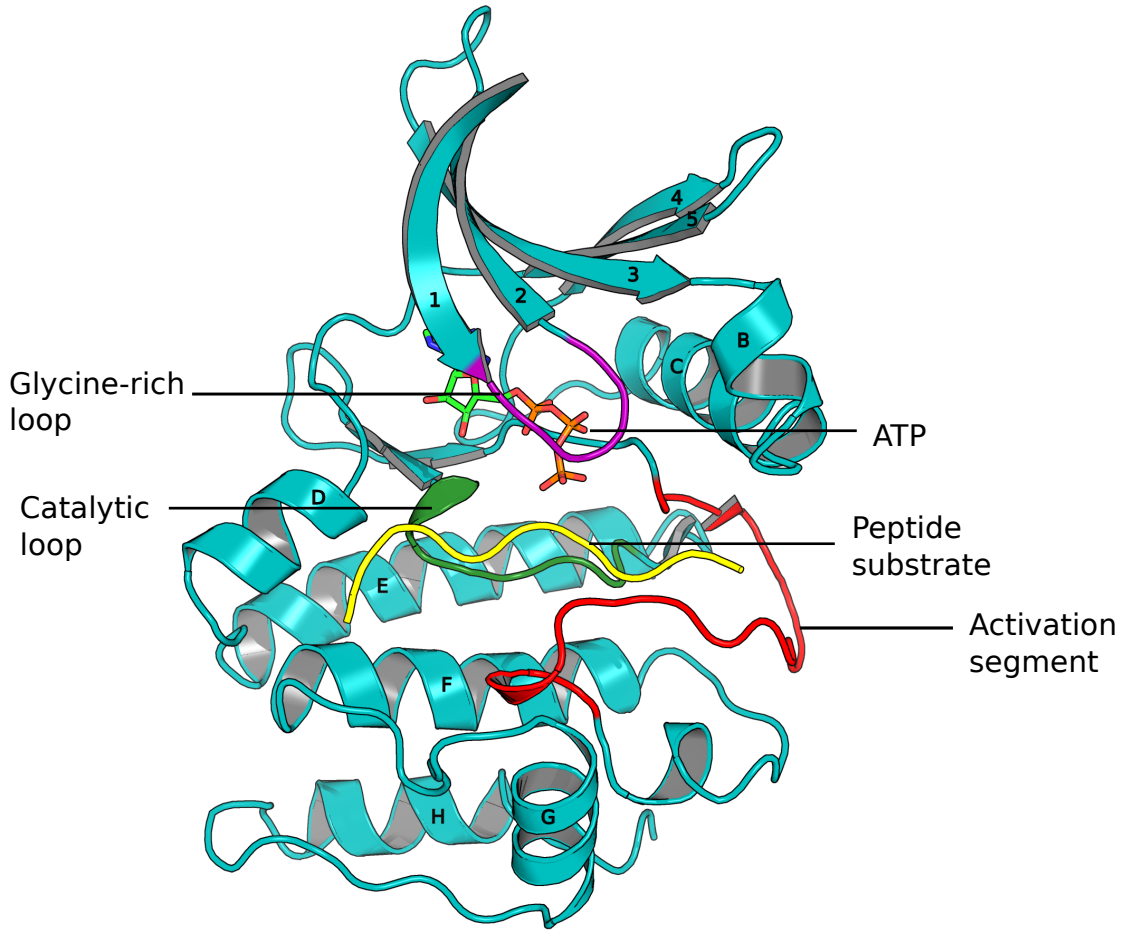


Fig. 1.2 The protein kinase domain is ~260 residues in length and consists of an N-terminal lobe (top) and a larger C-terminal lobe (bottom). The ATP molecule binds in the N-terminal lobe whereas the substrate (yellow) binds in the active site between the two lobes. Important structural features are labelled here and discussed further in the main text. *PDB: 1ATP*

the two kinase lobes (Xu et al., 2011). The highly conserved APE motif is C-terminal to the DFG motif, and both motifs together delimit an intervening sequence referred to as the 'activation segment'. In inactive kinases, a looped sequence ('activation loop') of the activation segment obstructs the kinase active site and thus prevents substrate phosphorylation (Nolen et al., 2004). Upon phosphorylation of the activation loop – either by autophosphorylation or an upstream kinase – the loop becomes bound to the catalytic core via ionic bonds to the arginine side chain of the HRD motif and the active site becomes accessible to substrates (Johnson et al., 1996). The activation segment is therefore highly important for the regulation of kinase activity, but also for substrate specificity as some activation loop residues contact the substrate C-terminal to the phosphoacceptor. The activation segment P+1 loop (approximately positions 159 to 164) for example is an important determinant of kinase 'P+1' specificity (Nolen et al., 2004).

The increased availability of data on kinase sequence and structure has proven these features to be largely universal across the protein kinase superfamily. Most research efforts are therefore now directed towards an improved understanding of *Family*- or *Subfamily*-specific modes of kinase regulation and substrate specificity (for peptides and nucleoside analogues). There have however been some advances in the global understanding of the kinase domain since the Hanks and Hunter review of 1995. The structural alignment of many kinase models for example enabled the identification of core catalytic and regulatory 'spine' residues that are spatially conserved and anchor functionally related residues (Kornev et al., 2008). An alignment-based analysis of thousands of kinase domain sequences has also suggested the existence of distinct 'catalytic', 'specificity', and 'regulatory' sectors, which are defined as networks of co-evolving residues within the kinase domain (Creixell et al., 2017).

Numbering of protein kinase residues

The kinase residues in this thesis are numbered according to domain positions in the protein kinase *Pfam* domain (PF00069), which represents a profile hidden Markov model (HMM) constructed from a sequence alignment of protein kinases (El-Gebali et al., 2019). This allows residues from several different family members to be mapped to a common reference, which is an important consideration for most comparative structural analyses. An alternative approach is to map the residues to a single reference protein that is considered representative or prototypical of the family in some way. In the kinases for example, kinase residues are often mapped to the first solved kinase structure (PKA catalytic subunit in mouse, PDB: *1atp*) (Knighton et al., 1991), with in-text references to both the native residue numbering and the PKA mapping (Mok et al., 2010). Other studies have made use of secondary structure elements of the family domain for residue mapping. This involves numbering the secondary

structure elements first (i.e. the α helices and β sheets), and then numbering the residues according to their position within the helix or sheet, as was done for comparative structural analyses of the Gα and arrestin families (Flock et al., 2015; Sente et al., 2018).

1.1.3 Protein kinase classification and evolution

Early evolution of the superfamily

Eukaryotic protein kinases are structurally distinct from many of the protein kinases found in prokaryotes. Unlike in eukaryotes, phosphate-based signal transduction can be mediated by different molecular systems in prokaryotes. The most well-characterised of these is the bacterial two-component system in which the sensor kinase first autophosphorylates on a histidine residue and then transfers the phosphate group to an aspartate residue on the response regulator (Stock et al., 2000). Prokaryotic Ser/Thr kinases also exist in which the kinase and phosphatase activity is encoded in the same molecule, such as for the bacterial isocitrate dehydrogenase kinase/phosphatase enzyme (Laporte et al., 1989). Some prokaryotic kinases also function in distinct phosphotransferase systems such as is the case for the phosphoenolpyruvate-protein transferase enzyme involved in sugar uptake (Kotrba et al., 2001). Notably, in all three cases the protein kinase lacks clear sequence homology to the eukaryotic protein kinase domain (ePK) domain (Cozzone, 1993; Pereira et al., 2011).

Eukaryotic PK-like protein kinases do however exist in some prokaryotes, as was first demonstrated with the sequencing of the pkn1 gene in *Myxococcus xanthus* in 1991 (Muñoz-Dorado et al., 1991; Pereira et al., 2011). Advances in sequencing technology later revealed the existence of such ePK-like kinases (ELKs) in many different prokaryotic species, including in the Archaea. Their sequence similarity to the ePK domain is typically low (7%–17%), although comparisons between crystal structures reveals a higher degree of conservation (Kannan et al., 2007b). Notably, a metagenomics sequencing project of marine prokaryotes in 2007 suggested that ELKs may be as prevalent as the histidine kinases in prokaryotes (Kannan et al., 2007b). This prevalence suggests either that the origin of the kinase superfamily predates the emergence of the eukaryotes, or that the horizontal gene transfer (HGT) of kinases from eukaryotes to prokaryotes occurred early after the separation of the archaea, bacteria, and eukaryotes (Kennelly, 2002; Scheeff and Bourne, 2005). A more recent analysis however argues strongly against the possibility of HGT, thus placing the origin of Ser/Thr kinases in the last universal common ancestor of life (Stancik et al., 2018).

Confusingly, so-called ePK-like kinases (ELKs) also exist in eukaryotes in the form of small molecule kinases (choline kinase, aminoglycoside phosphotransferase, etc.) without sequence homology to the ePK domain but with a shared bilobal fold (Scheeff and Bourne,

2005). The atypical kinases (APKs) also are more divergent than both ePKs and ELKs (Figure 1.3) and are thought to represent an intermediate fold between kinases and other ATPases (Kornev and Taylor, 2015; Oruganty and Kannan, 2012; Oruganty et al., 2016). An integrative analysis of ePK and ELK sequence and structure suggests that the ePKs diverged from the eLKs early during evolution, contrary to previous suggestions of ELKs as a polyphyletic clade that diverged intermittently (Oruganty et al., 2016; Scheeff and Bourne, 2005). Remarkably, this analysis also implies that some protein kinases exist – such as as the channel kinases – that are more closely related to small molecule kinases than other protein kinases (Scheeff and Bourne, 2005).

The word 'kinase' from this point on should be taken to mean 'proteins with a catalytically active ePK domain', unless otherwise stated, as the research component of this thesis concerns ePKs exclusively.

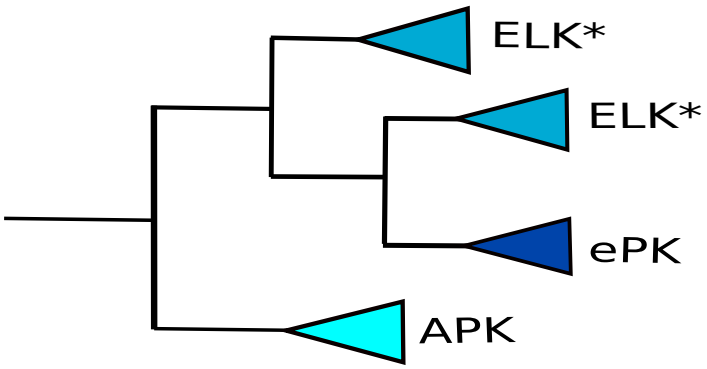


Fig. 1.3 Atypical protein kinases (APKs) are currently believed to serve as an outgroup to the ePKs (kinases with a eukaryotic protein kinase domain) and the ELKs (ePK-like kinases). The ePKs are also believed to have emerged from within the ELKs. *This is a simplified representation of the results given in Oruganty et al., 2016*

Classification of protein kinases

The protein kinase superfamily is unusually large, with around 500 canonical protein kinases in human alone (Manning et al., 2002b). This necessitated the development of a bespoke classification system for the kinases. The most striking partition within the kinase superfamily is between serine/threonine-phosphorylating kinases and tyrosine-phosphorylating kinases. Tyrosine kinases however constitute less than 20% of all human kinases and so account for a small fraction of the total sequence diversity within the superfamily.

The first systematic classification of the protein kinases in 1995 was based upon a global kinase phylogeny generated from a kinase domain alignment (Hanks and Hunter, 1995). The first tier of classification organised the kinases into five different *Groups*: AGC, CAMK, CMGC, tyrosine kinases, and 'Other' kinases that could not easily be categorised. Each *Group* in turn is then defined by the set of related kinase *Families* that constitute the particular clade in the phylogeny. Kinases of a particular *Family* (e.g. the cyclin-dependent kinases) by definition diverged from each other later in evolution, and in many cases also share a common function, specificity, and mode of regulation (Hanks and Hunter, 1995). The third tier of classification refers to kinase *Subfamilies*, which are similar in principle to kinase *Families* but in which the functional and sequence similarity is defined more finely (Hanks and Hunter, 1995). Not all kinase *Families* can be divided into *Subfamilies*.

The near-complete sequencing of the human genome over five years later enabled for the first time an attempted annotation of the human 'kinome' – the total set of protein kinases encoded in the genome. This could be achieved largely by querying genomic sequences with a hidden Markov model (HMM) of the ePK domain. Such queries revealed that kinases account for around 2% of all protein-coding human genes and thus constitute one of the largest protein superfamilies. The original classification of 5 *Groups*, 44 *Families* and 51 *Subfamilies* was thereby extended to 8 *Groups*, 134 *Families*, and 196 *Subfamilies* primarily on the basis of kinase domain sequences. The kinase protein sequence, domain architecture and known functions were also taken into account for the classifications (Manning et al., 2002b).

A brief summary of the 8 major kinase *Groups* is given below:

- AGC group (PK**A**, PK**G**, PK**C**): consists mainly of signalling kinases, including cyclic nucleotide-dependent kinases. They are generally basophilic (i.e. R/K-favouring), with the exception of the GRK and PDK1 *Families*.

- CMGC group (**C**DK, **M**APK, **G**SK3 and **C**asein Kinase II): associated with a diversity of functions, including cell cycle control, stress signalling, and metabolic regulation. Many have a strong preference for proline at the +1 position (i.e. the position directly C-terminal to the phosphoacceptor serine/threonine).

- CAMK group (CAMK1, CAMK2): characterised by calcium- and calmodulin-regulated kinases, although non-calcium regulated kinases are also present. Many have a strong preference for arginine at the -3 position (i.e. three positions N-terminal to the phosphoacceptor serine/threonine)

- STE group (STE7, STE11, STE20): Generally consists of those kinases that exist upstream of MAPK (i.e. MAPKK, MAPKKK, and MAPKKKK) in canonical MAPK

signalling cascades. Some STE kinases however have no reported role in MAPK cascades.

- RGC group: receptor guanylate cyclases – there are only five RGC proteins in human and each is catalytically inactive (i.e. is a pseudokinase). The term 'RGC' refers to the fact that they are receptors and have a 'guanylate cyclase' domain.
- CK1 group (CK1α,CK1δ, CK1ϵ): The CK1 group is highly divergent from most other protein kinases, and they function in a diverse set of cellular processes (cell cycle, transcription, cytoskeleton, etc). There are only 12 CK1 kinases in human
- TK group (Abl, EGFR, Src): Tyrosine kinase group (EDGFR, PDGFR) – kinases with exclusive specificity for tyrosine residues; includes receptor tyrosine kinases (RTKs) and cytoplasmic tyrosine kinases.
- TKL (BRAF, IRAK, LRRK): tyrosine kinase-like group – they exhibit sequence similarity to the tyrosine kinases but are generally specific for Ser/Thr.
- Other kinases (Wee, PLK, Aurora kinase): all kinases with an ePK domain that do not fit into any of the above *Groups*.

Non-human kinomes

The increased availability of genome-wide sequence data since 2002 has enabled similar characterisations for model organisms across the Tree of Life. The kinome of *Saccharomyces cerevisiae* for example was elucidated before the human kinome and revealed the existence of yeast-specific kinase *Subfamilies*, as well as a histidine kinase unrelated to the ePK domain (Hunter and Plowman, 1997). In line with expectation, it is generally the case that model organisms closely related to human have similar kinomes whereas more distantly related species have more divergent kinomes. *Mus musculus* for example was reported to have orthologues for 510 of 518 human kinases (Caenepeel et al., 2004). The more distantly related echinoderm *Stonglyocentrotus purpuratus* however, while having 183 out of the 187 *Subfamilies* found in human, features multiple unique *Families* including the 'Urch' *Family* with 29 members (Bradham et al., 2006). For *Caenorhabditis elegans* and *Drosophila melanogaster*, there are likewise significant overlaps with the human kinome but also idiosyncrasies at the *Family*- and *Subfamily*-level (Manning et al., 2002a; Plowman et al., 1999).

The analysis of more distant species helps to clarify the relationship between kinome evolution and the emergence of the metazoa. The basal metazoan *Amphimedon queens-*

landica and the choanoflagellate *Monosiga brevicollis* for example both feature tyrosine kinases to the exclusion of species outside the filozoa (choanoflagellates and metazoa), and so implicate tyrosine kinases in the evolution of animal multi-cellularity (King et al., 2008; Srivastava et al., 2010). This finding accords well with the known functions of tyrosine kinases in cell signalling, cell migration, and differentiation. Both species however contain more tyrosine kinases than human despite their apparent biological simplicity. In the case of *M. brevicollis*, most TKs have no clear homologue in human, suggesting that tyrosine kinase proliferation occurred independently in the choanoflagellate and metazoan lineages (Manning et al., 2008).

The functional annotation of kinomes for species outside the metazoa and fungi is more difficult because a large proportion of kinases have no homology to experimentally characterised kinases in human or budding yeast. The ciliate *Tetrahymena thermophila* for example contains over 1000 kinases but less than 40% can be placed into a canonical *Group*, *Family* or *Subfamily* (Eisen et al., 2006)(KinBase). Most therefore are of unknown function. To a lesser extent this is true also for species with smaller kinomes. *Plasmodium* for example has 91 kinases and around 21% of them belong to a *Group* (FIKK) that is thought to be unique to the genus (Talevich et al., 2012). Some species are marked also by large-scale expansions of particular kinase *Groups* relative to what is observed in human or *S. cerevisiae*. For *Dictyostelium* and the plants, this is the case for the TKL *Group* (Goldberg et al., 2006; Lehti-Shiu and Shiu, 2012). Since some TKL kinases have demonstrated tyrosine kinase activity in both *Dictyostelium* and plants (Goldberg et al., 2006; Jaillais et al., 2011), it is tempting to speculate that these TKL expansions were at least partially analogous to the emergence and proliferation of the TKs in the filozoa (Goldberg et al., 2006).

Methods for the automated annotation of kinomes

Many of the kinome studies referred to above required significant levels of manual curation. Since then, efforts have been made to develop tools for the automated annotation of a kinome. This requires first that all kinases in a proteome are identified, and second that each identified kinase is correctly classified. The identification of the kinases themselves is usually achieved using profile-based HMMs, which in the context of proteins are statistical models of sequence variation derived from multiple sequence alignments (MSAs) (Eddy, 1996). Profile-based HMMs generally identify sequence homologues with a greater degree of sensitivity and specificity than BLAST does (Park et al., 1998). They are also used for the classification of kinases via the generation of *Group*-based HMMs (AGC, CAMK, CMGC, etc.) from carefully constructed seed alignments. An unclassified kinase for example would be assigned to the *Group* corresponding to the HMM with the strongest similarity to the kinase domain

sequence. Many such profile-based HMMs constructed at the *Group*, *Family*, and *Subfamily* level are provided in the kinase database KinBase (*http://kinase.com/web/current/*). Their use can obviate the need for time-consuming manual kinome curation provided that the underlying genome has been fully sequenced and that the HMMs can be assumed to be accurate. This is less likely to be the case if the examined species is distantly related to the species from which HMM seed alignment was generated.

The *Kinomer* tool developed in 2007 advances this concept by making use of multilevel HMM libraries so that each kinase *Group* is represented by multiple HMMs (Miranda-Saavedra and Barton, 2007). Its development was motivated by the finding that protein homologues can be detected more reliably by using multiple subfamily HMMs rather than a single HMM to represent a given protein family (Brown et al., 2005). The benchmarking of *Kinomer* revealed this to be the case for the protein kinases also as the multilevel HMMs were found to classify protein kinases with a higher degree of sensitivity than was found for BLAST or single-level HMMs. This is likely achieved by the ability of multilevel HMMs to more accurately represent the distinguishing features of each constituent subfamily (Miranda-Saavedra and Barton, 2007).

The *Kinannote* tool developed in 2013 for the same purpose is a more sophisticated method that incorporates BLAST and HMM-based homology searches (Goldberg et al., 2013). *Kinannote* differs from *Kinomer* also in the sense that it attempts kinase classifications at the *Family* and *Subfamily* level also whereas *Kinomer* does not attempt classification beyond the *Group* level. The algorithm can be divided into three phases. In the first phase, the proteome is queried with a single-level kinase HMM using lenient cut-off values; the putative kinases are then scored with a kinase-domain PSSM and also used for a BLAST-based query of the KinBase reference database. In the second stage, low-scoring kinase domains and aPKs are identified on the basis of their sequence homology (via BLAST) to reference kinases in KinBases. Any remaining sequences below the PSSM threshold are then discarded, and kinases above the PSSM threshold but below the HMM threshold are identified as 'twilight hits' but not further classified. Finally, the remaining sequences with the ePK domain are then classified at the *Group*/*Family*/*Subfamily* level on the basis of their homology (via BLAST) to kinases in KinBase with known classifications. The benchmarking of Kinannote suggests that it performs similarly to *Kinomer* in the identification protein kinases, but outperforms *Kinomer* when classifying kinases at the *Group* level, in spite of Kinannote's use of single-level HMMs (Goldberg et al., 2013).

Such automated tools for kinome annotation are advantageous in the sense that they enable the comparison of kinomes between previously uncharacterised species. The application of *Kinomer* to eukaryotic species across the Tree of Life for example demonstrated that

the kinase *Groups* AGC, CAMK, CK1, CMGC, and STE are universal across the eukaryotes and therefore were likely present in the ancestor of all eukaryotes (Miranda-Saavedra and Barton, 2007). Within the metazoa, the results of *Kinomer* queries have also been used to demonstrate a strong correlation between the frequency of tyrosine kinases and of proteins with an SH2 domain (Liu et al., 2011). Similarly, *Kinomer* has also been used to test for associations between particular kinase *Groups* and phosphomotif usage in 18 fungal species (Studer et al., 2016). However, in the majority of cases these tools have been used simply for the study of a single species kinome at a time. A few other publications notwithstanding (Miranda-Saavedra et al., 2012; Talevich et al., 2012), there are surprisingly few studies that attempt to correlate kinome differences with phenotypic differences across several species.

1.2 Protein kinase specificity

1.2.1 Spatial and temporal factors

The 'specificity' of a protein kinase refers to the set of substrates that can be phosphorylated by the kinase under physiological conditions. There are multiple factors that constrain kinase specificity. At a cellular level, the kinase and substrate must be co-expressed and co-localised for them to interact *in vivo*. Most kinases also are not constitutively active, and so the activation period of the kinase must overlap with the expression interval of any putative substrate. More subtle contextual factors could also prevent any meaningful substrate phosphorylation *in vivo*. For example, if a kinase is expressed at low levels and in the presence of a preferred substrate, then the high-affinity substrate is likely to competitively inhibit phosphorylation of low-affinity substrates (Ubersax and Ferrell, 2007). In multicellular organisms, all of the aforementioned requisites must be satisfied for kinases and substrates in the same cell or tissue type.

There are multiple examples of kinases with similar sequences but different specificities owing to differential expression patterns. Protein kinase Cγ for example is similar in sequence and peptide specificity to the other conventional PKCs but is expressed exclusively in the brain and spinal cord (Saito and Shirai, 2002). At the subcellular level, proteins with identical primary sequences can differ in specificity because of spatial segregation. For example, cyclin B1-CDK1 complexes are localised primarily in the cytoplasm or nucleus whereas cyclin B2-CDK1 complexes are targeted towards the Golgi apparatus (Ubersax and Ferrell, 2007). Such subcellular targeting will have additional quantitative effects in the sense that the effective concentration of the kinase and substrate is increased due to com-

partmentalisation, and that there are fewer potential substrates to compete for the active site of the kinase.

The cellular gene regulatory network (GRN) also impacts upon kinase specificity in the sense that the kinase and substrate must be expressed during the same phase of the cell cycle and/or in response to the same stimuli. Notably, a global network analysis of kinase-substrate relations in human suggested a strong enrichment relative to the null expectation of kinase-substrate pairs in which both proteins are expressed by the same transcription factor (Hu et al., 2014; Newman et al., 2013). Kinases and substrates therefore seem to be significantly co-regulated at the transcriptional level. Sequence changes outside the kinase protein-coding sequence therefore also seem to be important for the evolution of specificity.

1.2.2 Adaptor and scaffold proteins

Kinase specificity is also influenced by adaptor and scaffold proteins that bind to the kinase but have no catalytic activity (Figure 1.4). These proteins can modulate the specificity of kinases either by directing their subcellular localisation or by promoting their physical interaction with potential substrates (Schechtman and Mochly-Rosen, 2001). Both mechanisms serve to increase the local effective concentration of the kinase with respect to a particular substrate, and also to reduce the probability of kinase 'inhibition' by other substrates. The RACK (receptor for activated protein kinase C) proteins for example differentially localise different PKC isozymes (Mochly-Rosen et al., 1991), whereas scaffold proteins of the MAPK cascade (e.g. Ste5 in *S. cerevisiae*) serve to physically link the MAP3K, MAP2K, and MAPK enzymes (Dhanasekaran et al., 2007). Both mechanisms can be combined in a single class of adaptors, as is the case for the cyclin proteins, which can promote substrate interaction and also determine the kinase localisation (Jackman et al., 1995; Schulman et al., 1998). In co-crystal structures of the CDK2-cyclin complex, a backbone carbonyl of the cyclin subunit interacts with the preferred lysine residue at position +3, suggesting that adaptors can in some cases also influence kinase specificity directly at the active site (Alexander et al., 2011; Brown et al., 1999).

1.2.3 Substrate docking

Substrate 'docking' sites refer to sequence motifs on the substrate that bind to allosteric sites on the kinase (Figure 1.5). This mechanism is similar in principle to adaptor and scaffold binding except that interaction occurs with the kinase directly rather than through an intermediate protein (Miller and Turk, 2018). The substrate interaction with the kinase can occur either through the kinase domain or an external one, but by definition cannot com-

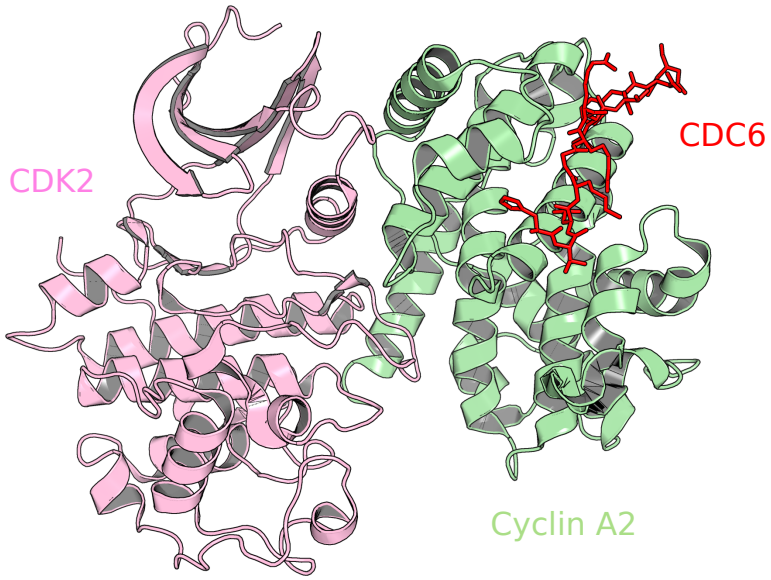


Fig. 1.4 Adaptor proteins (cyclin A2) often bind to short linear motifs (SLiMs) in the substrate (CDC6), thereby recruiting the kinase (CDK2) to its target. *PDB: 2CCH*

pete with the phosphorylation interface (Biondi and Nebreda, 2003). Docking for Ser/Thr kinases occurs more commonly via the kinase domain itself whereas Tyr kinases tend to rely upon accessory domains for substrate docking (Ubersax and Ferrell, 2007). In the Ser/Thr kinase PKCι, for example, a short linear motif located in the C-terminal lobe of the kinase domain can be used to recruit substrates (Linch et al., 2013). For tyrosine kinases, the SH2 domain is often used as a docking module for the binding of phosphorylated tyrosine residues (Roskoski, 2004).

Docking interactions can in some cases also be integral to the activation of the kinase. PDK1 kinases for example are responsible for the activation of many other AGC kinases, but employ allosteric sites to interact with the substrate prior to active site binding (Biondi and Nebreda, 2003). The docking site on the substrate however must be phosphorylated on a hydrophobic motif before the binding of PDK1, which then activates the substrate via activation loop phosphorylation (Pearce et al., 2010). Such 'phospho-primed' docking also occurs for kinases of the GSK (glycogen synthase kinase) *Family*. However, for these kinases the primed phosphorylation occurs close to the substrate P0 site and so is more important for defining GSK specificity than substrate activation. Specifically, the priming phosphorylation is only four residues C-terminal to the phosphorylatable Ser/Thr and so is considered part of the GSK peptide motif (S/T-x-x-x-pS/T), which is traditionally defined for substrate sites 5 residues N- and C-terminal to the phosphoacceptor (Biondi and Nebreda, 2003).

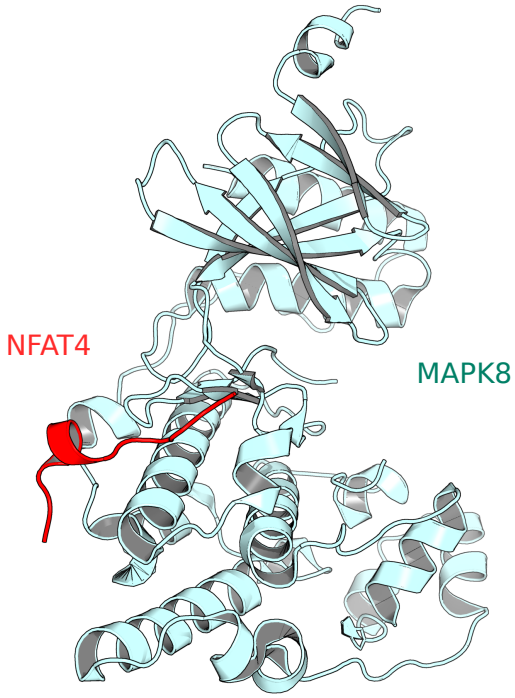


Fig. 1.5 Docking interactions between the kinase domain (MAPK8) and substrate (NFAT4) are also often an important determinant of kinase specificity. Many different binding pockets on the kinase can be used for this purpose (Biondi and Nebreda, 2003). *PDB:2XRW*

1.2.4 Phosphoacceptor specificity

The most discriminating feature of target sites is usually the phosphoacceptor itself – serine, threonine, or tyrosine. Of these, tyrosine phosphorylation is likely the most derived trait as it accounts for < ~1% of all phosphorylations in human and similar levels in other non-metazoan species (Pease et al., 2013; Sharma et al., 2014; Stark et al., 2010). Tyrosine phosphorylation in the metazoa, choanoflagellates, and the filastaerea can be accounted for by the proliferation of dedicated kinases in the TK group that phosphorylate tyrosine exclusively (Pincus et al., 2008). These kinases are characterised by a deep catalytic cleft to accommodate the bulk tyrosine side chain (Ubersax and Ferrell, 2007).

Dual-specificity kinases also exist that can phosphorylate any one of serine, threonine, or tyrosine. This trait likely evolved multiple times independently among the Ser/Thr kinases as the DSKs do no constitute a single monophyletic group like the tyrosine kinases do. DSKs for example are found among disparate *Families* such as the MAP2Ks (STE), the DYRKs (CMGC), and the Wee kinases ('Other') (McGowan and Russell, 1993; Roskoski, 2012; Walte et al., 2013). In the majority of cases the tyrosine phosphorylation takes place on an activation loop via a transactivation or autoactivation mechanism. However, for kinases in general target sites within the activation loop are not representative of the kinase peptide specificity (Miller et al., 2008; Pike et al., 2008), and so it is not clear if dual-specificity is determined by sequence changes in the kinase active site. Notably, a search in 1992 to identify sequence-based determinants of dual-specificity failed to return any positive results (Lindberg et al., 1992)

Many Ser/Thr kinases are marked by a strong preference for either serine or threonine. For example the STE kinase PAK4 strongly prefers serine phosphorylation at the active site whereas a different STE kinase, MST4, strongly prefers threonine phosphorylation. An analysis of aligned kinase domain sequences reveals that such preferences co-vary with the residue directly C-terminal to the DFG motif ('DFG+1'), which is proximal both to the phosphoacceptor residue and the ATP γ-phosphate and thus likely to modulate the rate of catalytic transfer (Chen et al., 2014a). Notably, multiple other studies have implicated 'DFG+1' as a determinant for substrate preference at the +1 position (Brinkworth et al., 2003; Howard et al., 2014; Kannan and Neuwald, 2004), suggesting that a single kinase SDR can in some cases determine substrate preference at more than one site (Figure 1.6).

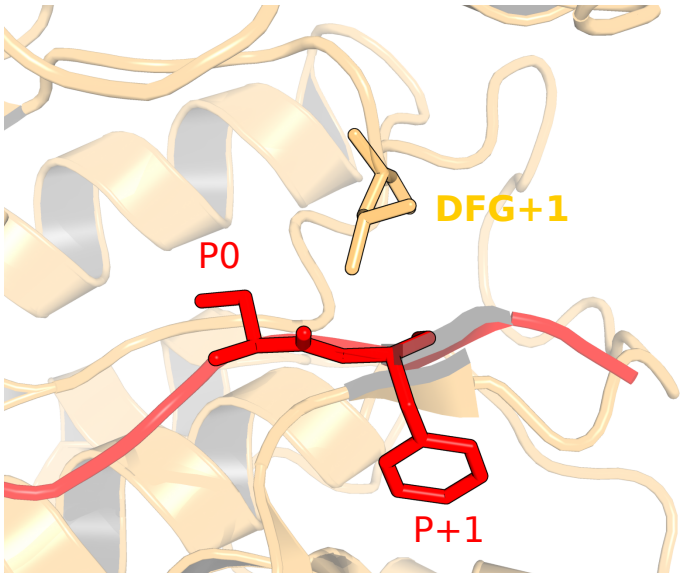


Fig. 1.6 The 'DFG+1' residue refers to the site directly C-terminal to the DFG motif. DFG+1 is proximal both to the 0 and +1 substrate positions and may influence specificity at both sites. *PDB:3CQW*

1.2.5 Peptide specificity

Kinase specificity at the active site is often referred to as 'peptide specificity', as it is traditionally defined by the phosphoacceptor residue and the five residues N- and C-terminal to it. It is often represented in the form of a sequence motif that describes the residues in the target site that are present significantly more often than would be expected by chance. The sequence motif of AKT1 for example is (R-x-R-x-x-S/T-x-x-x-x-x) as the arginine at positions -5 and -3 (relative to the phosphoacceptor) exhibit a significant enrichment in frequency in target sites relative to their background frequency in the proteome.

A protein kinase substrate motif can be derived either from the analysis of known target sites or from *in vitro* biochemical analyses of the kinase in the presence of phosphorylatable synthetic peptides. For most kinases the two orthogonal approaches are mutually validating in the sense that the most conserved substrate sites will also be the most important for the efficient phosphorylation of the peptide. This was indeed the case for the first discovered motif (PKA: R-R-x-S), in which the mutation of either arginine in a synthetic peptide led to the severe reduction in the phosphorylation levels of the peptide (Cohen, 2002; Kemp et al., 1976; Zetterqvist et al., 1976). Protein kinase specificity could therefore be defined very gradually from the *in vitro* identification of new target sites or from the analysis of synthetic peptides (Clarke and Hardie, 1990). Such reductionists approaches however were

slow to yield insights; for example, 15 years after the identification of the first motif, only 10 additional motifs were revealed (Pearson and Kemp, 1991).

1.3 Identification and prediction of new kinase substrates

1.3.1 Experimental detection of new substrates

More high throughput methods have since been developed for the rapid identification of kinase substrates (Xue and Tao, 2013). One approach is to knockdown or inhibit a kinase of interest and then to perform phosphoproteomic profiling on the mutant cell (de Oliveira et al., 2016). However, technical challenges notwithstanding, any phosphorylation changes measured only confirm a kinase-substrate association at best and do not provide evidence for a direct kinase interaction. Putative substrates can also be mutated in a genetic screen and then assayed to determine whether they suppress or phenocopy a kinase knockdown of interest (Xue and Tao, 2013). Again, such tests reveal kinase-substrate associations only and usually require follow-up validation experiments to demonstrate a direct relationship (Kettenbach et al., 2011).

Protein arrays can alternatively be used to infer kinase-substrate relationships. Here the kinase of interest is incubated with an array of whole-length putative substrates in the presence of radiolabelled ATP, and reactions are detected using autoradiography (Mok et al., 2009; Newman et al., 2013). Such arrays however are imperfect *in vitro* models for interactions that may occur *in vivo*, resulting in a number of false positive and false negative predictions. The 'phage display' display method may be used in a similar manner for whole proteins (Fukunaga and Hunter, 1997), with the advantage that multiple selection rounds can be used to amplify weak signals. Improper substrate folding in the bacterial system however may result in false negative predictions (Pillay, 2004; Xue and Tao, 2013). Some related methods (yeast two-hybrid, tandem affinity purification, etc.) seek to identify kinase-protein interactions rather than kinase-substrate interactions *per se*, though transient kinase-substrate interactions cannot reliably be detected by such methods, and any detected interaction will not necessarily be a kinase-substrate relationship (de Oliveira et al., 2016).

MS/MS-linked *in vitro* methods on the other hand are becoming increasingly popular. These involve the incubation of the kinase of interest with a peptide or cell extract, followed by shotgun proteomics to identify target phosphorylations. In many cases the extract will be pre-treated with a phosphatase to remove endogenous phosphorylations in the proteome. Following incubation with the kinase, the resulting phosphopeptides will then be enriched (TiO_2, IMAC, etc) and analysed using mass spectrometry (Douglass et al., 2012; Imamura

et al., 2014). An alternative approach is to generate by mutation a kinase allele sensitive to nucelotide analogues; substrates of the mutant kinase can then be detected by immunopurifying proteins with a synthetic chemical tag that results from the covalent modification of the substrate (Allen et al., 2005).

1.3.2 Experimental determination of kinase specificity

Other methods exist that characterise the kinase peptide specificity without identifying any physiological substrates. The phage display method mentioned previously for example can be adapted so that the phage present randomised S/T/Y-containing peptides rather than full-length proteins (Dente et al., 1997; Schmitz et al., 1996).

The use of synthetic peptide arrays however is generally more popular. Orientated peptide libraries for example are large mixtures of synthetic peptides containing a central serine, threonine, or tyrosine residue flanked by degenerate amino acids (Songyang et al., 1994). After incubation with a kinase, kinase substrate preferences are revealed during MS/MS analysis by an over-representation of one of the degenerate flanking residues among the phosphopeptides. Positional scanning peptide libraries (PSPLs) are similar but makes use of several different mixtures in which one of the flanking positions is fixed for a particular amino acid. Substrate preferences are then revealed after exposing the 9x20 different mixtures to a phosphor screen and quantifying the signal intensity for each spot, which correspond to a kinase preference for a given amino acid at a given position (e.g. arginine at position -3). This method is more sensitive than the use of OPLs and is better able to identify negatively-selected residues (Hutti et al., 2004).

More recently, heterologous systems have been used to determine the peptide specificity of a kinase of interest (Corwin et al., 2017; Lubner et al., 2016, 2017). In these systems, the kinase (usually human) is expressed transgenically in a model organism where the background phosphorylation of interest is likely to be low (e.g. S/T phosphorylation in prokaryotes or Y phosphorylation in *S. cerevisiae*). The phosphoproteome that results from the induced kinase expression is therefore likely to reflect the peptide specificity of that kinase. In 2017 such an approach was implemented in *S. cerevisiae* to define the peptide specificities of 16 human non-receptor tyrosine kinases (Corwin et al., 2017). In 2018, an extensive pool (the SERIOHL peptides) of serine-phosphorylated peptides (>100,000) was expressed heterologously in *E. coli* to generate a serine proteome library (Barber et al., 2018b). This pool of peptides can be incubated with protein kinases *in vitro* for the MS-based determination of substrate specificity, as has been recently demonstrated using the 'SERIOHL-KILR' approach (serine-oriented human library–kinase library reactions) (Barber et al., 2018a).

1.3.3 Computational prediction of kinase substrates and specificity

Despite recent advances, the experimental determination of kinase substrates and kinase specificity remains time-consuming and expensive (Kobe et al., 2005; Trost and Kusalik, 2011). A number of *in vivo* phosphoproteins are also though to remain 'hidden' from current methods due to their low abundance (de Oliveira et al., 2016). For these reasons, a large amount of research effort has been invested in developing computational models of kinase specificity capable of target prediction.

Phosphosite sequence -based predictors

The earliest predictive methods tended to be entirely sequence-based. Such methods use known *in vivo* targets to construct a computational model of kinase specificity that can then be used to query a list of candidate substrates for potential target sites. In the case of the tool *Prosite*, the 'model' is simply a string of letters representing the consensus motif, which can be used to query candidate substrates using pattern matching (Sigrist et al., 2002). Matrix-based methods (*Scansite*, *PHOSITE*, *PhoScan*) are similar in principle but assign weights to all 20 amino acids rather than representing only the most frequent amino acids (Koenig and Grabe, 2004; Li et al., 2008; Obenauer et al., 2003). Both methods are apparently limited in the sense that neither represents any inter-positional dependencies that may exist between the substrate positions. A popular tool developed in 2008 named *NetPhorest* did use artifical neural networks (ANNs) for some kinases to model inter-positional effects (Miller et al., 2008). However, in a later study the incorporation of such inter-positional dependencies into complex models did not improve predictive performance relative to regular PWMs for the 3 kinases (ATM/ATR, CDK1, and CK2) tested (Joughin et al., 2012). This suggests that inter-positional effects in kinase substrates either do not exist or are too weak to be detected given the low sample size for most kinase substrates.

Phosphosite structure -based predictors

Other tools make use of complex machine learning methods (neural networks and support vector machines) to model kinase specificity (Trost and Kusalik, 2011). The relevance of these approaches is that they can incorporate many different features for the prediction of new kinase targets. This is especially important considering the degeneracy of most substrates motifs, which alone would result in many false positive predictions. Some tools use the three-dimensional structure around the phosphorylation site additionally to guide predictions of potential substrates. *NetPhos* for example uses artificial neural networks for predictions (Blom et al., 1999), while *Phos3D* and *PhosK3D* uses a support vector machine (Durek

et al., 2009; Su and Lee, 2013). In the latter two cases, kinase specificity is represented by the radial pattern of amino acids biases in the vicinity of the phosphosite (3-12 Å away), rather than the sequence logo conventionally used for sequence-based approaches. While the incorporation of structural data does seem to add modest predictive value to some kinase models, it is thought that the lack of phosphosite structures currently present in the PDB may for now preclude the training of models significantly better than sequence-dependent predictors (Durek et al., 2009; Plewczyński et al., 2005; Trost and Kusalik, 2011).

Network-based predictors

So far the use of network-based data has been more successful than the use of structural data to inform substrate predictions. The popular *NetworKIN* tool for example makes use of the STRING functional association network in the sense that putative kinases 'closer' (separated by fewer edges) to the phosphorylation site in the network are prioritised over more 'distant' kinases when making predictions, everything else being equal (Linding et al., 2008). During benchmarking, the combined use of substrate motifs and cellular context (i.e. STRING network) was found to generate more accurate predictions than the use of substrate motifs alone. A more recent study by Wagih and colleagues is similar in principle except that the STRING functional association network was used to predict the kinase specificity logo directly rather than the kinase target sites (Wagih et al., 2015). Their analysis assumes that target sites are more likely than not to be proximal to their effector kinase in the STRING network, and so an approximate specificity model could be constructed by sampling phosphomotifs (using general MS data) from proteins that neighbour the kinase of interest. Finally, the CEASAR method represents a more holistic approach as it predicts both kinase specificity models and kinase-phosphosite relations (Newman et al., 2013). In this case, the protein association network was generated *in vitro* from protein microarray data rather than from STRING, but the approaches used to generate the logos and the kinase-phosphosite relations are analogous to the methods used in Wagih et al and NetworKIN, respectively. These approaches are expected to become more accurate in the future as *in vivo* phosphorylation data and protein association data becomes more readily available.

Kinase sequence-based predictors

The final class of predictor to be discussed uses the kinase primary sequence as an input to predict the kinase peptide specificity. None of the existing methods in this class predict kinase specificity completely *de novo* but instead rely upon homology of the query kinase to a set of reference kinases for which specificity has been experimentally determined.

The *Predikin* method was established first and predicts specificity independently for positions -3, -2, -1, +1, +2, and +3. For each substrate position, a small set of kinase domain specificity determining residue (SDR) is thought to confer substrate preference at that position (Brinkworth et al., 2003). It was therefore reasoned that the specificity of novel kinases at a given substrate position could be predicted by sampling target sites from reference kinases homologous to the query kinase at the relevant SDRs (Saunders et al., 2008). In this case, the SDRs for each position were selected on the basis of their observed proximity to the substrate peptide from available kinase-substrate crystal structures at the time. *Predikin* is therefore still a sequence-based method primarily although its development (i.e. SDR selection) was guided by a detailed structural analysis of kinase-substrate complexes.

The *KINspect* approach is similar in some ways to *Predikin* although its implementation is more sophisticated (Creixell et al., 2015a). Like *Predikin*, the specificity of the query kinase is predicted on the basis of its homology to characterised reference kinases. However, all positions within the kinase domain are first assigned weights randomly between 0 and 1 that represent the importance of a given residue as a specificity determinant. Many iterations of a machine-learning based optimisation algorithm are then implemented to generate a combination of ~250 weights (the 'specificity mask') that is optimal for the prediction of specificity within the cross-validation set. Kinase specificity is modelled using PWMs, where each query PWM is generated from the weighted average of all PWMs belonging to homologous kinases, and where reference kinases more similar to the query kinases are assigned a higher weight than more distant kinases. Importantly, when assessing homology between kinase domains, each kinase domain position is not considered equally but is weighted between 0 and 1 (as described above), so that positions with a score of 0 are essentially ignored when calculating homology. Therefore, *Predikin* represents a binarised form of the *KINspect* algorithm in which select positions within the domain (SDRs) are assigned a value of 1 but every other position is 0, and in which a different specificity mask (i.e. set of SDRs) is used to calculate matrix values for each substrate position.

The method developed by Safaei and colleagues is also similar to *Predikin* in some respects (Safaei et al., 2011). For example, a set of SDRs is assigned to each substrate position in this study instead of to the whole kinase domain. Unlike *Predikin*, however, SDRs for this analysis were selected based on their correlation (mutual-information based) with particular substrate preferences rather than by the inspection of kinase-substrate complexes. In this case a training set of 224 kinases of known specificity was used, and the mutual information values were weighted by the probability that the observed amino acid pairs would interact *in vivo* (e.g. acid-acid pairs were down-weighted but acid-base pairs were up-weighted). The SDRs selected for each position represented the 7 kinase positions that correlated the

most strongly with the observed substrate preferences. A PWM for any given query kinase could then be predicted by finding, for each substrate position, the average conditional probability of the amino acids given the amino acid identity of the 7 SDRs, where the 7 mutual information values were used to generate a weighted average of the conditional probabilities.

Computational methods that predict specificity from the kinase primary sequence could be employed for the evolutionary analysis of kinase specificity. Such analysis is not currently feasible for network- and structure-based methods – or even standard phosphosite-based methods – where the requisite data is limited to a few model organisms. It is therefore surprising that none of the three methods described above have been leveraged towards an evolutionary analysis so far. There may be concerns however about the accuracy of these methods and/or their applicability to species distantly related to the model organisms (human and baker's yeast) from which these methods are based. It should also be noted that even for the two model organisms mentioned, no more than 50% of the kinome has been experimentally profiled, thus imposing a limit on the coverage of models that could be trained with current data.

1.4 Identification of protein kinase specificity determinants

Multiple experimental methods exist for the identification of kinase specificity determining residues. Perhaps the most simple approach is to perform a detailed structural analysis of a kinase-substrate complex so that the chemical basis for specificity can be rationalised. However, such rationalisations can only contribute qualitatively to the understanding of specificity, and still ultimately constitute a set of hypotheses that need to be verified experimentally. Validation of a predicted SDR typically requires that the kinase of interest be mutated at the relevant SDR position, followed by independent experimental characterisations of the wild-type and mutant kinases. This approach can yield good quantitative insights into specificity determination, but is limited in the sense that the results apply only to the kinase assayed. For a broader overview of specificity determination it is necessary to cross-reference the experimental and/or structural analyses with a multiple sequence alignment of kinases of known specificity. This would allow the researcher to then determine whether the identified SDR is particular to the kinase analysed or is more broadly relevant for specificity. Finally, it is also possible to investigate kinase specificity by manually generating kinase-substrate models given that the structure and specificity of the unbound kinase is known. These approaches take advantage of the fact that the number of non-redundant kinase structures in the PDB far outnumbers those structures with peptide substrate bound, although any models

constructed will be associated with a high level of uncertainty. Examples for each approach are described below.

1.4.1 Structures

Note: when referrring to kinase specificities in this thesis, the name of the preferred amino acid is used in addition to its position relative to the phosphoacceptor e.g. R-3 kinases are kinases that prefer substrates with an arginine (R) residue 3 positions N-terminal to the phosphorylated serine, threonine, or tyrosine.

Surprisingly few non-redundant examples exist of crystal structures with the kinase bound to the peptide substrate. It has been suggested previously that the low affinity of kinases for substrates – a physiologically necessary feature for most signalling complexes – in particular poses a problem for kinase-substrate co-crystallisation (de Oliveira et al., 2016; Endicott et al., 2012; Goldsmith et al., 2007). The relatively few complex structures that do exist however cover some of the most common substrate preferences. Three (R-3, P+1, and R-2/R-5) are discussed below as examples:

R-3

Protein kinase A (PKA) was the first kinase to have its structure revealed, which occurred in 1991 and happened to include the kinase in complex with a peptide substrate (Knighton et al., 1991). The structure revealed the binding mode of the kinase with respect to the R-3 determinant, which is a substrate preference often found in kinases belonging to the AGC and CAMK *Groups*. This structure in particular implicated the glutamate at kinase position 84, which contacts the positively charged arginine side chain. Contacts are also made however with the ATP molecule and a backbone carbonyl group of the N-terminal lobe. In 2012 a structure of Protein kinase Cι in complex with its substrate revealed a very similar binding mode for R at the -7 position in the linear sequence (Wang et al., 2012a); in this case the classical 'R-3' motif had been reconstituted from the the three-dimensional folding of the substrate (Figure 1.7).

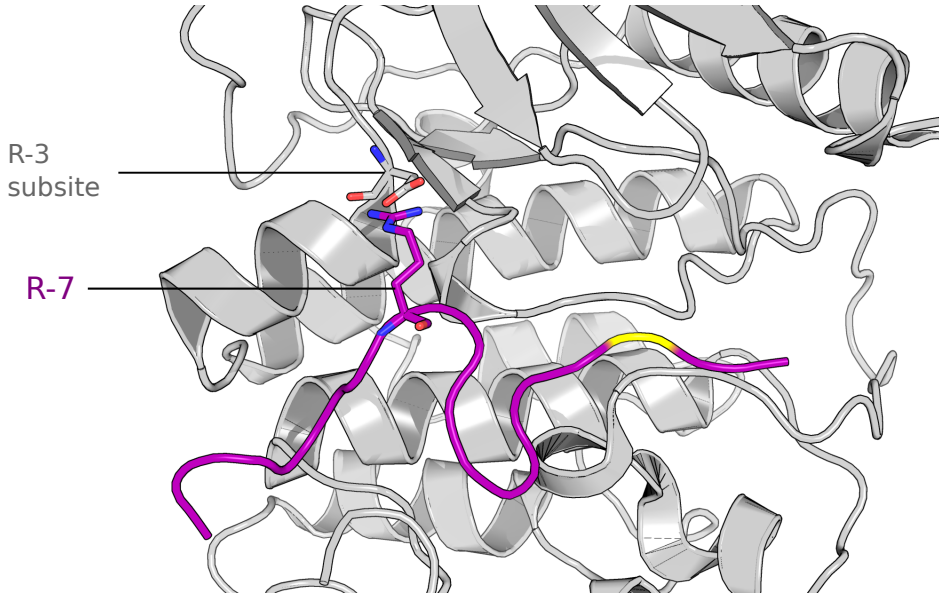


Fig. 1.7 An example of a kinase binding to the substrate in which the substrate peptide is not found in the canonical extended-linear conformation. Here, binding of the arginine at position -7 is structurally analogous to R-3 binding for linear substrates. The P0 position is represented in yellow. *PDB: 4DC2*

P+1

Most CMGC kinases feature a strong preference for proline at position +1. Proline is not a charged amino acid and so cannot be easily rationalised on the basis of opposite-charge interactions between the kinase and substrate. The release of the first CMGC kinase-substrate structure in 1999 suggested that its selectivity arises from its unique side-chain structure (Brown et al., 1999). In particular, it appears that for proline +1 there is no amide group to form a hydrogen bond with the backbone carbonyl group of the activation segment residue at position 159 (as is observed in PKA). Unlike for other kinase *Groups*, the 159 position in CMGC kinases is not represented by glycine; this allows the backbone carbonyl group to instead be stabilised by the arginine at position 164 rather than substrate amino acid at +1 (Figure 1.8). This explains why proline at +1 is highly disfavoured in most non-CMGC kinases i.e. because there would be no arginine at 164 or backbone amide at +1 to stabilise the glycine carbonyl at position 159 (Zhu et al., 2005a).

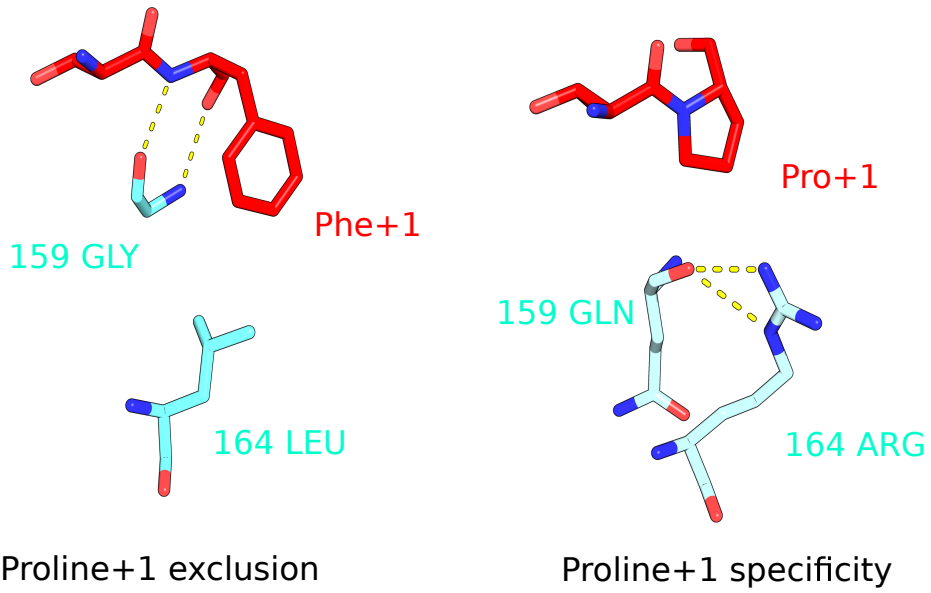


Fig. 1.8 The negative selectivity (disfavour) of non-CMGC kinases for proline at the +1 position can be explained by the inability of the proline side group to stabilise the backbone carbonyl of the residue at domain position 159. *PDB: 3CQW (left), PDB: 2WO6 (right)*

R-2

The substrate peptide bound to the 1991 structure of PKA also features an arginine at the -2 position. This is a fairly common substrate preference although not quite as strong or as prevalent as the R-3 preference. From the PKA structure, it is apparent that the arginine at position -2 forms specific contacts with glutamates at positions 127 and 189 (Figure 1.9). The release of the AKT1-peptide structure in 2008 later demonstrated that the R-5 subsite in AKT overlaps closely with the R-2 subsite of PKA (Lippa et al., 2008). Ben-Shimon and colleagues in 2011 conducted a more holistic study by analysing several different kinase structures (unbound) with either an R-2 or R-5 preference (Ben-Shimon and Niv, 2011). This was achievable using the novel 'AnchorsMAP' method that scans an arginine probe along the surface of the kinase, and identifies putative binding sites from computed Gibbs free-energy dG calculations. The importance of particular residues for binding could then be investigated by coupling the 'AnchorsMAP' approach with targeted *in Silico* mutagenesis. This approach yielded many insights that would not have been possible using the narrow set of kinase-substrate complexes with arginine present at -2. For example, a number of secondary SDRs were predicted, and it was shown that the 127/189 pair is important for R-2 binding but much less so for R-5 binding.

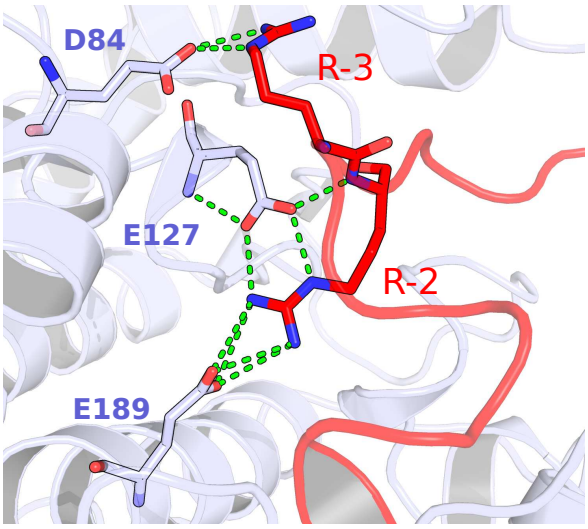


Fig. 1.9 Opposite charge-interactions help to determine selectivity for arginine residues at positions -2 and -3 in the substrate. This is generally determined by SDRs at positions 127 and 189 for R-2 selectivity, and an SDR at position 84 for R-3 selectivity

1.4.2 Homology models

Kinase-substrate models can also be created for a kinase of interest where the relevant structure is not publicly available. Such models are typically constructed by superposing an unbound structure of the kinase of interest against a homologous kinase-substrate complex. The homologous kinase can then be removed and the substrate peptide mutated *in Silico* to match a known target site of the kinase of interest. The kinase-substrate model is then generated by a structural energy minimisation (EM) step, followed by an optional short molecular dynamics (MD) step.

Kinase-substrate models have historically been used to rationalise more obscure motifs associated with only a single kinase *Family*. The CK2 *Family* (CMGC) for example prefers to bind acidic residues at positions +1 and +3. Two different kinase-substrate models previously generated both suggest that the +1 binding pocket overlaps with that of the proline-directed CMGC kinases, but that lysine/arginine substitutions at positions 164, 161, and 157 now confers the D/E preference (Niefind et al., 2007; Sarno et al., 1997). However, models generated by different research groups can be radically different from each other. For human SRPK1 (CMGC), which prefers arginine at +1, Nolen also places the +1 residue in the canonical +1 binding pocket for the yeast orthologue (Nolen et al., 2001), but Kannan and Neuwald place the +1 residue outside the canonical +1 binding pocket and instead proximal to the SRPK-specific asparagine at position 144 for the human kinase, and place the serine at position +2 in the canonical +1 binding pocket (Kannan and Neuwald, 2004). This example

is an important reminder that kinase-substrate models, although useful, should be treated with caution.

Most of the models constructed assume that the peptide binds to the kinase in an extended-linear conformation. At least one published model however suggests that the substrate peptide may assume a 'hairpin' conformation when binding to the kinase. Specifically, a model generated for the MPSK (NAK *Family*; 'Other') kinase suggests that the presence of an α-helix at the C-terminal end of the activation segment (ASCH) would preclude the binding of the peptide in an extended-linear conformation. Instead, the peptide is speculated to form a half-turn about the +1 position so that more C-terminal positions (+2, +3, +4, etc.) bind between the ASCH and the α-G helix at the C-terminal lobe (Eswaran et al., 2008). This is a bold prediction but not completely without precedent, as the kinase-substrate crystal structure of haspin kinase published in 2014 shows that the peptide forms a U-shaped hairpin structure at the active site, with only the P0 Thr residue present in the expected substrate position (Maiolica et al., 2014). Both haspin kinase and MPSK belong to the 'Other' kinase *Group*. However, a kinase closely related to MPSK (GAK, NAK *Family*) was predicted to have the substrate bind in the extended linear conformation, with the C-terminal side resting 'above' the ASCH in the groove between the two kinase lobes (Chaikuad et al., 2014). This highlights once again the high level of variation that can exist even between models of closely related kinases.

1.4.3 Kinase sequence alignments

Alignment-based approaches are more appropriate for the global analysis of kinase specificity. The relevant tools for this approach use an alignment of protein kinase sequences as an input and for each alignment position give a probabilistic indication as to whether or not that position is a specificity determinant. The output however is unlikely to be completely accurate, and a number of false positive and false negative predictions are expected. For this reason, alignment-based approaches are often coupled with experimental and/or structural analyses to identify the likely true-positive predictions.

All alignment-based approaches to some extent test for statistical associations between a particular kinase sequence position and a particular substrate preference. In most cases this will involve first grouping the kinase alignment sequences on the basis of their specificity; putative SDRs are then marked as positions with maximal similarity within groups but minimum similarity between groups. Two of the most relevant previous studies that adopt this approach however do not group the kinases directly on the basis of specificity. Li et al., 2003 for example collects kinases sequences into *Families* (PKA, PKC, RAC, GRK, S6PK, and PDPK1) and compares sequences between them under the assumption that specificity will

be conserved within kinase *Families* (Li et al., 2003). For a different study (Kannan and Neuwald, 2004), the objective was to identify the residues responsible for the functional divergence of the CMGC *Group* and for the divergence of different *Families* within the CMGC (CDK, DYRK, GSK, etc). Many of the results were relevant to kinase specificity however as CMGC membership correlates strongly with the proline+1 preference and the CK2 and SRPK *Families* both have distinct specificities, as discussed previously.

Many of the results reinforce the structural predictions discussed previously. The 164 position (R) discussed above with respect to the proline+1 preference for example emerges as a strongly predicted CMGC determinant from an alignment-based approach (Kannan and Neuwald, 2004). The same study also strongly implicates lysine at the 164 position discussed above (with respect to CK2 homology models) as a residue responsible for the functional divergence of CK2 kinases. For the Li et al., 2003 study, 7 of the 16 SDRs identified for the 6 different AGC *Families* map to residues in close proximity to the peptide inhibitor (PKA, *PDB:* 1ATP). One of these (position 86) is a likely determinant of R-3 specificity. The earliest studies therefore confirmed the value of using alignment-based approaches to infer SDRs.

Where kinase PWMs can be constructed it is not strictly necessary to divide the kinases into discrete groups. Instead a covariance analysis can be performed between the kinase sequence and the numerical representation of kinase specificity (i.e. the PWM). This was the approach used for the Mok et al., 2009 study of 61 Saccharomyces cerevisiae kinases that had been characterised using a peptide screening approach (Mok et al., 2010). Again, many of the predictions accorded with results from previous structural analyses. The R-3 determinant at position 84 discussed previously for example was predicted using this approach. The R-2 determinant at position 127 was predicted also in this study. Two of the previous methods (Creixell et al., 2015a; Safaei et al., 2011) used for the prediction of kinase specificity in effect also make use of this covariation approach, and involve the prediction of SDRs as an intermediate step. For the *KINspect* algorithm, all kinase domain positions with a predictive weight equal to 0.9 or above were designated as SDRs by the authors, with some being verified experimentally (discussed below).

1.4.4 Mutational analysis

Experimental methods for the identification of kinase SDRs have become more advanced with time. In the earliest studies, a putative SDR would be mutated to alanine and then the activity of the mutant would be assayed against an optimal substrate using autoradiography. This approach confirmed in 3 independent studies that a glutamate/aspartate at position 84 could indeed serve as an R-3 determinant (Batkin and Shaltiel, 1999; Gibbs and Zoller, 1991;

Huang et al., 1995), as is also suggested by structural analyses. Specific positions in the substrate will usually also be mutated so that the SDR can be assigned to its cognate substrate position. In this case, SDR-substrate interactions are indicated by non-additive effects of the kinase-substrate double mutant (Moore et al., 2003; Sarno et al., 1997; Scott et al., 2002). A more direct approach may be to 'rescue' a deleterious kinase mutation with a complementary mutation in the substrate. For example, in the human kinase AMPK1, it was found that the effect of mutating the preferred M at position -5 to the negatively charged aspartate could be suppressed by mutating leucine at position 195 to arginine, thereby implicating position 195 as an L/M-5 determinant (Scott et al., 2002).

More effective approaches are now in use, as any existing method for the experimental characterisation of wild type kinases may also be directed towards mutant kinases. In the Mok et al., 2009 study for example, peptide library screens were used to demonstrate that mutation of the Kss1 kinase (CMGC) at position 189 from serine to glutamate confers a preference for R at position -3 (Mok et al., 2010). Similarly, an approach coupling kinase mutagenesis with PSPL was used in 2014 to provide weak evidence for the role of domain position 42 and 44 in conferring the R/K+2 preference found in the protein kinase C *Family* (Creixell et al., 2015a). The big advantage of such methods is that they will generate a specificity profile for the mutant kinase; this allows the researcher to identify changes in specificity arising from mutation that may not have been predicted computationally. This should be contrasted with the approach described above, which can usually (i.e. without an extensive amount of experimentation) only validate an interaction between an SDR and substrate position that is already suspected. The profile will also represent kinase preferences for all 20 amino acids and not just the two amino acids involved in the substitution. Finally, more recent MS-based approaches have been developed that will profile a mutant kinase's specificity on the basis of its activity *in vivo* against a proteome substrate (i.e. cell lysate). In one specific example a mutant human kinase (PKA) was expressed in *E. coli* so that its specificity could be inferred from the resultant phosphoproteome signatures present in the *E. coli* proteome. This approach was used to demonstrate that mutation of PKA at position 164 would diminish the preference of the kinase for hydrophobic residues at position +1 (Lubner et al., 2017). A very similar protocol from the same group was later used generate a double mutant in DYRK1a (at positions 159 and 164) to convert the wild-type proline+1 specificity into the D/E specificity usually found in CK2, thus generating a synthetic substrate motif (R-x-x-S/T-D/E) not previously observed in nature (Lubner et al., 2016).

1.5 Identification of functionally divergent residues

Alignment-based approaches for the detection of kinase SDRs belong to a broader set of methods for the general identification of protein subfamily SDRs. These methods can be divided roughly into those that make use of a sequence-based phylogeny and those that rely just on the MSA of the protein family of interest (Chagoyen et al., 2016; Chakraborty and Chakrabarti, 2015; Studer et al., 2013). The MSA-exclusive approaches tend to incorporate concepts from information theory and machine learning whereas the phylogenetic approaches are based on explicit evolutionary models. The latter generally partitions identified SDRs into 'Type I' and 'Type II' sites. Type II sites, as described previously, represent alignment positions with high conservation within functional groups but low conservation between them; Type I sites on the other hand refer to positions marked by rate asymmetry between groups, with one group being conserved and the other being non-conserved (Gu, 2006).

In the literature, phylogenetic approaches tend to be viewed favourably in comparison to purely sequence-based approaches (Chakraborty and Chakrabarti, 2015; Soyer and Goldstein, 2004; Studer et al., 2013). Here the assumption is made that the number of spurious predictions generated could be reduced by accounting for the phylogenetic non-independence between sequences. For example, MSA-exclusive approaches do not account for evolutionary distance when calculating per-site conservation, and so would not be able to distinguish cases in which a site is fully conserved within an ancient subfamily or a newly-emerged one. There is also the added advantage that phylogenetic approaches incorporate multiple features of evolution – distance between sequences (branch lengths), rate heterogeneity (*alpha* parameter), and amino acid composition (equilibrium frequency vector) – into a robust statistical framework that is well-established. Such approaches usually compare homogeneous models – where parameters are held constant across the phylogeny – to inhomogeneous models where the parameters are allowed to vary between Subfamilies (Gaston et al., 2011; Tamuri et al., 2009). Bayesian- or ML-based model selection approaches can then be used to identify sites where the inhomogeneous model is significantly more likely given the data, as represented by p-values. These statistical methods stand in contrast to MSA-exclusive approaches, where the sites in alignment are normally ranked on the basis of a custom scoring function that has no general interpretation.

Phylogenetic approaches however suffer from multiple disadvantages. One of them is that the input phylogeny is unlikely to be completely accurate, and so a layer of uncertainty is added when making SDR predictions. A second problem is that these approaches assume that the function of interest is exclusive to a monophyletic clade, and has not evolved paraphyletically or polyphyletically (i.e. convergently). The R-3 preference for example repre-

sents a paraphyletic trait as it was likely present in the ancestor of AGC and CAMK *Groups*, but is not found in all of its descendants. A more practical problem with these methods is that their relative difficulty of implementation makes them less amenable to an automated analysis of multiple different functions or protein families. This may be one of the reasons why, to the author's knowledge, phylogenetic methods have not yet been systematically benchmarked but alignment-based methods have.

Across these two groups, SDR-prediction tools can also be divided into the 'supervised' and 'unsupervised' methods (Chagoyen et al., 2016). Supervised methods rely upon sequence classifications provided by the user whereas unsupervised methods will try to infer functional sequence groups from the input data provided (using PCA, SOM, etc.). Other methods exist that will also explore systematically many possible sequence partitions. The 'evolutionary trace' method for example take a phylogenetic tree as an input and successively generates partitions of increasing specialisation, so that the functional sites identified may be fully conserved or just conserved within subfamilies (Lichtarge et al., 1996). The use of unsupervised methods however is thought to be largely unnecessary if the protein family can already be clustered on the basis of experimental data or previous characterisations.

In 2014, many of the MSA-based methods discussed above were benchmarked by generating predictions for 20 different MSAs where the SDRs had been identified experimentally (Chakraborty and Chakrabarti, 2015). A surprising result of this investigation was that many of the highest-score methods were based on relatively simple algorithms. A previous benchmarking had also shown that combining predictions from the three top-performing methods can achieve a higher specificity than the use of a single method alone, highlighting the importance of ensemble-based approaches (Chakrabarti and Panchenko, 2009).

1.6 Ancestral sequence reconstructions

Ancestral sequence reconstruction (ASR) methods predict the most likely sequence for every ancestral node in a phylogeny. Posterior probabilities are also assigned to every amino acid at every site so that the uncertainty in ancestral sequence predictions can be quantified. Currently, the PAML/CodeML and FastML programmes represent two popular tools for this purpose (Ashkenazy et al., 2012; Yang, 1997). Both methods reconstruct ancestral sequences within an ML (maximum likelihood)-based framework. Alternatively, Bayesian approaches can be used to account for the uncertainty in the phylogenies underlying ASR, although a simulation-based benchmark suggests that Bayesian approaches do not improve the accuracy of ASR (Hanson-Smith et al., 2010). Comparison between ML-based methods however has so far been limited. To the author's knowledge there has been only one bench-

marking study, which involved the experimental evolution of red fluorescent protein (RFP). This study concluded that ML-based approaches generally perform to a similar high degree of accuracy (Randall et al., 2016).

For most ASR studies, the computational and experimental analyses are closely integrated. The general objective is to study the evolutionary trajectory of a phenotype of interest by first 'resurrecting' ancestral proteins and then characterising them experimentally (Thornton, 2004). Such studies have yielded important insights into evolutionary cell biology. One study for example examined the evolution of a molecular scaffold protein (GKpid) needed for the orientation of the mitotic spindle, and revealed that a single historical substitution was responsible for an important innovation (binding of a cortical protein) (Anderson et al., 2016). A different study probed the divergence of steroid hormone receptor (SR) specificity following gene duplication, and revealed that the new specificity was largely determined by negative substitutions preventing binding to the ancestral DNA element (McKeown et al., 2014). More recently, ASR has been combined with deep mutational scanning to suggest that the reconstructed historical trajectory was one of many possible paths that could have yielded a similar outcome, with the actual trajectory depending strongly on the starting (ancestral) genotype (Starr et al., 2018).

The application of these methods to the protein kinase superfamily has so far been limited to one study (Siddiq et al., 2017). In 2014, Howard et al performed an ancestral sequence reconstruction on the CMGC *Group* of kinases. Most members of this *Group* strongly prefer target sites with a proline at the +1 position, but fungal kinases of the Ime2/RCK/LF4 *Family* prefer arginine at this position. Experimental analysis of the Ime2/RCK/LF4 ancestor suggests that this kinase was intermediate in specificity between the extant fungal (arginine+1) and metazoan (proline+1) kinases of this *Family* (Figure 1.10). This analysis therefore suggests a general model of kinase specificity in which specificities diverge proceeding from an ancestor of intermediate specificity. It remains to be seen whether this form of subfunctionalisation will apply to kinases of other families and specificities.

Ancestral sequence reconstruction approaches exist as part of a broader set of methods for the reconstruction of ancestral states (Joy et al., 2016). A broad range of phenotypes could be classified as a 'state' in this context. If the phenotype is discrete – the type of diet of Galapagos finches, for example – then a simple maximum parsimony (MP) approach can be employed (Schluter et al., 1997). A similar approach can be applied to predict the minimal sequence of structural alterations (inversions, deletions, and transpositions) required to produce multiple extant genomes (Bourque and Pevzner, 2002). In 2016, ancestral state reconstruction was applied in the context of signalling when used to predict across 18 fungal species whether or not an ancestral site would have been phosphorylated (Studer et al., 2016).

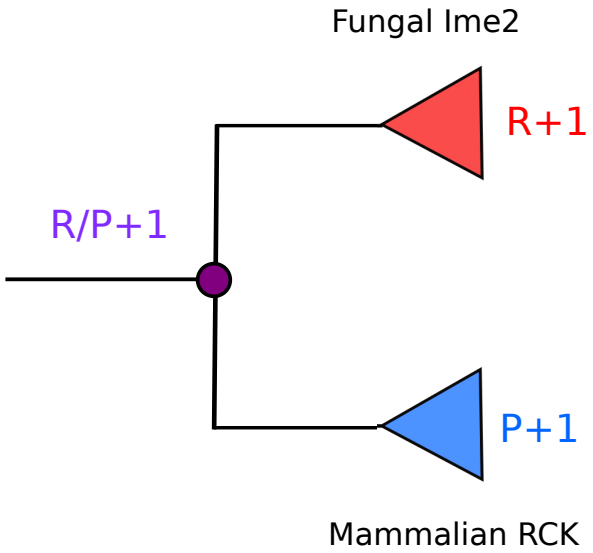


Fig. 1.10 The most recent common ancestor of the fungal Ime 2 kinases and the mammalian RCK kinases was likely intermediate in specificity between the two kinase clades at the +1 substrate position

1.7 Phosphoproteome analysis

Broad knowledge of kinase specificity is currently limited to only a few model organisms (human and budding yeast). While sequence-based predictors of specificity do exist (discussed above), they are not completely accurate and rely upon homology of the query sequence to reference sequences of known specificity. Such tools cannot therefore predict kinase specificities different from those already characterised. This would be especially problematic for species distantly related from human and budding yeast, many of which feature clade-specific kinase *Groups* and *Families*.

The analysis of phosphoproteomes is an alternative option, as they reflect an ensemble of kinase activities (and therefore specificities) present in the cell. This is achievable by coupling phosphopeptide enrichment methods (TiO_2, IMAC, Phos-Tag, immuno-based, covalent modification-based, etc) with tandem mass spectrometry (MS/MS) (Fíla and Honys, 2012). Phosphomotifs can then be identified as those sequence patterns that are over-represented among phosphorylation sites relative to randomly sampled S/T-centred sites in the proteome. The *motif-x* tool is commonly used for this purpose (Schwartz and Gygi, 2005). *motif-x* works by finding statistically significant residue-position pairs in a foreground set of phosphorylation sites using a binomial distribution. For each round of the algorithm, the tool iteratively searches for all such pairs until no more significant ones can be found. For CDK phosphorylation sites, for example, three such pairs (proline/-2, proline/+1, and lysine/+3) may be found until a motif is called (P-x-S/T-P-x-K). This is repeated many times

– each time with sites matching the motif iteratively removed from the foreground and background set – until no further motifs are identified .

Phosphoproteome data now exists for a broad range of eukaryotic species. For most species, motif enrichment analysis reveals motifs that can be broadly placed into one of three categories: acidic, basic, and proline-based/hydrophobic (Figure 1.11) (Amanchy et al., 2011; Lin et al., 2015a; Resjö et al., 2014; Studer et al., 2016). These are representative of known specificities for casein kinase II (D/E+1 and D/E+3), AGC/CAMK *Groups* (R-3), and the CMGC *Group* (P+1), respectively. This is the consistent with the observation that these kinase *Groups* are universal in the eukaryotes. More generally however, the relationship between the kinome and phosphoproteome has not been explored intensively, with a few exceptions. It has been noted for example that the depletion of the basic R-R-x-S/T and R-x-R-x-x-S/T motifs in *Arabidopsis* coincides with the depletion of their cognate effector kinases (PKA and AKT, respectively) (Resjö et al., 2014). In *Tetrahymena* also, the strong enrichment of the L-x-x-S/T motif correlates with the expansion of cognate Nek kinases in the genome (Tian et al., 2014). Also, as mentioned previously, in Studer et al., 2018, the enrichment of S/T-P in 18 fungal species was correlated with the number of predicted proline-directed kinases in each species.

Phosphoproteomic analyses have also been conducted for prokaryotic species (Lin et al., 2015b; Pan et al., 2015; Potel et al., 2018; Reimann et al., 2013; Wu et al., 2016). In most cases, however, the number of identified S/T phosphorylation is typically low, owing to the difficulty of phosphosite extraction in these species. At the time of writing, only the bacteria *E. coli* and the archaeon *Sulfolobus* have more than 1,000 known phosphosites. In the case of *E. coli*, neither of the classic S/T-P or R-x-x-S/T motifs were found to be enriched in this species, and the motifs identified in *E. coli* do not correspond to any known motifs in the eukaryotes (Lin et al., 2015b).

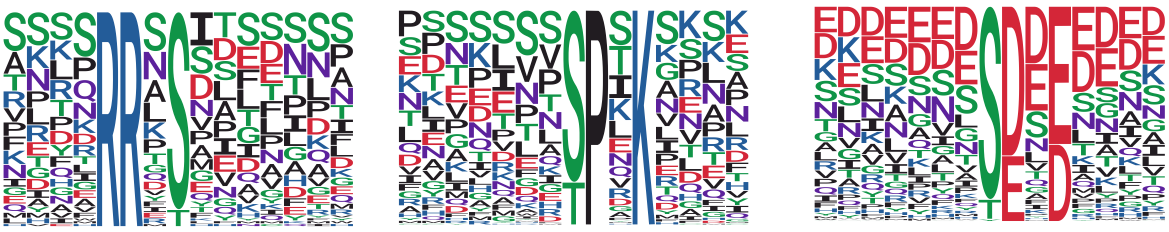


Fig. 1.11 Many phosphorylation motifs extracted from phosphoproteome data contain either basic residues (R or K), proline (P), or acidic residues (D or E)

1.8 Mutation of kinases in disease

Mutations to the ePK domain are often pathological. Diseases caused by the germline mutation of protein kinases are referred to as kinasopathies. Kinasopathies can have a range of physiological effects, but overlap particularly strongly with developmental and metabolic disorders (Lahiry et al., 2010). Mutations to AKT2 for example can lead to Type 2 diabetes (George et al., 2004). Moreover, mutation to the phosphorylase kinase γ chain can lead to glycogen storage disease type 9c (Maichele et al., 1996). In at least some of these cases, the structure-function relationship of the kinase domain can be used to infer the likely mechanism of disease. Cushing's syndrome for example is caused by a mutation to a protein kinase A SDR that prevents its binding to a pseudosubstrate inhibitor, resulting in overactivation of the enzyme (Cao et al., 2014; Cheung et al., 2015). When many diseases are considered together, hotspots of mutation become evident. For pathogenic germline mutations, it was found that regions involved in substrate-binding and regulation are mutated especially often (Lahiry et al., 2010; Torkamani et al., 2008).

A similar analysis for cancer-associated mutations revealed hotspots localising to the activation loop, catalytic regions, and the ATP-binding loop (Dixit et al., 2009). Molecular modelling analyses suggest that some of these mutations mediate their effect by destabilising the autoinhibited kinase structure (Dixit et al., 2009). The mutational hotspot represented by the BRAF V600E mutation is one example of this. Sites in the kinase employed for substrate binding may also be mutated frequently in cancers, as has been demonstrated for a dibasic motif in the PKCι enzyme (Linch et al., 2013). A more recent analysis has adopted a network-based approach to classify kinase mutations into those likely to alter network dynamics (i.e. kinase activity), and those likely to rewire the network (i.e. change kinase specificity). When considered alongside direct mutations to target sites, it was found that signalling networks were rewired in cancer cells to an extent greater than previously expected. Mutations altering kinase specificity were also identified and validated (Creixell et al., 2015b).

1.9 Aims of the thesis

The primary aim of this thesis is to explore the evolution of protein kinase specificity at the active site. In this *Introduction* chapter, I have given a broad overview of previous research that is relevant to this subject. In the next three chapters (*Chapter 2*, *Chapter 3*, and *Chapter 4*), I describe in detail research performed by me that addresses the question of the thesis directly. At the end of these three chapters, the results are discussed in the context of previous literature that relates to the thesis topic. In *Chapter 5*, an overview of the results from

the thesis is given, and potential future directions for the field are discussed. Finally, the methodology underlying *Chapters 2, 3, and 4* are described in *Chapter 6*.

Each of the three *Results* chapters has its own objectives. For *Chapter 2*, the objective was to predict new specificity determining residues (SDRs) for protein kinases by leveraging availabe data on kinase sequence, kinase structure, and kinase specificity. A strong emphasis is placed on giving structural rationalisations for the SDRs detected, and on using the results for the sequence-based prediction of kinase specificity. The aim was then to use these results to interpret the effect of genomic variants on kinase specificity, using data from *The Cancer Genome Atlas* as an example.

The aim of *Chapter 3* was to explore the evolution of kinase specificity primarily from the analysis of kinase domain sequences. This research would follow directly from the advances made in *Chapter 2*. One objective was to use the sequence-based predictor developed in *Chapter 2* to examine the extent to which kinase specificities are conserved between orthologues. Another aim was to globally predict changes in kinase specificity following gene duplication. Finally, the SDRs identified from *Chapter 2* would be used for a detailed evolutionary analysis of a single kinase *Family* by way of ancestral sequence reconstructions.

Finally, in *Chapter 4* the objective was to explore the evolution of kinase specificity by analysing phosphorylation data from several eukaryotic species. This could be achieved by extracting sequence motifs from phoshorylation data and calculating their enrichment across several species. The likely origin of the phosphorylation motif could then be predicted. Many motifs correspond to kinase specificities, and so I sought to explore the co-evolution between phosphomotifs and their upstream kinases where this was possible.

Chapter 2

Global analysis of kinase specificity determinants

In this chapter, specificity determining residues (SDRs) for protein kinases were identified and then rationalised using available structural data. I performed all of the computational analysis under the supervision of Pedro Beltrao. For the validation experiments, the mutations and kinase activity assays were performed by Cristina Viéitez at the EMBL research campus in Heidelberg, and the mass spectrometery experiments were performed by Vinothini Rajeeve under the supervision of Pedro Cutillas at the Barts Cancer Institute, Queen Mary University of London. Some of the work presented in this chapter was included in a preprint manuscript:

David Bradley, Cristina Viéitez, Vinothini Rajeeve, Pedro Cutillas, and Pedro Beltrao (2018): Global analysis of specificity determinants in eukaryotic protein kinases. *bioRxiv*

2.1 Introduction

The specificity of protein kinases has long been the subject of intensive research efforts. This is for the simple reason that knowledge of kinase specificity allows one to rationalise and predict the kinase-substrate interactions that are fundamental to most cellular processes. This can be exploited for clinical purposes as it enables the likely effect of kinase inhibition (by drugs) to be predicted (Cheng et al., 2014; Colinge et al., 2014).

Research into protein kinase structure and substrate targeting has therefore been extensive. However, most of the research conducted so far has concerned only a single kinase or kinase subfamily at a time (Eswaran et al., 2008; Onorato et al., 1991; Sarno et al., 1997;

Scott et al., 2002; Wang et al., 1998), and little attempt is usually made to cross-reference the results generated with those of kinases from other subfamilies. Such studies will provide their own insights to specialists but are of limited utility by themselves for the training of predictive methods or for the understanding of kinase evolution. Part of the motivation for this study was therefore to offer a global analysis of kinase specificity using the full extent of data currently available for kinase target sites and kinase structure.

To some extent holistic analyses have been attempted before but each has been limited in some way. The first publication underlying the *Predikin* tool for example was primarily a structural analysis but was based on only three different kinase-substrate structure available at the time (Brinkworth et al., 2003). It is for this reason that descriptions are given only for substrate positions -3 to +3 but not for more distal substrate positions. Also, kinase SDRs were assigned to substrate positions and not to substrate preferences *per se*, which is a likely simplification of the actual kinase structure-specificity relationship as specificities at the same position (e.g. R-2 and P-2) may be determined by different SDRs (see *Results* section). Finally, the study relied upon a structural analysis without recourse to kinase multiple sequence alignments (MSAs), and so the extent to which the inferences made apply to kinases without structural representation in the original study is not clear.

MSA-based analyses were generally performed more recently. In the (Mok et al., 2010) study, a covariation analysis between *S. cerevisiae* kinase sequence and specificity was used to predict SDRs for several preferences. However, there was only limited structural analysis to distinguish likely true positive and false positive predictions, and in most cases a structural mechanism was not proposed for the SDRs suggested. The analysis was also limited to a single model organism (*S. cerevisiae*). A similar approach for human kinases was performed also in 2015 but the SDRs predicted were not assigned to substrate positions and there was limited discussion of the structural basis for specificity (Creixell et al., 2015a). Some examples do exist of studies in which the predicted SDR are cross-referenced with the available structural data. However, these projects are designed to identify residues that define a particular *Group*/*Family*/*Subfamily* rather than specificity determinants *per se* (Kannan et al., 2007a; Kannan and Neuwald, 2004; Li et al., 2003; Mohanty et al., 2016).

Here I attempt to provide the most comprehensive analysis to date of protein kinase peptide specificity by considering explicitly both kinase sequence variation and kinase structure. To this end I have exploited most of the available data relating to kinase structure and known kinase target sites. The objective of this study was to relate the predicted kinase SDRs to particular substrate preferences (R-3, P+1, etc) and not just to particular substrate positions (-3, +1, etc). Following from this, I place a particularly strong emphasis also on proposing structural mechanisms to account for the SDRs predicted.

As an application of these results, I use the putative SDRs to generate sequence-based predictors of kinase specificity. As discussed in the *Introduction* chapter, such tools can be leveraged towards the evolutionary analysis of kinase specificity, and this is explored further in *Chapter 3*. For this chapter I predict the specificities of human kinases only and use these results to understand the differential mutation of residues in cancer between the two most common preferences (P+1 and R-3). I also use the detailed analysis of SDRs to form the basis of a more general analysis of SDR mutations in cancer, following on from previous publications concerning kinase domain mutations and cancer progression (Creixell et al., 2017; Dixit et al., 2009).

2.2 Overview of protein kinase peptide specificity

2.2.1 Protein kinase specificity models

Protein kinase specificities at the active site are modelled here in the form of position probability matrices (PPMs). Each value in the 20 x 11 matrix represents the empirical probability of finding a particular amino acid (e.g. proline) at a particular position (e.g. +1) relative to the phosphoacceptor serine, threonine, or tyrosine residue. For this analysis, I generated specificity models for human, mouse, and *S. cerevisiae* protein kinases on the criterion that each PPM must be supported by at least 10 non-redundant and experimentally-verified target phosphorylation sites for a given kinase. I do not include protein kinases defined as 'Atypical' in the Manning classification of protein kinases, which have little to no sequence similarity to the eukaryotic protein kinase (ePK) domain (Manning et al., 2002b).

I generated 179 PPMs on this basis, representing 9,005 unique kinase-phosphosite relations in total. For each species studied, this represents a small proportion of the total annotated kinome (human: 126/478, mouse: 35/504, *S. cerevisiae*: 18/116). For further analysis, I selected high-confidence PPMs that could successfully discriminate between target and non-target phosphorylation sites during 10-fold cross-validation (see *Methods* chapter Section 6.1.8). This resulted in 136 PPMs in total to be used for all of the analysis discussed below (88 human, 30 mouse, 18 *S. cerevisiae*). Among the remaining serine/threonine models, I observed a strong over-representation of kinases belonging to either the AGC, CAMK, or CMGC *Groups*, as these *Groups* represent 78% of the serine/threonine models generated but only ~49% of annotated serine/threonine kinases (Manning et al., 2002b).

The extent of kinase selectivity per substrate position can be summarised by calculating the information content in units of bits (Figure 2.1). For serine/threonine kinases, consistent evidence of active site selectivity is only apparent for the -3 and +1 positions, and to a lesser

extent the -2 position. The -3 and -2 constraints are generally only evident in the non-CMGC kinases and +1 constraint only generally in CMGC kinases. These constraints correspond mainly to the well-established preferences for basic side chains (R or K) at the -3 and/or -2 position, and in most CMGC kinases for proline at the +1 position. For specific amino acid preferences outside these well-known cases, I find that the selectivity is either less stringent than for R/K-3 and P+1 and/or less pervasive across the kinome, usually in that the specificity is restricted to a particular kinase *Family* or *Subfamily*. Such instances are discussed in detail in the following sections.

For the tyrosine kinases there is no evidence in this dataset of a single position for which the substrate amino acids are generally constrained. This is consistent with systematic analyses demonstrating a lower average performance for tyrosine kinase specificity models (relative to serine/threonine models) when benchmarked (Miller et al., 2008; Saunders et al., 2008; Wagih et al., 2015).

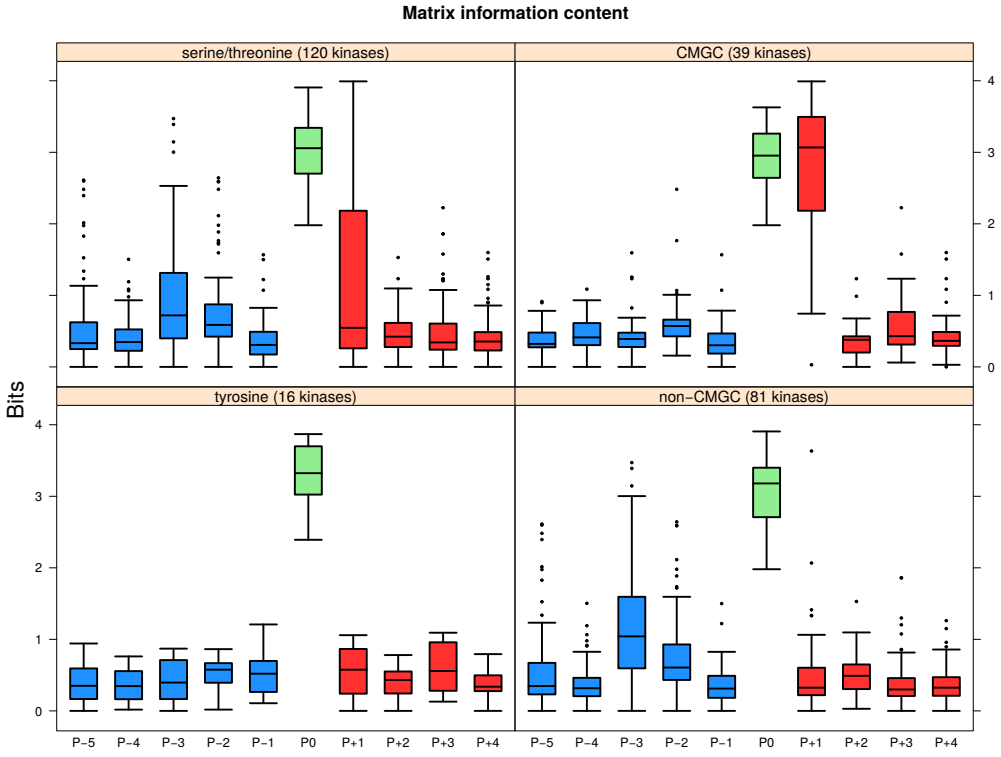


Fig. 2.1 Sequence constraint at substrate positions -5 to +4 is represented for all models of serine/threonine kinases (top-left), tyrosine kinases (bottom-left), CMGC S/T kinases (top-right), and non-CMGC S/T kinases (bottom-right). Sequence constraint is represented in terms of the matrix information content (unit bits).

2.2.2 Kinase-substrate interface at the active site

Retrieval of kinase-substrate complexes from the Protein Data Bank

Kinase-substrate cocrystal structures are an important resource for the inference of specificity-determining residues, as they provide direct physical evidence for residue-level kinase-substrate interactions. Here I have automated the retrieval of kinase-peptide complex structures at the active site from the protein data bank (PDB). The results for serine/threonine kinases and tyrosine kinases are listed in Tables 2.1 and 2.2, respectively.

In total I identified 9 different serine/threonine kinases with kinase-peptide co-crystal structures, and 7 tyrosine kinases, where kinase orthologues and close paralogues (e.g. AKT1 and AKT2) are counted once only. These numbers are close to what can be derived from the *ProtCID* database of common protein-protein interfaces (Xu and Dunbrack, 2011; Xu et al., 2015). I also identified cocrystal structures for the EGFR tyrosine kinase and for the SYK tyrosine kinase, in addition to the structures listed in *ProtCID*.

I also sought to automate the retrieval of inhibitory interactions at the active site that mimic substrate binding; such interactions are found for ABL1, CAMKII, INSR, MARK2, PAK4, and PKA cocrystal structures. For substrates and inhibitors, the P0 position and flanking residues (-5 to +5) are identified automatically here by considering the residue closest to the kinase HRD catalytic aspartate as 0. While useful, kinase-inhibitor structures should be interpreted with caution, however, as inhibitor binding can not always be expected to perfectly mimic substrate binding. In Pak4 kinase, for example, where both kinase-substrate and kinase-inhibitor structures are available, equivalent residues identified by structural superposition are offset by a single residue (inhibitor -3 position equivalent to peptide -2 position) (Ha et al., 2012).

Table 2.3 lists co-crystal structures where the substrate is greater than 35 amino acids in length. I detected only two different kinases (PKA and brassinosteroid insensitive 1-associated receptor kinase) where this is the case, in accordance with *ProtCID* data. I also detected an additional protein substrate (PRKAR1A) co-crystal structure for PKA. Two structures in which the kinase is in complex with a protein mimetic inhibitor (PKA-PRKA2B; EGFR-ERBB1) were also identified. With respect to kinase autophosphorylation, I used the structures identified from a comprehensive survey of autophosphorylation published in 2015 (Xu et al., 2015). In addition to these, I identified auto-inhibitory complexes for the putative orthologues of protein kinase G in *Plasmodium vivax* from a systematic query (Table 2.3).

Survey of kinase-substrate contacts

I next identified for each serine/threonine kinase-substrate structure the residues in the protein kinase in contact with the substrate. The residues identified were then mapped to the protein kinase domain and the results across all structural models aggregated (Figure 2.2a). Each kinase residue at the kinase-substrate interface was assigned to a particular substrate position by taking the substrate residue closest to the catalytic aspartate (HRD) as the 0 position. The results show that each substrate position can be bound by several different kinase residues, contrary to a previous study suggesting a more limited range of contacts between the kinase and substrate (Brinkworth et al., 2003).

Results for position +1 support the idea of a dedicated binding pocket (P+1 pocket), with most contacts occurring with kinase domain positions 157, 158, and 159 of the P+1 loop (Figure 2.2a). To a lesser extent, contacts also exist with the 161 and 164 positions previously implicated as important for proline specificity at position +1 (Kannan and Neuwald, 2004). Finally, ~40% of structures are contacted by the 'DFG+1' residue (domain position 144) that has been suggested previously as a +1 specificity determinant (Brinkworth et al., 2003; Howard et al., 2014; Kannan and Neuwald, 2004).

Positions +2 and +3 are bound most frequently by four residues of the activation segment (kinase domain positions 156-159). Position +2 however is also frequently bound by position 11 of the N-terminal glycine-rich loop, position 42 of the αC helix, and the 'DFG+1' residue mentioned above. The binding profile at position +4 seems more broadly distributed, although it should be emphasised that there are only 6 unique kinase-peptide complexes where position +4 is bound (Table 2.1)

Position -1 is contacted frequently by positions 160 and 161 of the P+1 loop although these represent non-specific backbone contacts. Sidechain contacts however do occur more frequently with position 125 of the **K**-x-x-N motif. Position -2 is frequently in contact with domain position 160/162 of the P+1 loop, and positions 125/127 of the **K**-x-**x**-N motif. Positions 125 and 160 however are unlikely to serve as an SDRs as they are highly conserved within the kinase superfamily.

Position -3 contacts are most common for domain positions 84, 86, and 87 on the αD helix and the preceding loop. The aforementioned 127 position of the K-x-**x**-N motif however is also a frequent binder. For substrate positions -4 and -5, position 86 of the αD helix binds most frequently. However, residues on the αF helix and the αF-αG loop are also common binders at position -5, and to a lesser extent at position -2 also.

Some of the most common binding residues in the kinase domain are represented structurally in Figure 2.2b.

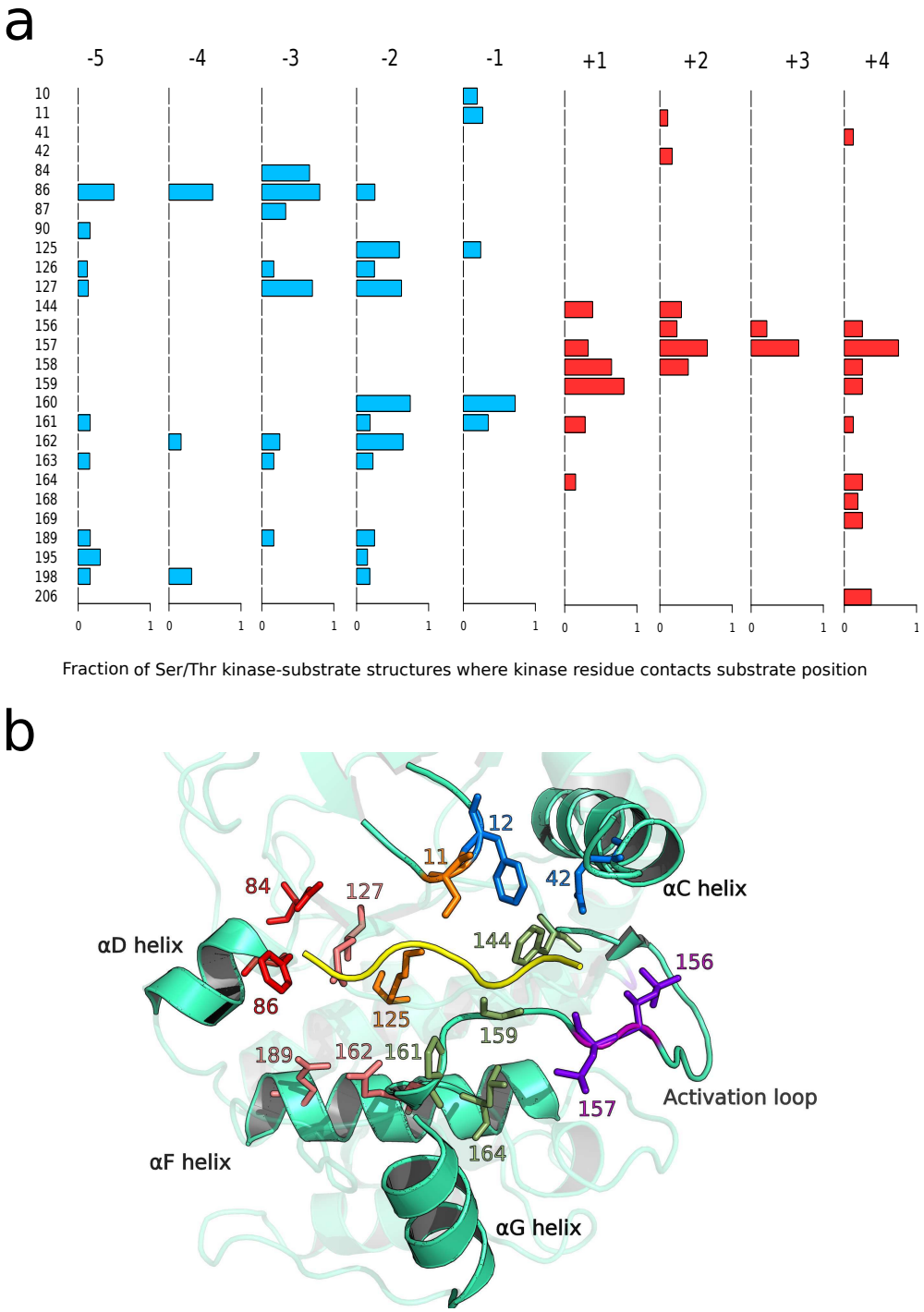


Fig. 2.2 a) Binding profile of Ser/Thr kinases for substrate positions -5 to +4. The numbering refers to the kinase *Pfam* domain (PF00069) position . b) Structural representation of some kinase domain positions most often in contact with the substrate *PDB: 1ATP*. -3: red, -2: pink, -1: orange, +1: green, +2: blue, +3: purple

PDB ID	Kinase	Group	Kinase species	Substrate	Resolution (Å)	-5	-4	-3	-2	-1	0	1	2	3	4	5
2CPK	PKA	AGC	Human	PKI-α^*	2.7	T	G	R	R	N	**A**	I	H	D	x	x
4O21	PKA	AGC	*M. musculus*	PKI-α	1.95	T	G	R	R	A	**S**	I	H	D	x	x
4WIH	PKA	AGC	*C. griseus*	PKI-α^*	1.1	T	G	R	R	Q	**A**	I	H	D	I	x
4XW5	PKA	AGC	*M. musculus*	PKI-α^*	1.82	T	G	R	R	A	**C**	I	H	D	x	x
3O7L	PKA	AGC	*M. musculus*	Pln	2.8	A	I	R	R	A	**S**	T	I	x	x	x
2PHK	PHKg	CAMK	*O. cuniculus*	Synthetic	2.6	x	x	R	Q	M	**S**	F	R	L	x	x
1QMZ	CDK2	CMGC	Human	Synthetic	2.2	x	x	H	H	A	**S**	P	R	K	x	x
3QHR	CDK2	CMGC	Human	Synthetic	2.2	x	x	x	P	K	**T**	P	K	K	A	K
1O6K	AKT2	AGC	Human	GSK3-β	1.7	R	P	R	T	T	**S**	F	A	E	x	x
3CQW	AKT1	AGC	Human	Synthetic	2.0	R	P	R	T	T	**S**	F	A	E	x	x
3MVH	v-AKT	AGC	Human	Synthetic	2.01	R	P	R	T	T	**S**	F	A	E	x	x
2C3I	PIM-1	CAMK	Human	Synthetic	1.9	R	R	R	H	P	**S**	G	x	x	x	x
4DC2	PKC-i	AGC	*M. musculus*	Par-3	2.4	G	F	G	R	Q	**S**	M	S	x	x	x
2WO6	DYRK-1A	CMGC	Human	Crb-2	2.5	x	A	R	P	G	**T**	P	A	L	x	x
4JDH	PAK-4	STE	Human	Synthetic	2.0	x	x	R	R	R	**T**	W	Y	F	G	G
4L67	PAK-4	STE	Human	Pak4*	2.8	A	R	R	P	K	**P**	L	V	D	P	A
4OUC	Haspin	Other	Human	Histone H3.2	1.9	x	x	x	A	R	**T**	K	Q	T	A	x
3KL8	CAMKII	CAMK	*C. elegans*	CAMK2n1*	3.37	I	G	R	S	K	**R**	V	V	I	x	x
3IEC	MARK2	CAMK	Human	cagA*	2.2	L	K	R	H	D	**K**	V	D	D	L	S

Table 2.1 Table of unique kinase-peptide models in the protein data bank (PDB) for serine/threonine kinases. The PDB entry associated with the earliest publication is shown in each case. Inhibitor peptides are marked with an asterisk.

PDB ID	Kinase	Family	Kinase species	Substrate	Resolution (Å)	-5	-4	-3	-2	-1	0	1	2	3	4	5
1IR3	INSR	InsR	Human	Synthetic	1.9	x	x	x	G	D	**Y**	M	N	M	x	x
1GAG	INSR	InsR	Human	Synthetic*	2.7	P	A	T	G	D	**F**	M	N	M	S	P
3BU3	INSR	InsR	Human	IRS2	1.65	P	Y	P	E	D	**Y**	G	D	Y	E	Y
1K3A	IGF1R	InsR	Human	IRS1	2.1	x	x	x	G	E	**Y**	V	N	I	E	F
2G1T	ABL1	Abl	Human	Synthetic*	1.8	x	x	x	E	I	**F**	G	E	F	E	A
2G2F	ABL1	Abl	Human	Synthetic*	2.7	x	x	E	A	I	**F**	A	A	P	F	x
2GS6	EGFR	EGFR	Human	Synthetic	2.6	x	x	x	E	I	**Y**	G	E	x	x	x
4R3P	EGFR	EGFR	Human	ERRFI1	3.25	x	x	x	T	H	**Y**	Y	L	L	P	x
5CZH	EGFR	EGFR	Human	Synthetic	2.90	x	D	E	E	D	**Y**	Y	E	I	P	x
5CZI	EGFR	EGFR	Human	SHC1	2.60	x	x	D	H	Q	**Y**	Y	N	D	x	x
2PVF	FGFR2	FGFR	Human	FGFR2	1.8	x	x	x	E	E	**Y**	L	x	x	x	x
3CBL	Fes/Fps	Fer	Human	Synthetic	1.75	x	x	x	x	I	**Y**	E	S	L	x	x
3FXX	EPHA3	Eph	Human	Synthetic	1.7	x	Q	W	D	N	**Y**	E	Y	I	w	x
5C26	SYK	Syk	Human	Synthetic	1.95	x	x	x	E	V	**Y**	E	S	P	x	x

Table 2.2 Table of unique kinase-peptide models in the protein data bank (PDB) for tyrosine kinases. The PDB entry associated with the earliest publication is shown in each case. Inhibitor peptides are marked with an asterisk.

PDB ID	Kinase	Group	Kinase species	Substrate	Resolution (Å)	-5	-4	-3	-2	-1	0	1	2	3	4	5
2QCS	PKA	AGC	*M. musculus*	PRKAR1A*	2.2	R	R	R	R	G	**A**	I	S	A	E	V
2QVS	PKA	AGC	*M. musculus*	PRKAR2A	2.5	x	T	R	R	V	**S**	V	C	A	E	T
3IDB	PKA	AGC	*M. musculus*	PRKAR2B	2.62	x	T	R	R	V	**S**	V	C	A	E	A
4DIN	PKA	AGC	*M. musculus*	PRKAR2B*	3.7	R	R	R	R	G	**G**	V	S	A	E	V
3TL8	BARK1	TKL	*A. thaliana*	HopAB2	2.5	I	D	L	G	E	**S**	L	V	Q	H	P
4JZV	EGFR	Tyrosine	Human	ERBB1*	2.7	K	V	C	S	T	**H**	Y	Y	L	L	P
4RZ7	PKG (putative)	AGC	*P. vivax*	Autoinhibition*	2.35	R	N	K	K	K	**A**	I	F	G	E	D
5DZC	PKG (putative)	AGC	*P. vivax*	Autoinhibition*	2.30	N	E	K	K	K	**A**	I	F	S	N	S

Table 2.3 Table of unique kinase-substrate models in the protein data bank (PDB) for substrates longer than 35 amino acids. The PDB entry associated with the earliest publication is shown in each case. Inhibitor peptides are marked with an asterisk.

2.2.3 MSA-based inference of kinase SDRs

I sought to complement the direct structural evidence discussed above with MSA-based insights into kinase specificity. To this end, I implemented a pipeline for the automated detection of putative kinase SDRs for each amino acid preference at the substrate positions flanking the phosphoacceptor residue. Each substrate position (-5 to +5) was considered in turn and the major specificities (e.g R/K at position -3) were identified by generating clusters from the corresponding PPM columns (across all PPMs generated). For each preference identified, the MSA of 119 kinases of known specificity was divided into a positive group (all kinases with the specificity of interest) and negative group (all kinases without the specificity of interest). Those alignment positions that best discriminated between the two specificity groups were then implicated as SDRs. This method is represented in Figure 2.3.

A previous study has shown that the mode of substrate binding differs between serine/threonine and tyrosine kinases; the -1 substrate position for tyrosine kinases binds to what would be the -3 pocket in serine/threonine kinases, for example (Brinkworth et al., 2003). I therefore intended originally to perform this analysis separately for the serine/threonine and tyrosine kinases. However, there were too few tyrosine kinase PPMs (16) for the reliable detection of SDRs, and so the analysis was applied to the serine/threonine kinases only.

The results of this analysis are summarised in Figure 2.4. I identified 30 putative SDRs across 16 different preferences. Notably, I find a significant over-representation of SDRs among residues close to the substrate peptide (10/30, Fisher $p < 0.01$). Many of these had been described previously and are discussed in more detail below. In some cases I associate a known SDR with a new substrate preference. Position 189 for example was previously identified as an R-5/R-2 determinant but here it emerges as a determinant also of L-5 specificity. There are also some cases where I assign an SDR to a new substrate position; kinase position 164 for example is considered to be a +1 determinant usually but here I predict it to be a determinant of leucine specificity at position +4 also. I also predict a number of putative distal SDRs that have not been described previously, which are discussed in more detail in the section below.

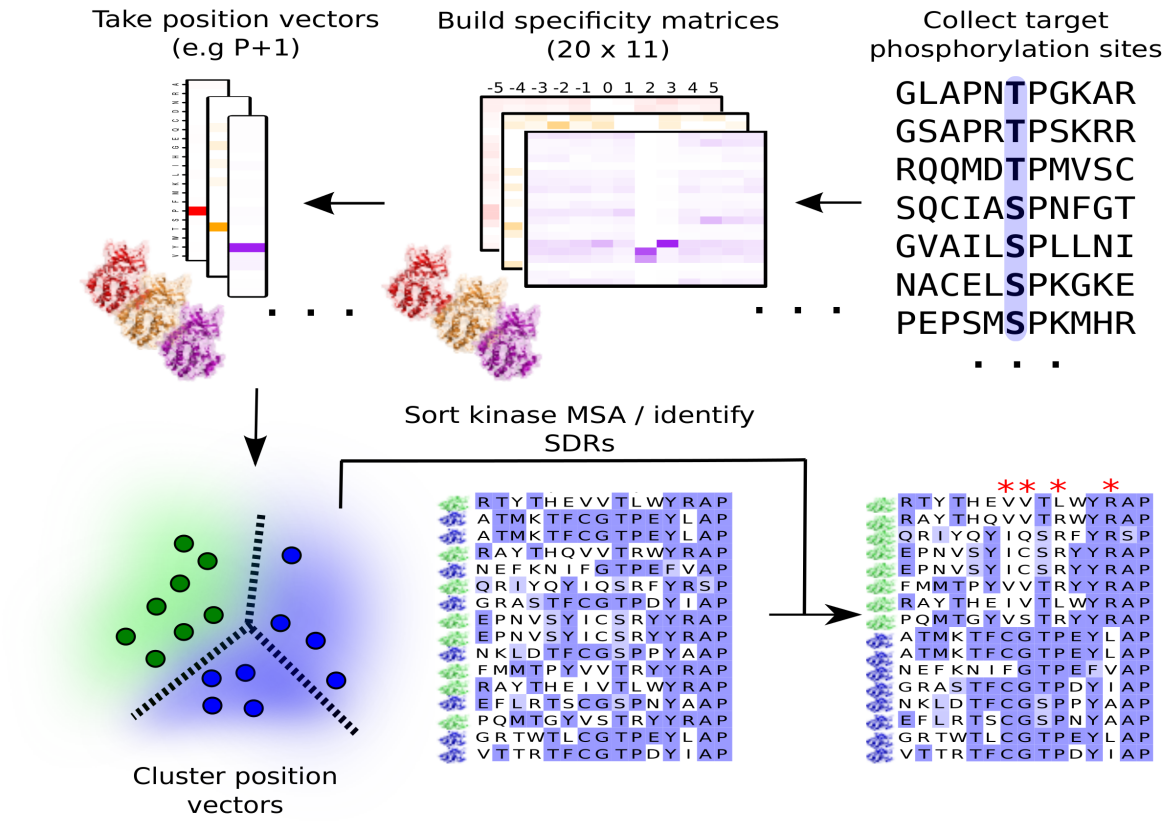


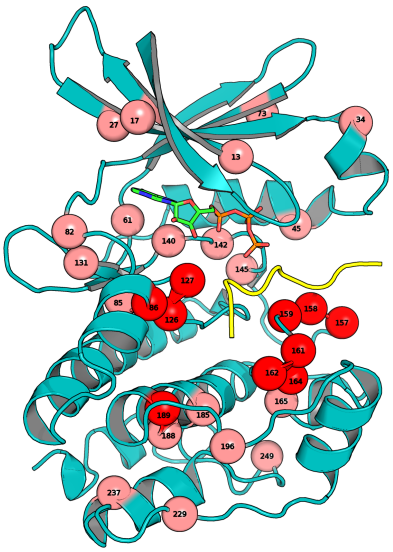
Fig. 2.3 Process for SDR detection from kinase domain multiple sequence alignments. Specificity matrices are first constructed for all kinases with at least 10 known target sites. The pipeline then iterates through each column of the matrix (-5, -4, -3, etc) and, for each position, clusters the columns to identify kinases with similar specificities at that position (e.g. for proline at +1). The pipeline then iterates through each of the clusters (e.g. proline+1, aspartate/glutamate+1, etc.) and, for each iteration, sorts the kinase domain MSA into two groups: kinases with the specificity of interest (e.g. proline+1) and those without. Those positions that best discriminate between the two sequence groups are predicted as SDRs (represented by a red asterisk)

.

a

Preference	SDR positions
Arg-3 (55)	17, 82, 86, 127, 158, 162, 185
Pro+1 (36)	158, 159, 161, 164, 188, 196
Arg-2 (27)	27, 162
Pro-2 (25)	82, 161, 162, 188, 196
Leu-5 (21)	86, 189
Arg/lys+2 (14)	45, 61, 126, 229
Asp/glu-2 (13)	157, 189
Asp/glu-3 (13)	86, 127, 140, 157
Arg-5 (12)	162
Arg/lys+3 (10)	161, 237
Arg-5 (12)	162
Asp/glu+1	85, 249
Leu-2 (8)	131
Asp/glu+2 (7)	13, 34
Pro+2 (7)	145
Leu+4 (6)	164
Asp/glu+4 (6)	73, 142, 165, 249

c



b

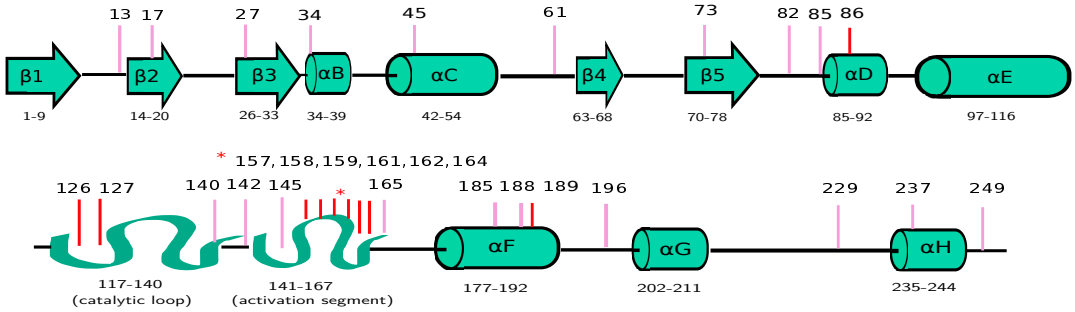


Fig. 2.4 a) Table of the SDRs predicted using the alignment-based procedure represented in Figure 2.3 b) Mapping of the results to a 1D-representation of the protein kinase domain. c) SDRs mapped to the a structure of protein kinase A. SDRs within 4 Å of the peptide substrate are coloured in red. *PDB: 1ATP*

2.3 Analysis of protein kinase SDRs

In the following section I make use of available kinase-substrate PDB models to rationalise the determinants that have been discovered. For some specificities not represented in Tables 2.1, 2.2, or 2.3, I generated models of the kinase-substrate interface using the method described in the *Introduction* (Section 1.4.2) and *Methods* chapters (Section 6.1.7). I also repeatedly cross-reference the results here with previous kinase SDR analyses in the literature. In the subsection headings, I state whether the inferences made derive from an alignment-based analysis or purely from a structure-based analysis or both. Results that are of only minor significance have been placed in the *Appendix*.

2.3.1 Position +1

Proline (alignment and structure-based)

I identified 6 SDRs for the proline+1 specificity, 5 of which had been predicted previously (Kannan and Neuwald, 2004). Four of the six residues I identified as SDRs map to the kinase P+1 pocket (Figure 2.5). Of these four, the results suggest that domain positions 159 and 164 are the most discriminating. The role of these two residues in proline-directed kinases was explained in the *Introduction* chapter.

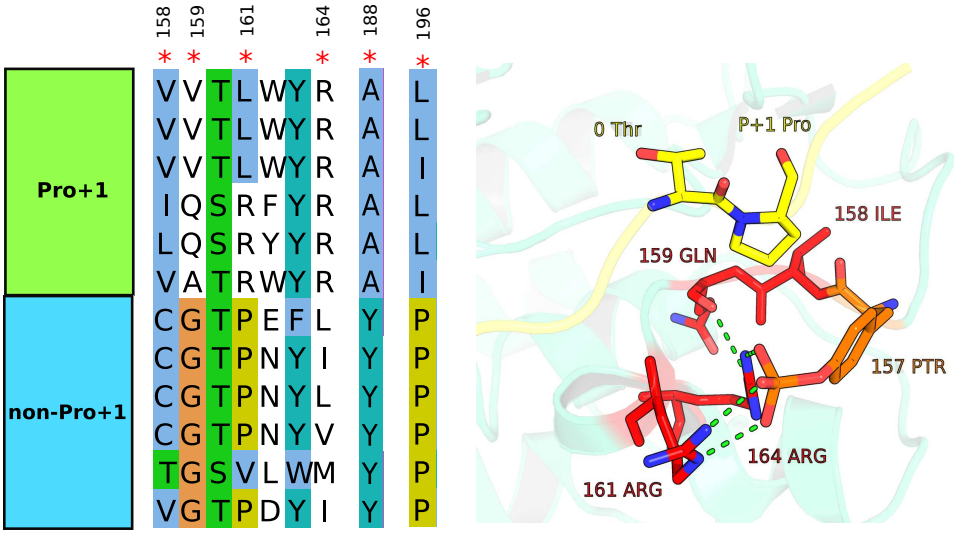


Fig. 2.5 The predicted SDRs for proline at the +1 position are marked by a red asterisks and coloured in red in the structure (right). Position 157 was not predicted to be an SDR but is discussed in the main text. *PDB: 2WO6*

At position 158, I observe valine and leucine/isoleucine exclusively in proline-directed kinases whereas in non- proline-directed kinases I observe mainly cysteine, but also alanine,

isoleucine, and valine. The residue side chain does not project towards the +1 position and is therefore unlikely to be a direct determinant of specificity. It does however participate in a network of contacts in the activation segment, involving also positions 122 (H**R**D), 168 (APE**I**) and position 174 (Y). I therefore suggest a role for position 158 in the CMGC-specific stabilisation of the activation segment in its active conformation. Consistent with this, *in silico* mutation of residue 158 from valine to cysteine is predicted to destabilise the kinase-peptide complex ($\Delta\Delta G = +1.93$)

At position 161, I observe mainly proline in the non- proline-directed kinases but arginine (non-CDK *Families*) and leucine (CDK *Family*) in the proline-directed kinases. The arginine residue is likely important for binding the second activation loop phosphorylation that can occur in CMGC kinases (Kannan and Neuwald, 2004; Soundararajan et al., 2013; Varjosalo et al., 2013), which functions to fully activate the CMGC kinase (Prowse et al., 2001). This does not however account for the role of leucine in the CDK kinases, which do not generally have a second activation loop phosphorylation. In CDKs, a 'phosphomimetic' glutamate at position 202 (αG helix) instead projects towards the P+1 pocket (Cheng et al., 2006) and binds to the arginine at position 164 (Figure 2.6). I note from the MSA that, in the absence of a phosphorylatable tyrosine at 157, that position 157 is either a glutamate/aspartate or position 202 is, suggesting that acid-mediated stabilisation of the P+1 pocket is a required feature for proline specificity.

At position 188 I observe mainly alanine in the proline-directed kinases and tyrosine otherwise, and at position 196 mainly leucine in the proline-directed kinases and proline otherwise. Position 188 is found on the αF helix and position 196 on the flexible loop connecting the αF helix with the αG helix (Figure 2.6). Both positions are distal from the +1 substrate site. It has been previously suggested that 188 and 196 – among several other functionally divergent residues in the αF to αG region – are important for connecting the substrate-binding residues with a CMGC-unique region ('CMGC insert') that has been implicated in scaffold/adaptor binding (Bax et al., 2001; Dajani et al., 2003; Oruganty and Kannan, 2012). I note here that the positioning of the αG helix in CMGC differs from that of non-CMGC in kinases, in that the N-terminus of the helix in the former is tilted more closely towards the P+1 pocket. As a consequence, the domain position at 202 discussed above is in closer proximity to the critical determinant at 164. It is possible that the presence of divergent residues in the α-F to α-G region contribute to the repositioning of the α-G helix, suggesting one possible factor linking the distal SDRs detected (188 and 196) with the proline+1 preference.

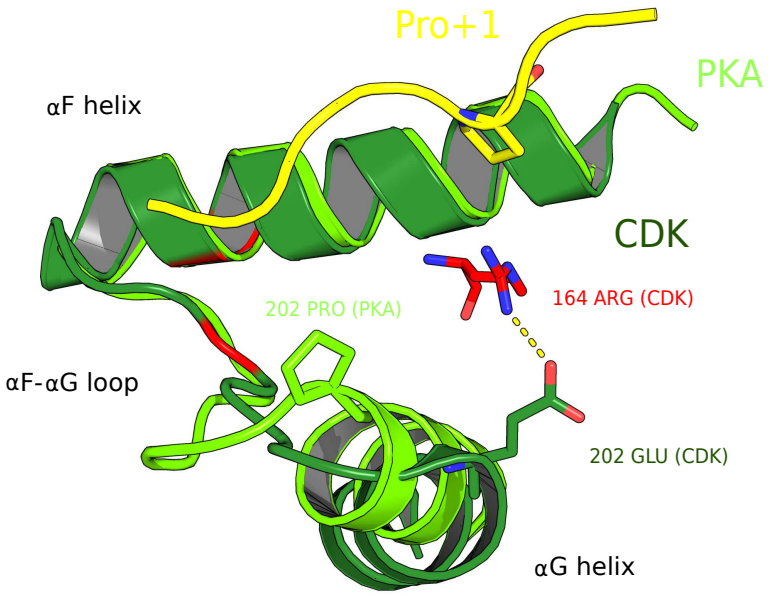


Fig. 2.6 The 188 and 196 positions (coloured red) are found on the αF helix and αF-αG loop, respectively. They may influence the positioning of the αG helix, which contains at least one substrate-binding position (202). *PDB: 2CCI (CDK) and PDB: 1ATP (PKA)*

Aspartate/glutamate (structure-based)

I identified eight kinases with a preference for aspartate and/or glutamate at the +1 position in the dataset. Of these eight, five belong to the acidophilic casein kinase 2 (CK2) *Family*, with varying levels of aspartate/glutamate preference spanning from the -4 to +4 positions of the substrate peptide (Kuenzel et al., 1987; Songyang et al., 1996). Here, I generated a kinase-subsrate model that implicates a lysine at position 164 as the primary substrate determinant (Figure 2.7). This is in agreement with experimental data and previously generated CK2-substrate models (Niefind et al., 2007; Sarno et al., 1997).

The kinase ADRBK1 (GRK2) was also found to be acidophilic for the +1 position. This kinase belongs to the GRK (G-protein coupled receptor kinase) *Family* of the AGC *Group*, and is distantly related to the CK2 kinases. A kinase-substrate model generated here for ADRBK1 however suggests a lysine at position 202 (αD helix) as the primary substrate determinant (Figure 2.7). These results therefore suggest that the D/E+1 preference evolved by different mechanisms in the two different *Families*.

Domain positions 85 and 249 were inferred as putative SDRs. Both however are distal from the P+1 binding pocket and no obvious structural mechanism that could link these sites to +1 selectivity was observed. These sites may therefore represent either indirect determinants of specificity or false positive identifications.

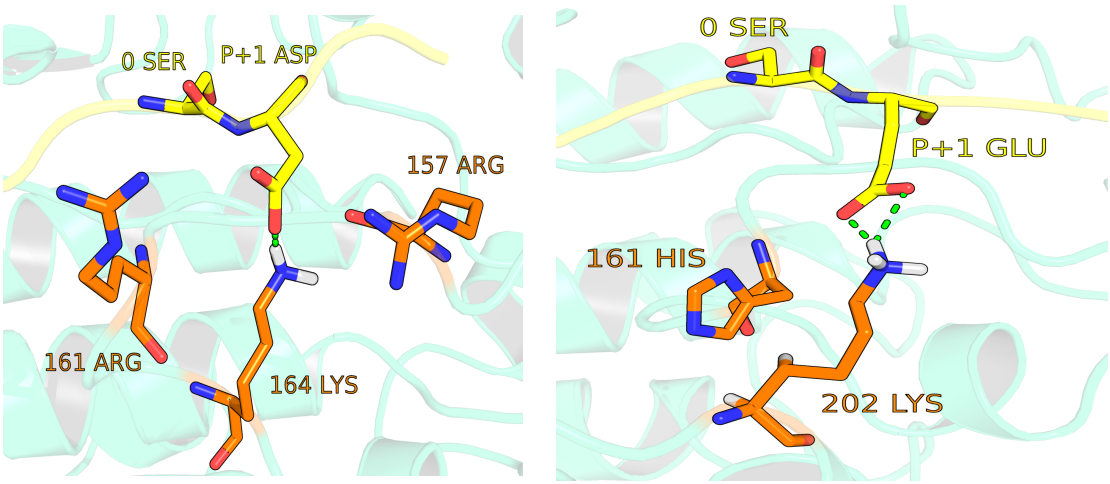


Fig. 2.7 The D/E+1 specificity for CK2 kinases (left) and the ADRBK1 (right) kinase is likely to be determined by different kinase domain positions. The kinase-substrate interfaces here were modelled as part of this analysis.

Arginine (structure-based)

I detected a single protein kinase (SRPK1) with a significant preference for arginine at the +1 position. This kinase belongs to the SRPK *Family* of the CMGC *Group*. The generation of a kinase-substrate model here places the +1 arginine in the P+1 pocket, and in specific contact with an aspartate at position 157. The placement of the +1 residue in the P+1 pocket is in agreement with a previous model generated for the *S. cerevisiae* homologue Sky1p (Nolen et al., 2001). However, it challenges a different model that places the +1 residue outside of the +1 pocket and proximal the DFG+1 residue, which was the suggested SDR (Kannan and Neuwald, 2004). The mode of determination of this specificity is therefore still contested.

Glycine (structure-based)

I observed a single kinase (PRK1) in the dataset that is highly selective (96%) for glycine at the +1 position. PRK1 is a yeast kinase of the NAK *Family* ('Other' *Group*). All yeast kinases of the NAK *Family* (AKL1, ARK1, PRK1) exhibit strong glycine specificity when tested against degenerate peptide libraries (Mok et al., 2010). All human NAK kinases contain an α-helical insert towards the C-terminus of the activation segment (named ASCH) (Sorrell et al., 2016), a feature not yet observed in any kinase outside the NAK *Family*. Two different kinase-substrate models in the literature suggest different consequences for peptide binding. The first suggested that the peptide undergoes a half-turn about +1 glycine so that the residues C-terminal to it are directed towards the hydrophobic groove between ASCH and the α-FG helices (Eswaran et al., 2008). The second model however places the substrate

peptide C-terminus 'above' the ASCH in an extended linear conformation (Chaikuad et al., 2014).

The kinase-substrate model generated here agrees with the former model by suggesting that the substrate peptide forms a kink at the +1 position (Figure 2.8a). This is likely enabled by the role of glycine as a frequent Ramachandran outlier. In particular, the model generated implicates arginine at 164 and threonine at 159 as G+1 specificity determinants (Figure 2.8b).

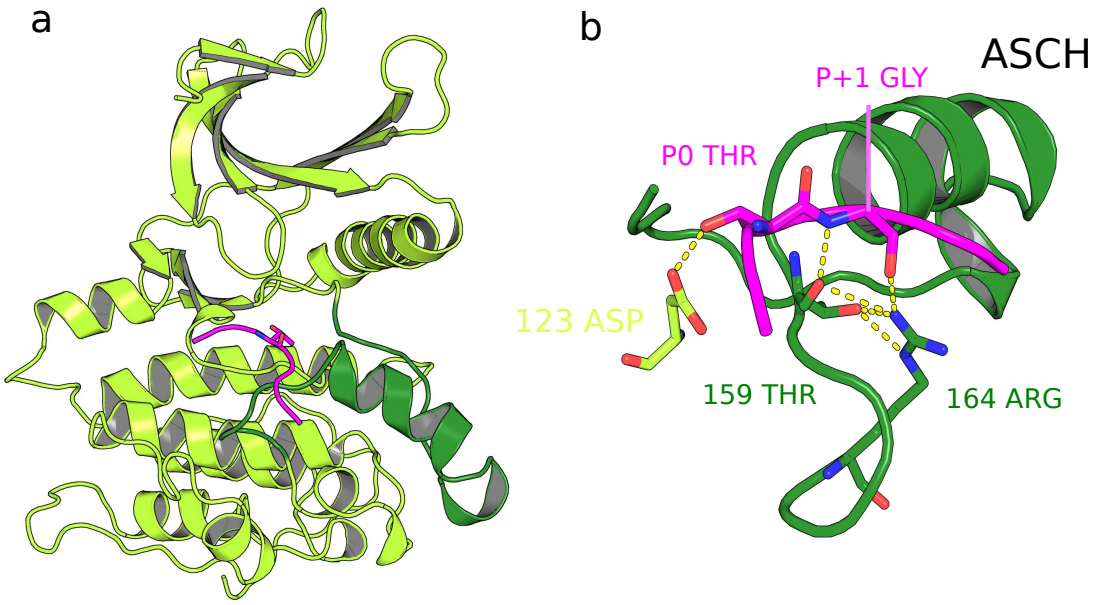


Fig. 2.8 a) The NAK *Family* of kinases contains an activation segment C-terminal helix (ASCH) in the activation segment (dark green) and is highly specific for glycine at the +1 position. The model generated here suggests that the substrate peptide does not bind in the extended linear conformation; b) kinase positions 159 and 164 are suggested as specificity determinants.

2.3.2 Position +2

Arginine (alignment and structure-based)

Of the 14 kinases I identified with an arginine/lysine +2 preference, 10 belong to the protein kinase C *Family* (AGC *Group*). The alignment-based analysis performed here implicates domain position 45 as a potential determinant, which is present mainly as a cysteine in R+2 kinases and as a polar/charged residue otherwise (particularly arginine/lysine). This residue projects towards the αB-αC loop and may serve as a negative determinant of arginine binding

when present as an arginine or histidine by attenuating the electronegative charge of the αB-αC loop. This loop has been implicated previously as an R+2 substrate determinant (Creixell et al., 2015a; Li et al., 2003; Wang et al., 2012a) and is again here for the kinase-peptide model generated (Figure 2.9). The other three domain positions (61, 126, and 229) inferred from the multiple sequence alignment (MSA) are unlikely to be direct determinants given their distal positions in the protein kinase domain (ATP-binding site, P-2 subsite, and C-terminal tail, respectively).

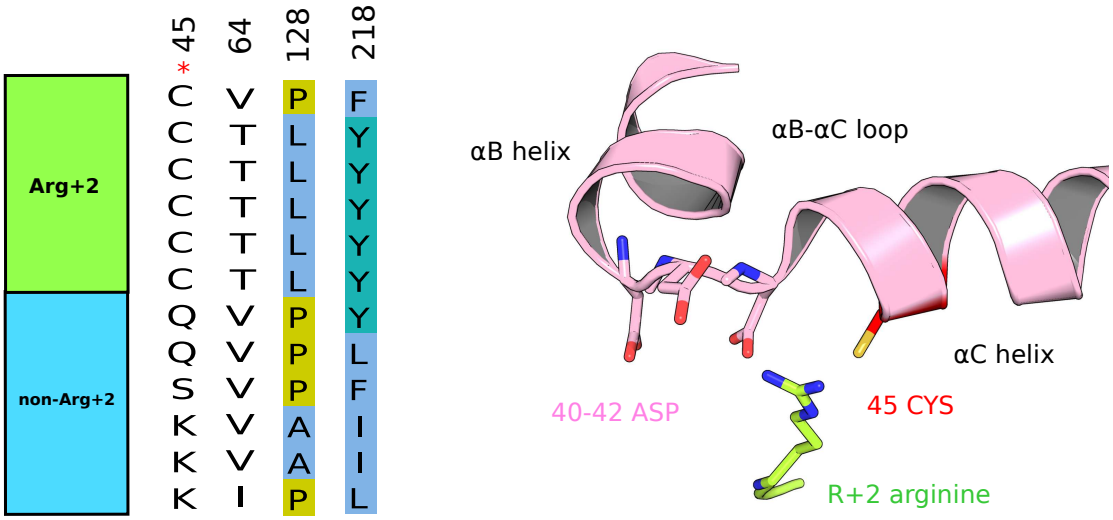


Fig. 2.9 I identified one residue (position 45) in the αC helix as a possible R+2 determinant. Previous studies have also implicated the aspartate residues at positions 40, 41, and 42

Aspartate/glutamate (structure-based)

I identified seven kinases here that are acidophilic at the +2 position. Four (human CSNK2A1, human CSNK2A2, *S. cerevisiae* CKA1, *S. cerevisiae* CKA2) belong to the casein kinase II *Family*, two to the CAMK2 *Family* (human CAMK2A and CAMK2D), and one to the Polo-like kinase *Family* (human PLK3).

A CAMK2 co-crystal structure for the *D. melanogaster* orthologue supports the role of R41 and K44 as D/E+2 determinants in this *Family* (*PDB: 5H9B*, unpublished). Both residues are located in the N-terminal lobe. Notably, for a PLK3-substrate model constructed here, I also find a lysine at domain position 44 that co-ordinates D at the +2 position. Both kinase *Families* are distantly related and so it appears that the kinases have converged upon a similar mechanism to determine the D/E+2 preference.

Although I find some preference for D/E+2 among the CK2 kinases, the +1 and +3 positions are thought to be the major determinants of specificity, with the other flanking

positions playing more subsidiary roles (Kuenzel et al., 1987; Niefind et al., 2007; Sarno et al., 1997). In accordance with this, I do not observe a clear candidate SDR from the CSNK2A1-substrate model constructed here.

2.3.3 Position +3

Arginine (structure-based)

SRPKs exhibits a strong preference for arginine at this position (Wang et al., 1998). The SRPK1 model that I have constructed suggests that an aspartate at domain position 47 forms a hydrogen bond with the arginine side chain at position +3. A glutamate at position 43 does not bind directly to the arginine but may contribute to selectivity by strengthening the negative charge of the +3 pocket.

Aspartate/glutamate (structure-based)

The acidophilic constraint observed for peptide substrates of casein kinases II are thought to be the most stringent at the +3 position. The model constructed here suggests that a lysine at domain position 44 and an arginine at position 47 form hydrogen bonds with the aspartate side chain. This is consistent with previously-constructed model and with the results of substrate peptide assays on mutant (K44A, R47A, etc) CSNK2A enzymes (Niefind et al., 2007; Sarno et al., 1997). Positions 43, 44, and 47 are all located on the αC helix.

2.3.4 Position +4

Leucine (alignment and structure-based)

I detected six kinases (MARK2, CAMK1, PRKAA1 (human), PRKAA1 (mouse), PRKAA2 (human), and Snf1) with a moderate preference for leucine at position + 4. All belong to the CAMK *Group*. Snf1 (*S. cerevisiae*), PRKAA1, and PRKAA2, are homologous enzymes involved in the metabolic regulation of AMP levels (Hardie et al., 1998). For Snf1, the phosphosite-based classification is supported by the results of a peptide library assay (Mok et al., 2010), and from the analysis of synthetic peptide variants (Dale et al., 1995). This is also the case for PRKAA enzymes (Dale et al., 1995; Weekes et al., 1993), and CAMK1 (Dale et al., 1995).

The alignment-based analysis performed implicates domain position 164 as the sole putative SDR. This position has been discussed extensively above as a determinant for specificity at the +1 position. This residue is an alanine in five of the kinases listed above (valine in

CAMK1), and otherwise will be an aliphatic hydrophobic residue (L/I/M) or arginine in the case of most CMGC kinases and NEK kinases.

The PDB file 3IEC listed in Table 2.1 contains the MARK2 enzyme in complex with a peptide featuring a leucine residue at position +4. The peptide bound to the substrate-binding site is not a substrate but a peptide mimetic inhibitor of the *Helicobacter pylori* protein cagA (Nesić et al., 2010). In the cocrystal structure, the peptide forms a turn about position +2 so that the +4 hydrophobic side chain projects towards the P+1 pocket and stacks against the +1 residue (Figure 2.10). The substitution of aliphatic residues for alanine at position 164 in these kinases therefore seems to generate a small binding pocket that allows the substrate +4 position to functionally substitute for position 164 by stacking against the +1 residue.

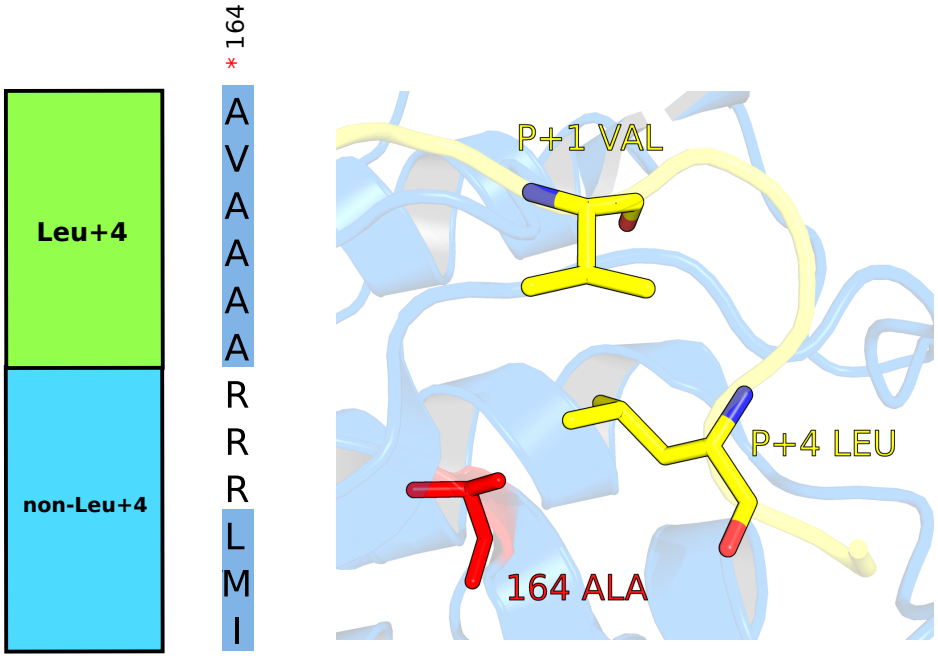


Fig. 2.10 The presence of a small side chain at kinase domain position 164 may promote the projection of a hydrophobic side chain at position +4 towards the P+1 pocket. *PDB: 3IEC*

2.3.5 Position -2

Proline (alignment and structure-based)

Of the 25 kinases I detected with a moderate proline preference, 22 belong to the CMGC *Group*. From the alignment-based analysis, I detected five positions (82, 161, 162, 188, 196) that are linked to this specificity.

Kinases with the preference usually contain a bulky hydrophobic residue (Y or W) at position 162, with a lack of conservation at this site for the non- proline-directed kinases (Figure 2.11). This residue has previously been named the 'P-3i-aromatic' as it was expected to bind most closely to the -3 residue (Kannan and Neuwald, 2004). However, for the two CMGC kinases listed in Table 2.1 with a residue at position -3 (PDB: 1QMZ and PDB: 2WO6), a larger contact area between 162 and position -2 was found than with 162 and position -3 (44.7 $Å^2$ and 26.9 $Å^2$, respectively).

The domain position 161 is part of the P+1 pocket and has been discussed previously in relation to proline +1 specificity, where it has been shown to bind to the 'secondary' activation loop phosphorylation moiety in some CMGC kinases (Bao et al., 2011)). However, it can also bind at the -2 position (Figure 2.11), as previously suggested (Kannan and Neuwald, 2004). In both structures referenced directly above, the residue at 161 forms non-bonded contacts with the -2 proline (17.90 $Å^2$ and 38.90 $Å^2$, respectively). For both domain position 161 and 162, I suggest that hydrophobic contacts with the pyrrolidine ring of the -2 proline confers this specificity. Position 161 was previously predicted to interact with the -2 substrate position (Kannan and Neuwald, 2004). Position 189 was also predicted to be a P-2 determinant but I find no basis for this from the analysis here (Mok et al., 2010).

The domain positions 188 and 196 were discussed previously (with respect to proline+1 specificity) and are again likely to serve as distal determinants here. Domain position 82 is marked by the absence of glycine relative to non-proline directed kinases. Its role as a proline-2 determinant is not immediately clear.

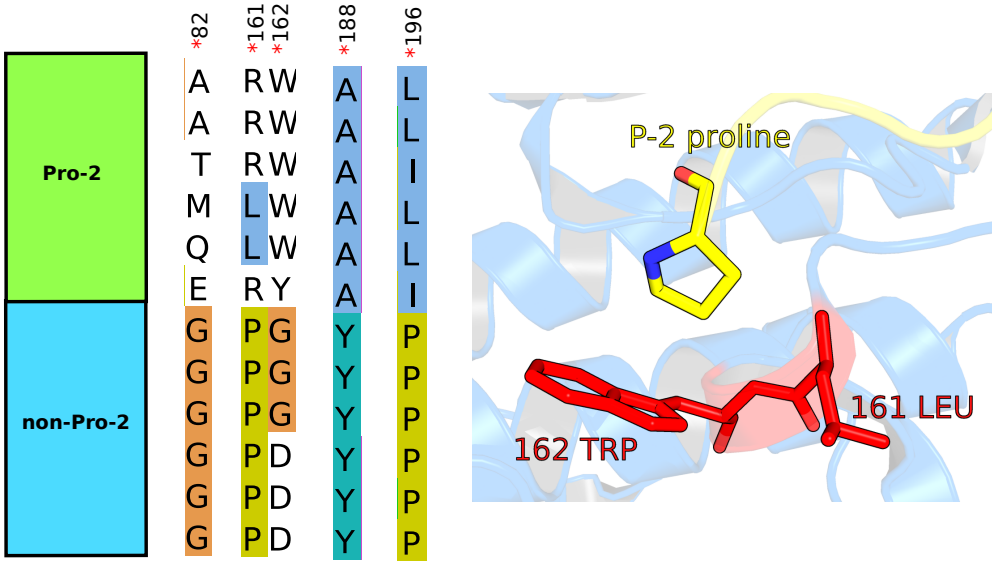


Fig. 2.11 Two of the five detected SDRs (161 and 162) are proximal to the -2 position in the structure shown. *PDB: 3QHR*

Aspartate/glutamate (alignment and structure-based)

I identified 13 kinases (7 Other, 1 AGC, 1 CAMK, 2 CK1, 2 CMGC) with a moderate preference for acidic amino acid residues. Domain positions 157 and 189 are implicated as specificity determinants here. Domain position 157 is located in the activation loop and so is unlikely to be a direct determinant for position -2. Position 189 however has previously been identified as a position -2 determinant in non-CMGC kinases from a number of structural analyses (discussed below). In support of this, a structural model generated here of an ADRBK1-peptide complex implicates this residue as a direct determinant of the -2 D/E preference (Figure 2.12). As expected, a CK2 kinase-peptide model generated here does not implicate position 189 for this role, as the CK2 *Family* belongs to the CMGC *Group*. This model suggests that a histidine at position 127 may instead influence position -2 specificity. Position 189 in these kinases is marked by the absence of a negatively charged D/E residue.

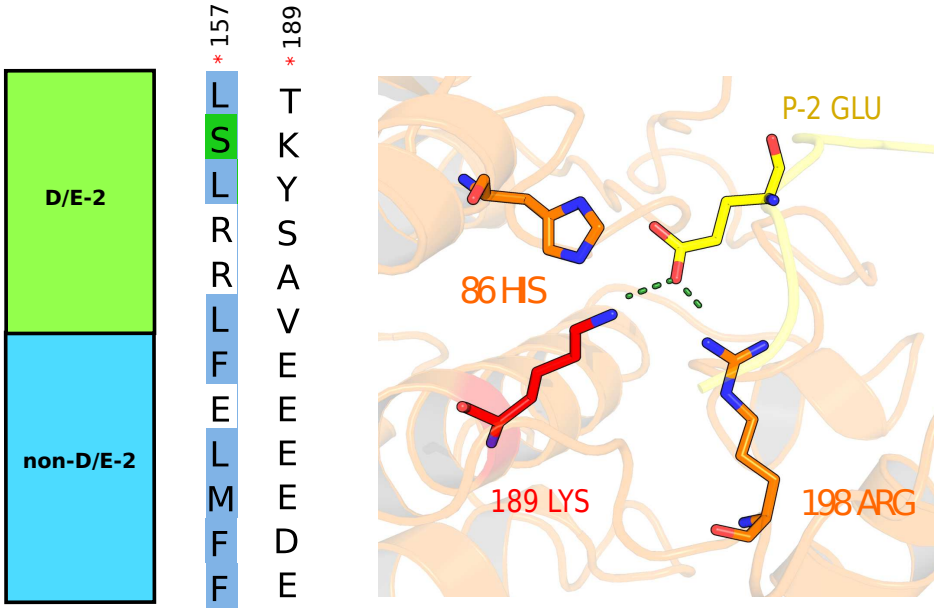


Fig. 2.12 A model constructed here for the ADRBK1-substrate complex places kinase domain position 189 in close proximity to the -2 substrate position

Arginine (alignment and structure-based)

The binding pocket for arginine at the -5 position has been observed to overlap structurally with that of arginine at the -2 position (Ben-Shimon and Niv, 2011; Zhu et al., 2005b), as was discussed in the *Introduction* section. I observe 6 kinases (PKN1, CHUK, PRKCB, PRKCD, PRKCE and PRKCG) with the acidic 127/189 pair that lack a significant preference for either arginine or lysine at positions -2 or -5 (CMGC kinases excluded, see below). This observation is in accordance with the suggestion of Ben-Shimon and Niv in 2011 that the acidic pair is neither necessary nor sufficient for the attainment of this specificity in some structural environments (Ben-Shimon and Niv, 2011).

In terms of R/K-2 binding, the mode of binding to arginine at -2 is similar between protein kinase A and PAK4/PAK1, in which both residues 127 and 189 co-ordinate the substrate arginine (Ben-Shimon and Niv, 2011). Protein kinase A and PAK1 belong to evolutionarily distant kinase *Groups* (AGC and STE, respectively), suggesting either that this binding mode evolved convergently or was an early event of protein kinase evolution (Zhu et al., 2005b). With respect to R/K-5 specificity, arginine binding in AKT structures is similar to that of R/K-2 arginine binding in protein kinase A, with the 127/189 acidic pair co-ordinating the substrate arginine along with a tyrosine residue at position 126 (Yang et al., 2002). In PIM kinases, the -5 arginine is also co-ordinated by two acidic residues but at positions 126 and 189, with the 127 glutamate not involved in binding (Ben-Shimon and Niv, 2011).

The alignment-based tools (see *Methods* chapter Section 6.1.3) employed here to identify specificity determinants assume independence between alignment positions and therefore would be unlikely to detect residues that determine specificity co-operatively such as the 127/189 acidic pair discussed above. Here, the alignment-based analysis implicates domain position 162 as a determinant for both the -2 and -5 arginine preferences (Figure 2.13). This position is represented mainly by an aspartate/glutamate for basophilic kinases and Y/W/N/G otherwise. This residue does not form polar contacts with the -2 arginine, but is proximal to the arginine-binding pocket (within 3.5 Å of -2 arginine in *PDB: 1ATP*). This residue has been implicated previously as an R-6 determinant but genetic analysis suggests a role for this SDR in the binding of other residues (Moore et al., 2003). Domain position 27 of the N-terminal β-pleated sheet is also implicated (tyrosine bias in arginine-directed kinases, valine otherwise), although is unlikely to be a direct determinant given its distance from the active site.

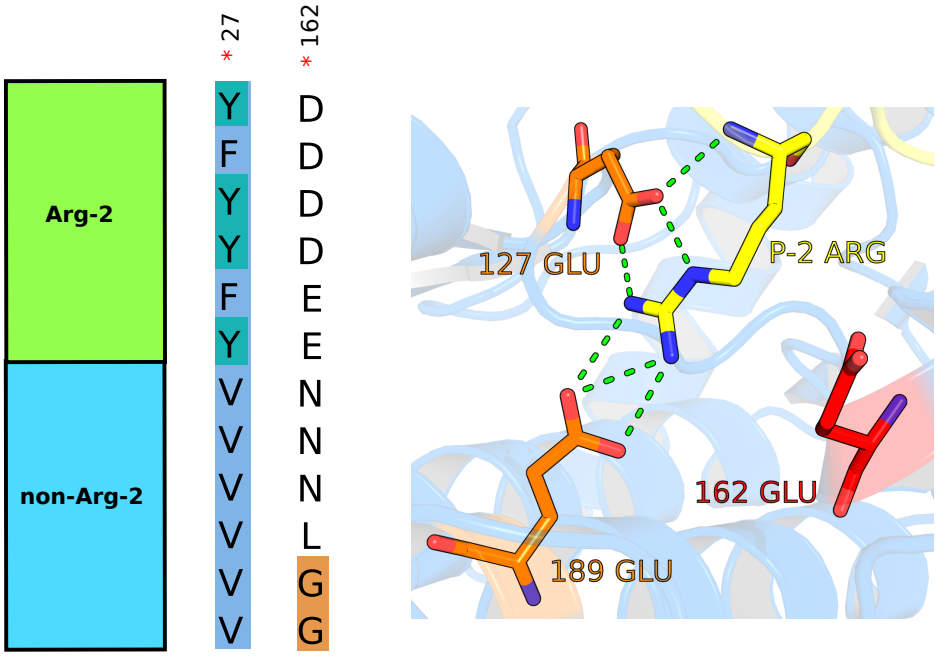


Fig. 2.13 Position 162 is predicted here as an R-2 SDR, and the 127/189 acidic pair has been described extensively in the literature

2.3.6 Position -3

Arginine (alignment and structure-based)

The basis for R-3 specificity was described in the *Introduction* section. While a universal mechanism for R-3 selectivity has sometimes been assumed on the basis of past studies (Ben-Shimon and Niv, 2011), the extent to which position 84 determines specificity in general is not fully understood. As previously discussed, sequence alignment-based analyses are required for a broader analysis of kinase specificity in the absence of kinome-wide experimental data. The multiple sequence alignment generated here of human, mouse, and *S. cerevisiae* protein kinases supports the earlier observation that E84 is not an obligate determinant (Mok et al., 2010), as 6 arginine-directed kinases do not feature E84. While I note that all arginine-directed kinases within the AGC and CAMK *Groups* feature E84, I identify 5 AGC/CAMK kinases with E84 but without a significant arginine preference at position -3.

The absence of the -3 R/K preference in these 5 kinases can likely be accounted for by domain positions 86 and 127, which line the -3 arginine binding-pocket of AGC and CAMK kinases. Both positions were identified as putative determinants from the alignment-based analysis performed here (Figure 2.14). For kinases with this preference, position 86 is mainly a tyrosine/phenylalanine and a polar amino acid otherwise, and position 127 is mainly a glutamate and serine/glutamine otherwise. In protein kinase A cocrystal structures,

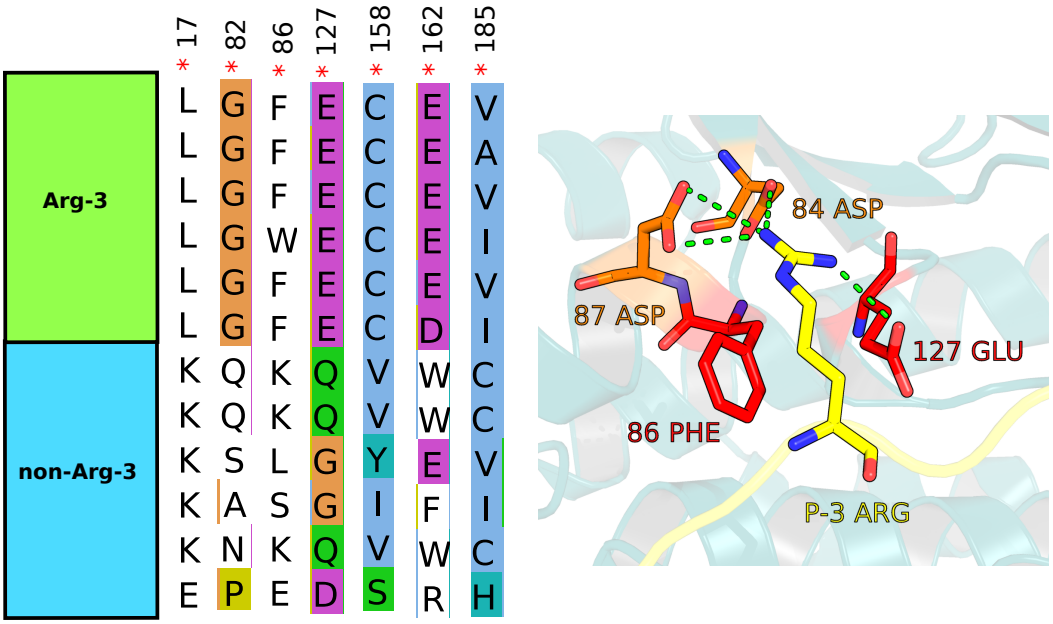


Fig. 2.14 Position 86 and 127 are predicted here as R-3 SDRs, and the aspartate/glutamate at position 84 has been described extensively in the literature

positions 86 and 127 pack against the hydrophobic moiety of the arginine side chain (Figure 2.14). I note that PDK1 features a leucine at position 86, GRK2 features a histidine at 86 and an alanine at 127, and that GRK5 features a lysine at position 86. These deviations from the canonical sequence patterns may account for the absence of the -3 arginine preference in these kinases. Kinase domain position 86 was predicted previously as an SDR for the Mok *et al* study (Mok et al., 2010). Domain position 127 was also suggested as an R-3 determinant but for CMGC kinases only, whereas it is suggested here that it can serve as an R-3 determinant for non-CMGC kinases also (Mok et al., 2010).

Glutamate at 127 has also been strongly implicated in the -2 arginine preference, as described above. I note however that, for AGC or CAMK kinases in which 127 is not involved in binding to either the -2 arginine or -5 arginine, that the glutamate can co-ordinate the arginine side chain at -3. I observe this to be the case for PKCι (*PDB: 4DC2*) and PIM1 structures (e.g. *2C3I*). Interestingly, in the AKT2 (-3 arginine and -5 arginine) PDB files *3E8D* and *3E87*, co-ordinates exist for the 127 glutamate side chain in polar contact with the -3 arginine and alternatively with the -5 arginine, suggesting that E127 can alternate flexibly between -3 and -5 co-ordination during the course of substrate binding and unbinding. I do not however observe the -3-binding conformation in any of the other 17 AKT cocrystal structures, suggesting that -5 contact is the dominant binding mode for E127 in AKT. Significantly, of the 22 kinases in the dataset with -3 arginine specificity but without a significant -2 or -5 arginine preference, all still feature E127, which suggests strongly that the observed

association between E127 and -3 is a direct association and not just an indirect result of compound -2 or -5 arginine preferences in -3 arginine kinases. Notably, a recent mutagenesis analysis of PKCβ has revealed that position 127 can indeed serve as an R-3 determinant in a non-CMGC kinase, and so confirms the prediction made here (Barber et al., 2018a).

The alignment-based tools employed here (see *Methods* chapter Section 6.1.3) also implicate domain positions 17, 82, 158, 162, and 185 in addition to the two other putative SDRs (86 and 127) previously discussed. Domain position 17 is marked by a leucine in arginine-directed kinases and by mainly by charged (K or E) residues otherwise. This residue side chain can be observed to pack against a tyrosine residue at position 79 in protein kinase A crystal structures. Notably, the tyrosine at position 79 is three residues N-terminal to the putative SDR at position 82, which is usually represented by glycine in R-3 kinases. I note also from manual observation of multiple sequence alignments that most arginine-directed kinases are characterised by a glycine insertion directly C-terminal to domain position 82. The alignment-based analysis performed here therefore suggests that divergent residues in the kinase hinge region (positions 79 to 83) between the N- and C-terminal lobes may contribute to the -3 arginine preference (Figure 2.15).

Domain positions 158 and 162 are part of the kinase activation segment and have been discussed previously. Position 185 is located on the kinase αF helix and is represented mainly by valine in arginine-directed kinases and mainly by cysteine otherwise. Given its position in the kinase domain, this position is unlikely to represent a direct determinant of specificity.

The R-3 kinases analysed belong to a total of 6 different kinase *Groups* (AGC, CAMK, CMGC, STE, TKL, and 'Other'). The structural analysis of some representative kinases revealed some *Group*-specific features for R-3 binding. I find for example that -3 binding in CMGC kinases is determined by the 127/189 acidic pair (discussed above), and therefore that -3 arginine binding in CMGC kinases corresponds to -2 arginine binding in non-CMGC kinases. For CAMK kinases also, it appears that an aspartate at position 87 binds to the R-3 in addition to the SDRs discussed above (Figure 2.14). For STE kinases, the mode of R-3 determination is not clear as the -3 arginine side chain is exposed to the solvent in the relevant co-crystal structures.

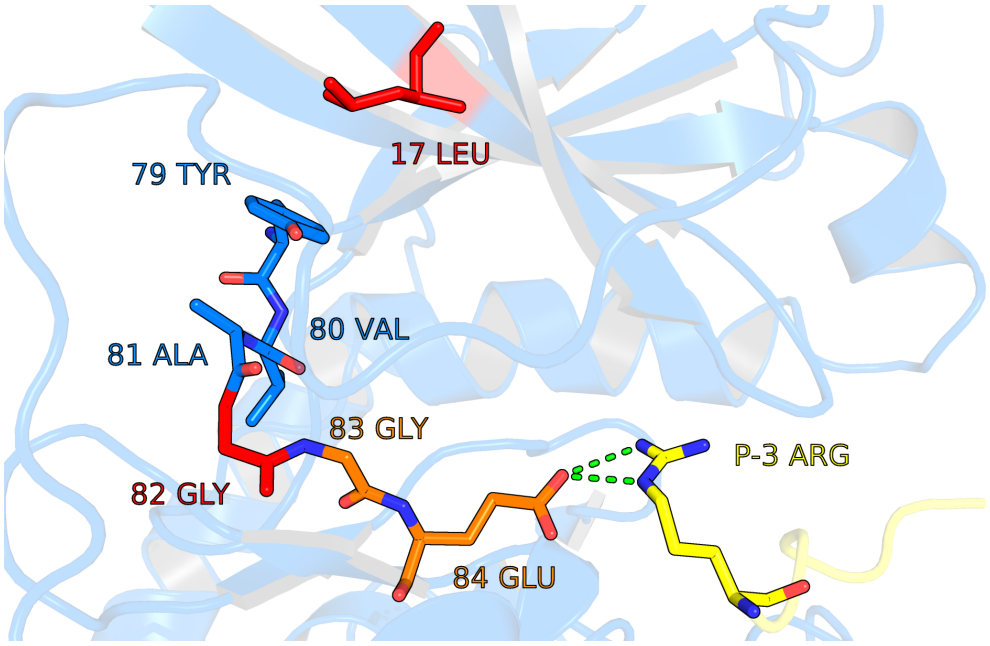


Fig. 2.15 The mechanism by which distal SDRs influence R-3 specificity is not currently clear. A network of contacts spanning the N-terminal lobe (position 17) and the active site (position 84) may affect R-3 selectivity. Positions 17 and 82 (coloured in red) were predicted as SDRs.

Aspartate/glutamate (alignment and structure-based)

I identified 13 kinases (8 Other, 2 CMGC, 2 AGC, 1 CKI) that are acidophilic at position -3.

For PLK1/PLK3/CDC5 (*S. cerevisiae* Polo-like kinase), a lysine at position 90 (αD helix) is a likely determinant as it forms a hydrogen bond with the -3 glutamate according to the model constructed here. Conversely, a lysine at position 86 (αD helix) is suggested as a determinant by the CSNK2A1 model constructed here as it forms a hydrogen bonds with the -3 glutamate (Figure 2.16). The model constructed here of the human CSNK1D protein suggest that arginine at 162 (activation segment) and a lysine at domain position 205 (αG helix) are the determinants of specificity. Position 162 was previously implicated as an pSer/pThr-3 determinant from an analysis of yeast kinases (Mok et al., 2010). In the constructed ADRBK1 model, the -3 glutamate side chain projects towards the solvent and does not form polar contacts with the protein kinase. The basis for the acidophilic character of ADRBK1 at position -3 is therefore not clear.

There is some support for the structural findings here in the form of the putative SDRs identified by alignment-based methods (see *Methods* chapter Section 6.1.3). Domains positions 86 (implicated in CK2 -3 D/E preference) was identified as a putative determinant, as was domain position 127 (implicated in CK2 -2/-3 D/E preference) by the alignment-

based tools employed here. Domain position 86 is marked in the alignment by an over-representation of basic residues relative to non-acidophilic kinases. Domain position 127 is marked in the alignment by an under-representation of glutamate residues relative to non-acidophilic kinases (Figure 2.16). In the latter case, it is likely that the loss of negative charge at this position contributes to a general acidic preference N-terminal to the phosphoacceptor. Domain position 157 was also identified as a putative SDR, although it is probable that this indirectly reflects the +1 acidophilic nature of some of the kinases listed rather than a direct contribution of 157 to -3 constraint, given the position of 157 in the kinase structure (kinase activation loop). Position 140 was also inferred as a determinant although the underlying mechanism is unclear as this residue is located closer to the ATP/inhibitor ligand than the substrate peptide in most kinase structures.

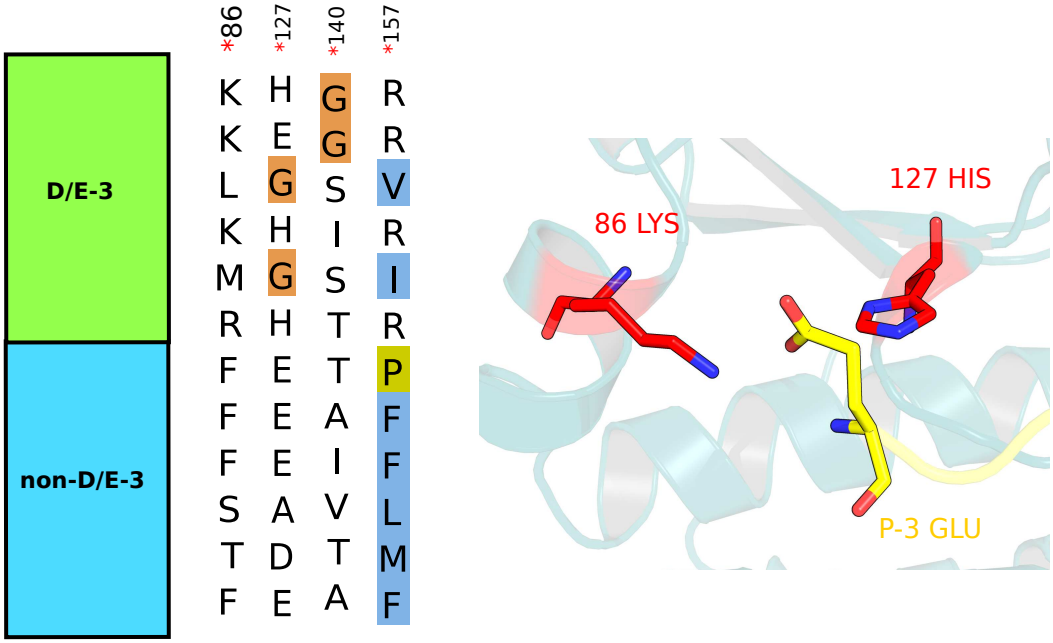


Fig. 2.16 Two of the predicted SDRs (86 and 127) for D/E-3 mapped onto a CSN2KA1-substrate model. Both kinase residues are in close proximity to the modelled position of the -3 substrate residue

2.3.7 Position -5

Leucine (alignment and structure-based)

I detected 21 kinases (14 CAMK, 5 AGC, 1 CMGC, 1PRK) with a moderate preference for leucine at position -5. The alignment-based analysis performed here suggests positions 86 and 189 as putative determinants (Figure 2.17). For the -5 leucine preference, most kinases with the preference feature a hydrophobic amino acid (mainly phenylalanine) at position 86. At position 189, most of the leucine-preferring kinases are marked by the absence of a glutamate at this position. However, I note that this trend relates to the CAMK kinases only – which account for 2/3 of leucine-preferring kinases – as all AGC and CMGC kinases with this preference still feature a glutamate at this position. Both residues are proximal to the -5 leucine in a MARK2-substrate structure (*PDB: 3IEC*), strongly suggesting that hydrophobic interactions between the kinase and substrate form the basis for this interaction (Figure 2.17).

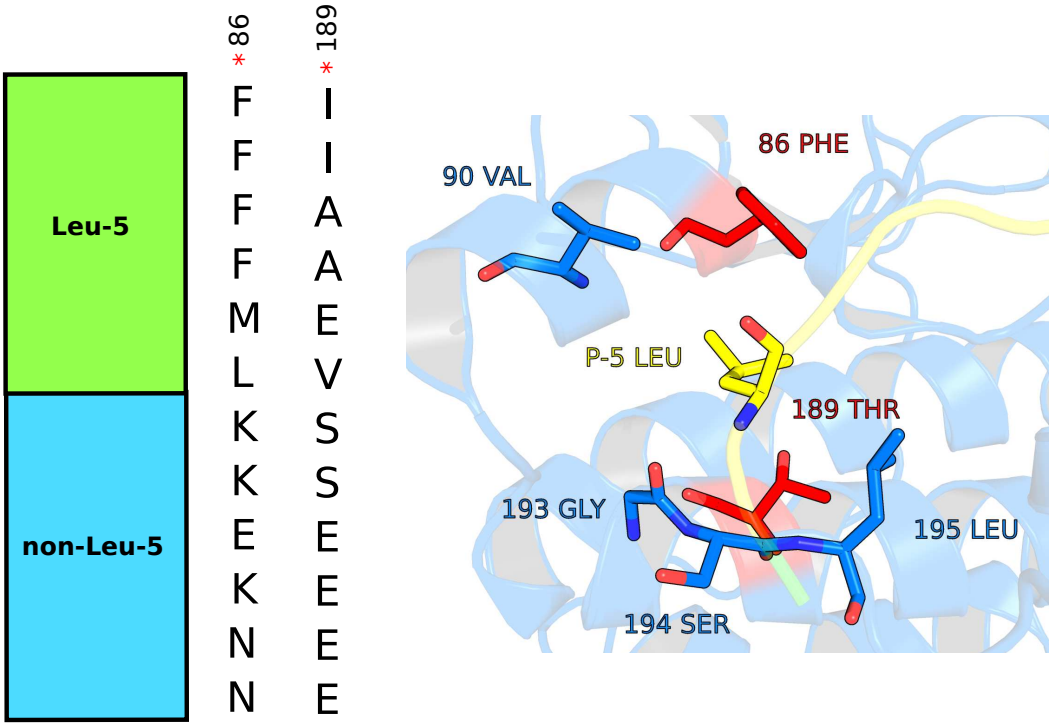


Fig. 2.17 The two predicted SDRs for the L-5 preference (positions 86 and 189) are proximal to the substrate and partly comprise a hydrophobic pocket surrounding the -5 substrate site

2.4 Experimental validation of SDRs

Kinase mutations and kinase activitity assays were performed by Cristina Viéitez at the EMBL research campus in Heidelberg. The mass spectrometry experiments were performed by Vinothini Rajeeve under the supervision of Pedro Cutillas at the Barts Cancer Institute, Queen Mary University of London. The author (David Bradley) selected the kinase to be mutated, the positions to mutate, and the three target peptides. Pedro Beltrao and Cristina Viéitez designed the assay. The data presented in Figure 2.18 was processed by Cristina Viéitez. Figure 2.18 was generated by the author (David Bradley).

Experiments were then undertaken to validate the MSA-based analysis described above. Here the objective was to mutate an SDR predicted from the analysis above and then to assay the peptide specificity of the wild-type and mutant kinases. For this purpose we analysed domain position 164 for the L+4 preference and 189 for the L-5 preference, as neither preference had been fully characterised at the time of the analysis. The *S. cerevisiae* kinase Snf1 was selected for mutation as both L-5 and L+4 preferences are found in this kinase (Dale et al., 1995). An overview of the experimental procedure is given in Figure 2.18. Briefly, the 164 and 189 positions were first subject to non-synonymous mutations (Snf1 A218L and Snf1 V244R, respectively). Each kinase was then incubated with 3 different peptides: WT, MutA (L+4 → A+4), and MutD (L-5 → D-5).

The WT peptide should represent an optimal substrate for the WT Snf1 kinases. The MutD peptide however would be expected to represent an optimal substrate for V244R due to complementary positive-negative charges between the kinase and substrate. For MutA, we expected this peptide to bind to A218L with equal or greater efficiency than the WT peptide, as the expected basis for L+4 specificity would be lost (see *Position+4: leucine*). From Figure 2.18, we do indeed observe these trends generally across the 3 time points (0 min, 7 min, and 20 min) for each of the three kinases assayed (WT, A218L, and V244L).

A recent study also concomitantly predicted position 189 as an L-5 determinant from a comparative structural analysis (Chen et al., 2017). However, while this residue was mutated and the specificity tested, mutation of 189 always occurred in combination with other kinase residues and so the role of position 189 *per se* as an L-5 SDR remained ambiguous. This study exists in addition to a previous one experimentally confirming the role of position 195 as an L-5 determinant (Scott et al., 2002). Domain position 126 has also been predicted as an L-5 determinant but without experimental confirmation (Mok et al., 2010). It is therefore likely that position 189 is one of multiple residues in the kinase hydrophobic pocket that confers L-5 specificity.

The L+4 specificity in comparison was completely uncharacterised before this analysis, and is significant because a traditional +1 determinant (164) has been linked to a distal substrate position (+4) for the first time.

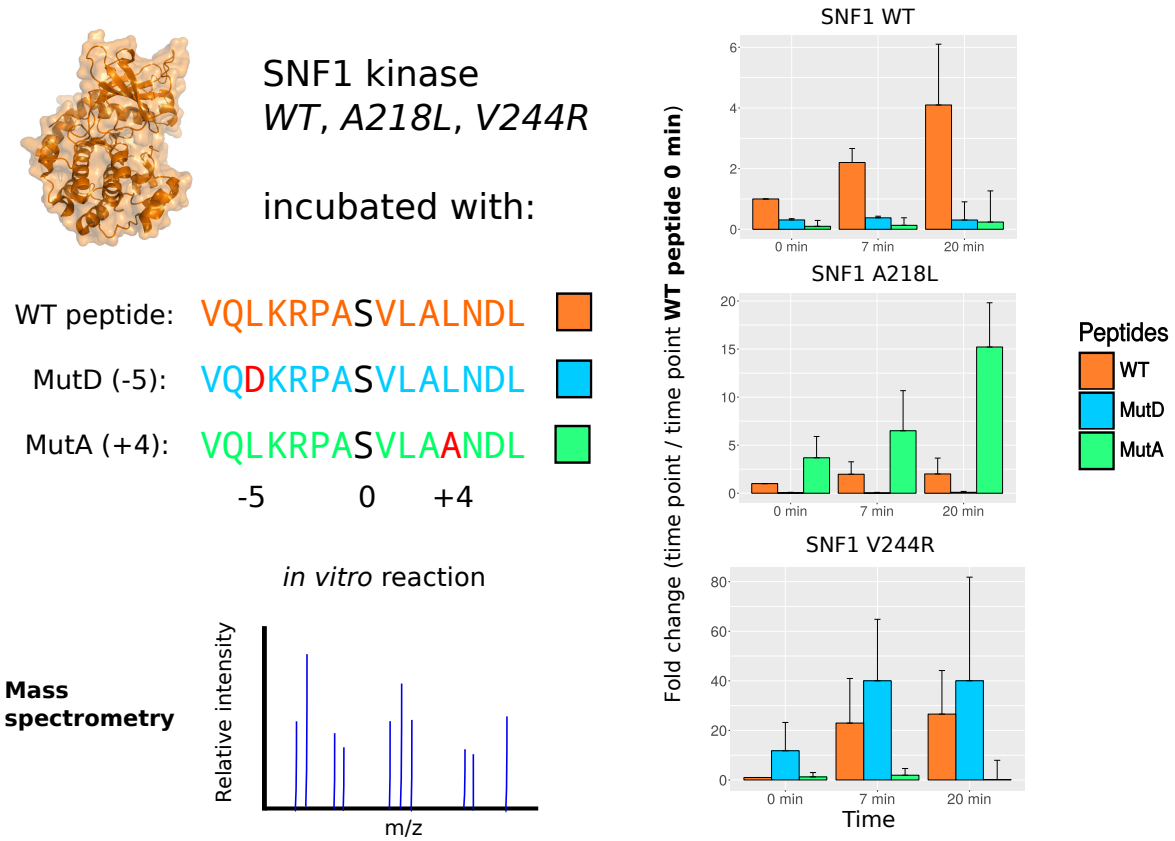


Fig. 2.18 *S. cerevisiae* kinase Snf1 was mutated at residues 218 (domain position 164) and 244 (domain position 189). The three SNF1 kinases (WT, A218L, and V244R) were separately incubated with the peptide substrates: one wild-type, one mutated at position+4 (MutA), and one mutated at position-5 (MutD). The extent of phosphorylation of each substrate peptide by each kinase was determined using mass spectrometry. The results of this analysis are explained in the main text.

2.5 Prediction of protein kinase specificity

Kinase sequence-based predictors of kinase specificity were discussed in the *Introduction* section. Here I was interested in using the results discussed above to inform a new predictor based on the kinase domain primary sequence. To this end, simple naive Bayes models were constructed to predict kinase specificity in a binary fashion. For the proline+1 specificity, for example, a given naive Bayes model would take a kinase primary sequence as an input and as

an output would give the posterior probability that the kinase of interest has this specificity. As a training set I use an MSA of all Ser/Thr kinase sequences with known specificity (i.e. the 119 Ser/Thr kinases used for the analysis in this chapter). Each position in the sequence alignment could potentially be used as a feature for model training. However, I limit feature selection for each model to the SDRs predicted for that model as described above and as listed in Figure 2.4a. I make only one exception for the R-2 model, as the 127 and 189 acidic residues are strongly implicated as determinants in the literature but mediate their effects inter-dependently and so were not predicted by the MSA-based procedure outlined in Figure 2.3

Models were generated for the five specificities with at least 20 positive examples in the data set used: R-3, R-2, P+1, P-2, L-5. A leave-one-out approach was then used for the cross-validation of models. This involved training a model on an MSA of all kinase sequences with one removed (119-1 = 118 kinases for training), and then testing the model on the excluded kinase. This procedure is repeated 119 times – once for each unique kinase. The performance of the models was assessed by calculating the area under the ROC curve and the area under the precision-recall curve, which are presented in Figure 2.19. The ROC curve AUCs were also used to to select the optimal subset of model features (i.e. kinase SDRs) to be used for training, which are listed in Table 2.4. The results presented in Figure 2.19 are based on these features.

Relatively high ROC AUC values were calculated for all five models (P+1: 0.99, P-2: 0.91, R-2: 0.82, R-3: 0.96, L-5: 0.82). The areas under the precision-recall curves (P+1: 0.98, P-2: 0.73, R-2: 0.47, R-3: 0.93, L-5: 0.55) were also higher than the baseline expectation for all five cases (P+1: 0.30, P-2: 0.21, R-2: 0.23, R-3: 0.46, L-5: 0.18). To put these results into perspective, I repeated this cross-validation procedure but for models trained using the *Predikin* method described in (Saunders et al., 2008) and the *Introduction* section. I then repeated cross-validation again also using a naive Bayes model but this time using the SDRs suggested by the *Predikin* method instead of those given in Table 2.4 (Brinkworth et al., 2003; Saunders et al., 2008). The same data was used for training and cross-validation for all 3 approaches (see *Methods* chapter Section 6.1.8). The results show that the naive Bayes model (with MSA-predicted SDRs) performs as well as the full *Predikin* method for all specificities except P-2, for which it performs better (Figure 2.19). The main advantage of the naive Bayes approach is that it can predict the specificity of any alignable kinase sequence, whereas *Predikin* requires that at least one of the reference kinases is sufficiently similar to the kinase of interest to make a prediction. Also, the *Predikin* approach is limited to substrate positions -3 to +3 and so is unable to represent preferences such as L-5.

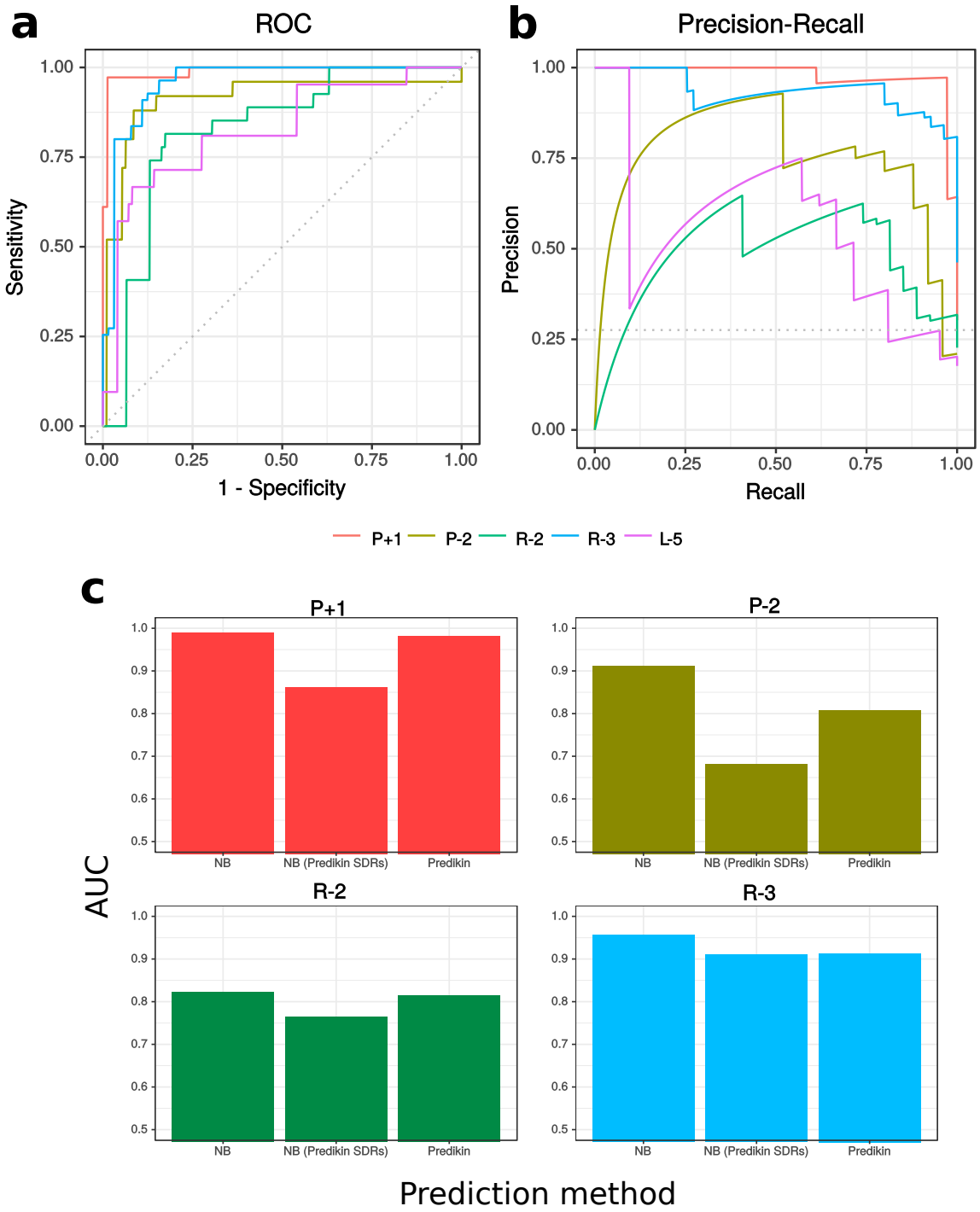


Fig. 2.19 a) ROC curves for naive Bayes models for five different kinase specificities: P+1, P-2, R-2, R-3, and L-5. b) Precision-recall curves for naive Bayes models for five different kinase specificities: P+1, P-2, R-2, R-3, and L-5. c) Area under the curve (AUC) values for the naive Bayes model (NB), the naive Bayes model trained upon the SDRs suggested by *Predikin*, and the full *Predikin* method implemented here. The data used for model training and cross-validation was the same in all three cases. The L-5 specificity is not represented because the *Predikin* method only considers substrate positions -3 to +3

Preference	Kinase positions
P+1	159, 188, 196
P-2	82, 162, 188
R-2	127, 162, 189
R-3	non-CMGC: 82, 86, 127, 162 CMGC: 86, 127, 189
L-5	86 , 189

Table 2.4 Kinase SDRs that were used to train the naive Bayes models discussed above. All positions were predicted as SDRs based on the approach presented in Figure 2.3 except from the residues highlighted in red, which were derived from previous studies

2.6 Mutation of kinase SDRs in cancer

Kinase SDRs are often targeted in cancer and congenital diseases (Berthon et al., 2015; Creixell et al., 2015b), and regulatory and substrate-binding regions are known to be frequently mutated in non-cancerous diseases also (Dixit et al., 2009; Torkamani et al., 2008). Here I have used mutation data from The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov/) to quantify the extent to which SDRs are targeted relative to kinase domain residues with other functional roles.

Overall, 20,032 unique mutations were mapped to the protein kinase domain. Some of these kinases had been identified previously as the products of oncogenes or tumour-suppressors and are recurrently mutated in these TCGA samples. The Ser/Thr oncokinase BRAF for example was mutated 749 times in these samples, and 116 unique mutations were found in the domain of the tumour suppressor CHEK2. In some cases, the cancer-associated kinases are mutated at residues that are likely to affect the peptide specificity of the kinase. For example, the cancer-associated kinase ERBB4 is found to be mutated twice at domain position 159 (in the +1 binding pocket), and its oncogenic paralogue ERBB2 was found to be mutated 17 times at position 164, which also forms part of the P+1 pocket. However, the mechanism (if any) linking these mutations to cancer progression awaits further experimental analysis.

For further analysis, I divided the kinase domain into one of four functional categories: ‘regulatory’, ‘catalytic’, ‘SDR’, and ‘Other’. ‘Regulatory’ positions are defined as those residues that are either located in the activation loop or are proximal in space to the primary activation loop phosphate. I include the kinase regulatory spine and the APE motif also as both regions have been linked to kinase regulation (Steichen et al., 2010; Taylor and Ko-

Category	Kinase positions
Catalytic	8, 10, 13, 15, 28, 30, 48, 85, 123, 125, 128 129, 130, 131, 140, 141, 186, 190
Regulatory	44, 52, 63, 121, 122, 142, 144, 145, 146, 147, 148, 149, 150, 151, 152, 155, 156, 157, 158, 165, 166, 167
SDR	6, 126, 127, 157, 158, 159, 161, 162, 164, 189

Table 2.5 Kinase domain positions that were used to define the functional kinase categories described in the main text

rnev, 2011). The 'catalytic' positions were defined primarily from the literature and were discussed in the *Introduction* chapter; residues of the catalytic spine were included also (Taylor and Kornev, 2011). I considered as 'SDRs' those residues that are both listed in Figure 2.4 and are proximal to the substrate (within 4 Å of the substrate peptide). This therefore represents a 'high-confidence' set of predicted SDRs that will minimise the number of false positive SDR predictions. 'Other' represents the complement of the kinase domain to these other three sets. These functional categories are represented structurally in Figure 2.20 and listed in Table 2.5.

Analysis of this mutation data reveals that the 'SDRs' are more frequently mutated than 'Other' residues with an unassigned function (Mann-Whitney $p = 0.0006$, one-tailed; Figure 2.20). This corresponds well to previous findings that kinase SDRs are often targeted in cancers. The finding that 'SDRs' are also more frequently mutated than 'catalytic' residues (Wilcoxon $p = 0.010$, one-tailed; Figure 2.20) contradicts a previous analysis suggesting a stronger enrichment of non-synonymous mutations in catalytic residues (Dixit et al., 2009). However, it is in agreement with a more recent analysis suggesting that kinase mutations altering kinase specificity are more common than mutations that would be predicted to inactivate the kinase (Creixell et al., 2015b).

Finally, I was interested to determine whether kinases SDRs are differentially targeted between kinases of different specificities. To this end, residue mutation frequencies were compared between predicted P+1 kinases and predicted R-3 kinases in human (posterior probability > 0.9 for both P+1 and R-3 predictors, respectively), as P+1 and R+3 represent the most common human specificities. For this comparison I identify 3 kinase SDRs (159, 161, and 164) that are differentially mutated between the two groups (Figure 2.20). I find this to be the case also when analysing just kinases of known specificity and therefore exclude the possibility that this result could be an artefact of inaccurate naive Bayes specificity predictions (*Appendix* Figure A.1). Domain position 164 and 161 are located in the P+1 loop and were mutated more often in proline+1 kinases. For position 161, the MAP kinases in particular are recurrently mutated in independent samples (MAPK1: 3, MAPK8: 3, MAPK11: 2, MAPK1: 1). This position is known to bind to the phosphotyrosine at 157 that exists in

MAPKs (Varjosalo et al., 2013). However, a previous study suggests that this this secondary activation loop phosphotyrosine is more important for kinase activation than kinase specificity (Prowse et al., 2001). Conversely, position 159 and 164 are both expected to be critical for kinase specificity and are highly conserved within their respective subgroups (high 164 R conservation in P+1 kinases and high 159 G conservation in R-3 kinases). I therefore find some limited evidence that the frequency of SDR mutations can differ depending upon the kinase specificity.

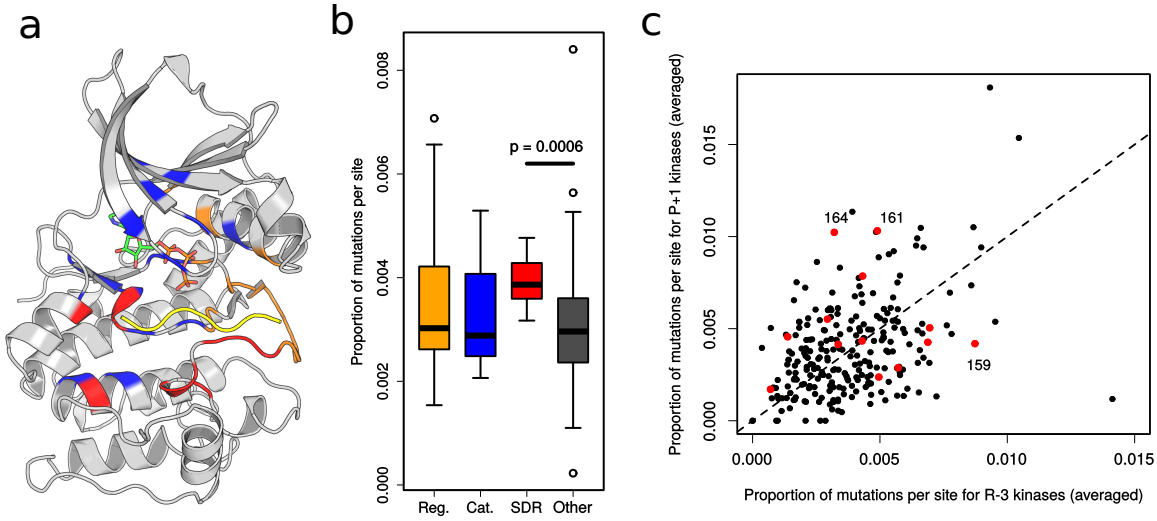


Fig. 2.20 a) Division of the kinase domain into functional categories 'regulatory' (orange), 'catalytic' (blue), 'SDRs' (red), and 'Other' (grey). b) The frequency of kinase domain mutation by functional category. c) The frequency of residue mutations for predicted R-3 kinases (x-axis) and P+1 kinases (y-axis). SDRs listed in Figure 2.4 for R-3 and P+1 are coloured in red.

2.7 Discussion

The analysis of kinase sequence and kinase structure has been combined to provide a comprehensive study of protein kinase specificity. Specificity models were first constructed for over 100 kinases belonging to either human, mouse, or *S. cerevisiae*. An unsupervised learning approach was then adopted to identify recurrent specificities for substrate positions -5 to +5. For each recurrent preference, MSA positions with the strongest association to that preference were predicted as SDRs. Where possible, available kinase-substrate structures in the PDB have been used to rationalise the results obtained. However, several kinase-substrate were also constructed where the specificity of interest was not represented by the list of cocrystal structures given in Tables 2.1-2.3. Two of the SDRs predicted were validated

experimentally using a mass-spectrometry based approach. I also use the results here to interpret data from the The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov/), and demonstrate that SDRs are more frequently mutated even than catalytic residues, and that SDRs can be differentially mutated between kinases of different specificities. Finally, I aggregated contact data from all known kinase-substrate cocrystal structures to give an overview of physical kinase-substrate interactions.

The physical analysis of kinase-substrate contacts is largely concordant with a previous analysis suggesting well-defined pockets in the kinase ('subsites') required for binding to the substrate positions -3 to +3 (Brinkworth et al., 2003). Remarkably, most of the kinase-substrate contacts described for that analysis of only three structure – CDK, PKA, and PHKg – are recapitulated among many of the newer structures analysed here and therefore attest to the structural conservation of the kinase domain in its active conformation. The analysis of newer structures however demonstrates a broader range of contacts than first described, and that most kinase-substrate contacts are not universal even if common. These results imply that the structure-guided prediction of kinase specificity (as implemented in *Predikin*, for example) should still provide a reasonable estimate for many canonical protein kinases. However, for some kinases these structure-based predictions will not be valid. The haspin kinase-substrate structure and the NAK-peptide model for example deviate strongly from the structures used to guide the *Predikin* approach (Eswaran et al., 2008; Maiolica et al., 2014), where the substrate binds in an extended linear conformation. Both kinases belong to the non-canonical 'Other' *Group* whereas most of the kinases listed in Tables 2.1-2.3 belong to the AGC/CAMK/STE *Groups*. The extent to which the binding profile presented in Figure 2.2 will apply to other *Groups* is not clear at present, but for the 'Other' kinases in particular a broad range of binding conformations is likely given the current evidence. Structural approaches also do not account for the looping of the substrate discussed in the *Introduction* chapter, but from the *Results* above it is clear that this can have a strong effect on kinase specificity. L+4 looping for example (Figure 2.10) allows a traditional +1 determinant (position 164) to influence +4 specificity. The prevalence of substrate looping at the active site is not currently known.

The 119 specificity models constructed suggest strong constraint for the -3, -2, and +1 positions. This is consistent with the structural evidence as dedicated pockets for binding exist for these positions, whereas binding to -4 and -1 is generally weak, and the +2 and +3 residues bind usually in a groove between the αC helix and the activation loop. However, as with the structural data, the specificity data is incomplete (only ~20% of human kinases covered) and biased towards the canonical AGC/CAMK/STE *Groups*. Whether the results will be robust to the influx of new data from human and/or other species remains to be

seen. It should be noted however that the results presented here are largely consistent with a previous study surveying the specificity of 61 *S. cerevisiae* kinases (~50% of kinome) (Mok et al., 2010).

The analysis of specificity here covers previously characterised preferences (P+1, R-3, R-2) but also under-studied ones (L-5, P-2, L+4, etc). Generally, the results generated for the well-characterised species are in agreement with past studies, but some new features were discovered also. I suggest for example that position 127 is a determinant for R-3 also and not just for R-2, a prediction that has recently been validated (Barber et al., 2018a). For P+1, the structural analysis also implicates domain position 202 as a determinant in the CDKs specifically. For R-3, it was noted also that the mode of kinase binding differs between the AGC, CAMK, CMGC and STE *Groups*. This should be contrasted with the R-2 preference, where the mode of binding is identical between the distantly related AGC and STE *Groups* (Zhu et al., 2005b). Finally, a number of SDRs distal from the kinase active site were also predicted. I propose a mechanistic role for a few cases, but these have yet to be validated experimentally. It will be interesting to probe in the future the mechanism (if any) linking these distal positions to the kinase active site.

Insights were also revealed for many specificities that were poorly characterised previously. In most cases, the predicted SDRs were not entirely new but represent 'new' roles for SDRs previously assigned to a different specificity. Position 189 for example was previously linked to R-2/R-5 specificity (Zhu et al., 2005b), but its role as an L-5 determinant was predicted and experimentally validated also. As described above, position 164 is traditionally considered as a +1 determinant (Kannan and Neuwald, 2004; Zhu et al., 2005a), but we predict and experimentally validate its role as a +4 determinant also. Other specificities are not represented by the structures in Tables 2.1-2.3 and so models have been manually constructed here for the kinase-substrate interface. In many cases the same specificity is found in distantly related *Families*. For some cases – such as D/E+1 in in the CK2 (CMGC) and GRK (AGC) – the models suggest that the same specificity has evolved by different mechanisms in the two *Families*. In this case, the SDRs for CKs had been experimentally verified before (Sarno et al., 1997), and so this conclusion is supported by extension. However, in other cases the opposite is observed; that the same specificities in unrelated *Families* had converged upon the same mechanism to determine that preference. This is suggested to be the case here for D/E+2 in the PLK and CAMK2 *Families* upon the basis of the modelled interface for PLK3 and its substrate and a kinase-substrate crystal structure of CAMK2 in the PDB. As more structural data becomes available, it will be interesting to determine whether or not the model-based predictions were correct. A summary of all SDRs discussed in this

chapter, based either on structural models or kinase sequence evidence, is given in *Figure 2.21*

The results listed in Figure 2.4 were used to construct sequence-based naive Bayes predictors for P+1, R-3, R-2, P-2, and L-5. Benchmarking of these models reveals that they perform similarly to the *Predikin* method for specificity prediction (Saunders et al., 2008), except for the P-2 specificity, which is modelled more successfully using the naive Bayes approach. There are three advantage overall to this approach relative to *Predikin* 1) predictions can potentially made for any specificity and not just those within the -3 to +3 window. 2) Predictions can be made for any query kinase sequence, although the accuracy of prediction will vary (depending on whether any kinases in the training set were similar in sequence to the query kinase) and 3) the predictions are probabilistic, and so the level of confidence in the prediction can be quantified. All three of these are significant advantages for research into the evolution of protein kinase specificity, and this is explored further in *Chapter 3*.

Finally, as an application of these findings, I used the results here to interpret cancer genome data from The Cancer Genome Atlas (TCGA: http://cancergenome.nih.gov/). I first discovered that kinase SDRs are mutated more frequently than random residues in the kinase domain, and even more often than catalytic residues. This accords with a previous analysis and suggests that signalling pathways are more likely to be rewired than inactivated during tumourigenesis (Creixell et al., 2015b). It was also found that SDR mutations can occur differentially between kinases of different specificities, which is in line with expectation considering that the impact of SDR mutation will differ depending upon the specificity of the kinase. Grouping all SDRs regardless of kinase specificity would therefore over-estimate the effect of mutation for some kinases. Overall, the results generated here could be expected to help further in classification of cancer mutations into 'drivers' and 'passengers'. More generally, multiple other pathologies can be triggered also by kinase domain mutations (Lahiry et al., 2010), and I expect the results presented here to aid in the interpretation of such variants.

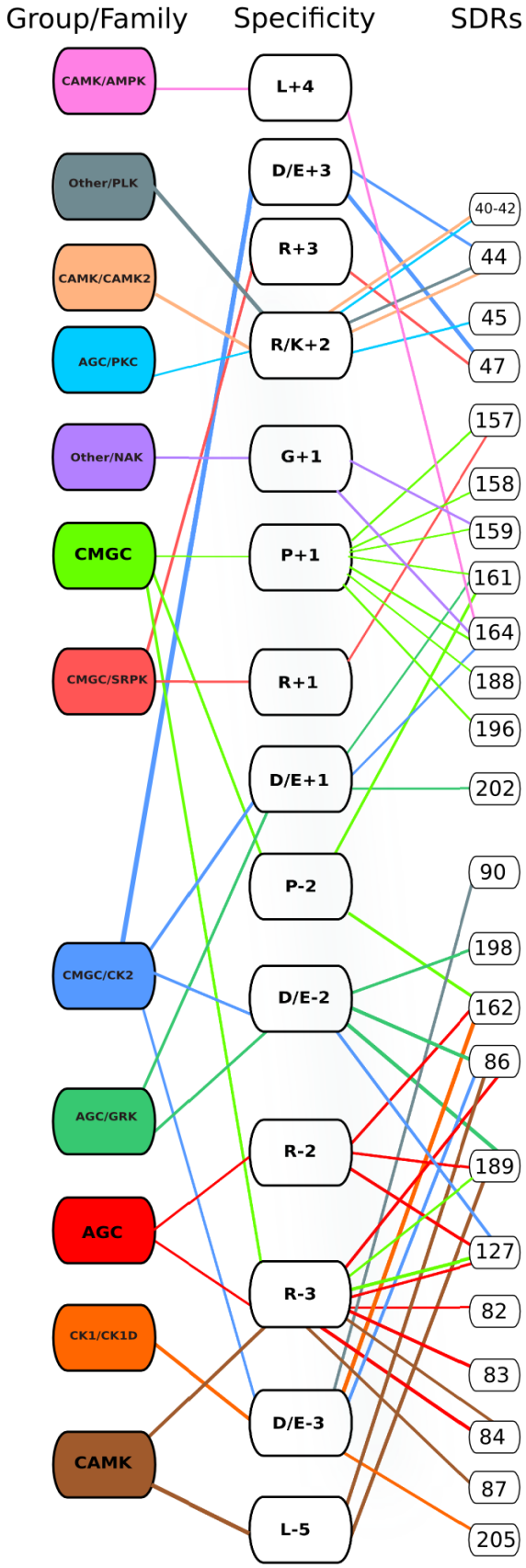


Fig. 2.21 Summary of all SDRs discussed in this chapter, either on the basis of structural evidence or sequence evidence (or both). The left-hand column refers to the kinase *Group* and *Family* where applicable. The middle column refers to the kinase specificity. The third column refers to the SDRs associated with a particular kinase specificity and *Group*/*Family*.

Chapter 3

The evolution of kinase function

In this chapter, I investigated the evolution of kinase specificity following speciation and following gene duplication. All of the analysis was performed by the author (David Bradley) under the supervision of Pedro Beltrao. Some of the work presented in this chapter was included in a preprint manuscript:

David Bradley, Cristina Viéitez, Vinothini Rajeeve, Pedro Cutillas, and Pedro Beltrao (2018): Global analysis of specificity determinants in eukaryotic protein kinases. *bioRxiv*

3.1 Introduction

Protein kinase peptide specificity was discussed extensively in *Chapter 2* from a structural perspective. The objective of *Chapter 3* instead is to investigate the evolution of protein kinase specificity and the evolution of kinase SDRs. This is a key question in evolutionary cell biology as kinases often serve as 'hubs' in complex signalling pathways (Albert, 2005; Zhu et al., 2007). Changes in kinase specificity are therefore likely to represent dramatic changes in their corresponding kinase-substrate networks. However, this subject has received surprisingly little attention in the literature compared to studies regarding the evolution of transcription factor specificity (Babu et al., 2004, 2006; Howard et al., 2014; Teichmann and Babu, 2004). Moreover, while some studies do exist for the evolution of phosphorylation sites (Beltrao et al., 2009; Freschi et al., 2011, 2014; Landry et al., 2014, 2009), the phosphorylating kinases and their evolution are not usually considered in these analyses. For this chapter, I focus on protein kinases as potential alternative drivers of phenotypic diversity and divergence. The evolution of kinase orthologues and paralogues are first studied separately and then both are considered together for a detailed case study of kinase evolution within one *Family*.

The extent to which kinase specificity is conserved between orthologues is not understood. However, previous experimental studies suggest indirectly that kinase specificities are conserved between distantly related species. Most notably, a genetic complementation study in *Schizosaccharomyces pombe* revealed that the human CDK1 could functionally substitute for the fission yeast cdc2 cyclin-dependent kinases (Lee and Nurse, 1987). This implies that the specificities of these kinases are also conserved between *S. pombe* and human. Since then, a number of *Saccharomyces cerevisiae* kinases have been similarly complemented genetically, including CDC28, HOG1, CDC15, MPS1, CAK1, and HRR25 (Elledge and Spottswood, 1991; Hamza et al., 2015; Yang et al., 2017). In 2010, more direct evidence for the specificity of conservation was provided when 61 *S. cerevisiae* kinases were assayed using peptide libraries (Mok et al., 2010). This experiment revealed a number of examples where kinase motifs are conserved between budding yeast kinases and their human orthologues (Miller and Turk, 2018). For example, the S/T-P-x-K motif is found in both CDC28 *S. cerevisiae* and CDK2 (human), and the R-R-x-S/T motif in both Tpk1 *S. cerevisiae* and PKA (human). However, the extent to which kinase specificities are conserved between orthologues has not been investigated systematically. In this chapter, the extent of specificity conservation is quantified for the first time using experimental data from both *S. cerevisiae* and human. The predictive models presented in *Chapter 2* are leveraged here also for a pan-taxonomic conservation analysis. These models take a primary kinase sequence as an input and therefore can be used to predict kinase specificities from a wide range of species.

Orthologues are generally assumed to be more conserved in function than paralogues – the so-called 'orthologue conjecture' – and this assumption extends to protein kinases also (Koonin, 2005; Studer and Robinson-Rechavi, 2009; Thomas et al., 2012). Paralogue divergence in protein kinases can be tested by comparing kinases belonging to different *Families* or *Subfamilies*, which are most often generated following gene duplication events. Therefore, by identifying divergent residues between sister families in a kinase phylogeny, it would be possible to discover the residues responsible for the functional specialisation of kinases following duplication. This is attempted in this chapter for 99 kinase *Families* and 88 kinase *Subfamilies*. I adopted a phylogenetic approach for functional residue prediction to explicitly account for residue changes occurring only following gene duplication. The results are then aggregated between *Families* to gain a global perspective of kinase evolution at the *Family* level. Sequence changes that map to kinase SDRs imply the divergence of kinase specificity. These analyses are performed at the *Subfamily* level also for this study to gain insight into more recent evolutionary divergence.

Finally, ancestral sequence reconstructions are used for the detailed evolutionary analysis of a single kinase *Family*. This analysis follows from the previous ancestral sequence

construction of the CMGC *Group* in 2014 (Howard et al., 2014), which was concerned with the R+1 and P+1 specificities in particular. The central conclusions from that analysis were 1) that substitution of a single residue can induce divergence in specificity 2) divergence can occur without a corresponding drop in kinase activity and 3) evolution of a new specificity occurred through an intermediate of broad specificity and therefore occurred via a process of subfunctionalisation. The purpose of the research conducted in this chapter was to test the generality of these conclusions with respect to kinases of a different *Family* and specificity. It was also hoped that a detailed case study would strengthen the understanding of kinase evolution at the active site. Towards this end, an evolutionary reconstruction of the GRK (G protein-coupled receptor kinase) *Family* was performed, which is an acidophilic kinase that likely emerged from a basophilic ancestor (Lodowski et al., 2006).

3.2 Conservation of specificity between orthologues

3.2.1 Conservation of predicted specificities

As discussed in the *Introduction* section of this chapter, the extent to which kinase peptide specificities are conserved between orthologues is not currently understood. To explore this further, the sequence-based naive Bayes models described in *Chapter 2* have been used to predict the specificity of kinase orthologues. All five models – P+1, P-2, R-2, R-3, and L-5 – performed well when assessed by cross-validation (Figure 2.19) and are therefore appropriate for the evolutionary analysis performed in this chapter.

I use the *Ensembl Genomes* pan-taxonomic *Compara* resource for the prediction of kinase orthologues across a taxonomically broad dataset, including sequences from fungi, plants, metazoa, and protists (Herrero et al., 2016; Kersey et al., 2018; Yates et al., 2015). In Figure 3.1, conservation values are plotted for the SDRs across orthologues for each of the five specificities listed above. The conservation values were averaged across three functional categories: 'kinase domain', 'SDR', and 'catalytic'. In all five cases, the SDRs predicted in *Chapter 2* are highly conserved and are also more highly conserved than random (i.e. non-SDR) residues in the protein kinase domain. These results suggest that sequence changes likely to affect kinase peptide specificity have been subject to purifying selection. However, the average conservation of kinase SDRs is less than that of kinase catalytic residues, suggesting that a modest level of orthologue divergence may have occurred during evolution.

To test this, the naive Bayes models described above were used to determine whether the variation observed is likely to alter the kinase specificity. Towards this end, average posterior probabilities were calculated for each orthologous group of kinases, where the human

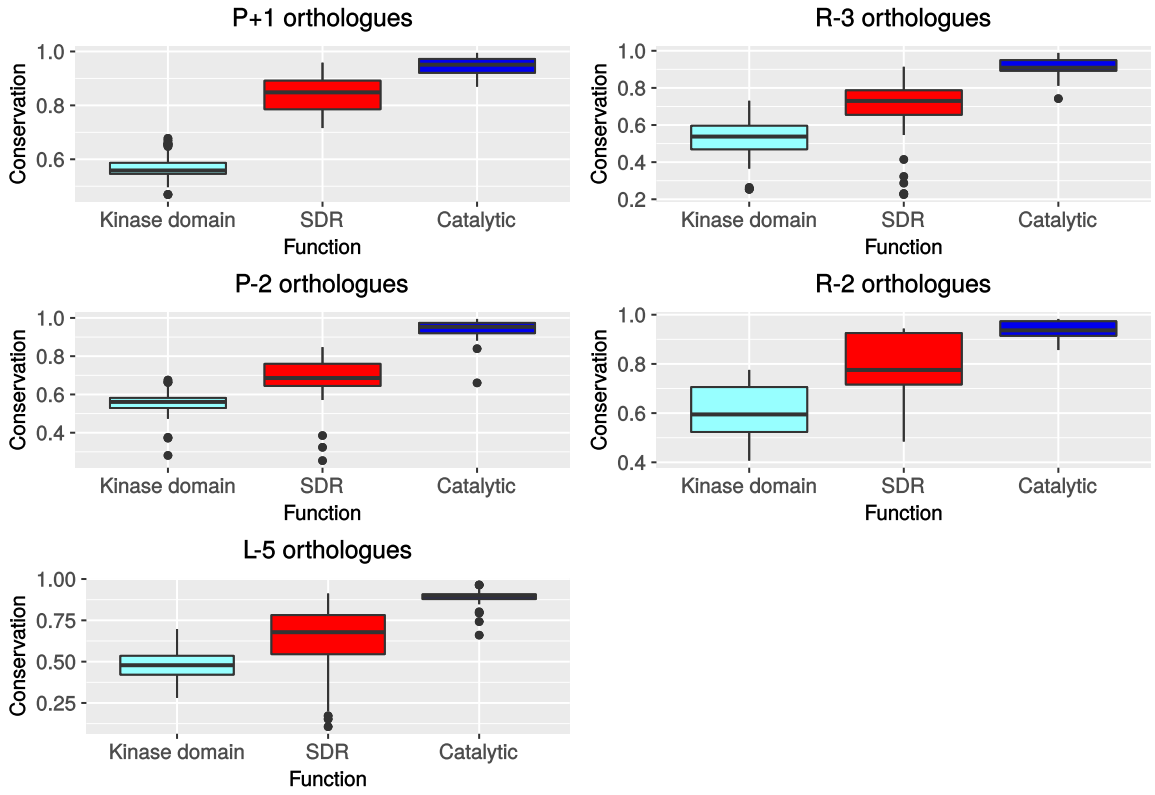


Fig. 3.1 Conservation of kinase domain residues, SDRs, and catalytic residues across orthologues. Each data point represents the average conservation (among kinase domain positions, SDRs, or catalytic residues) for an alignment of orthologous kinases where the human kinase is of corresponding specificity (P+1, R-3, R-2, etc.). The conservation values were calculated using a protein substitution matrix (see *Methods* chapter Section 6.2.1)

kinase is of corresponding specificity (P+1, R-3, P-2, etc.) As above, the results suggest that the specificity of orthologues is highly conserved for all five specificities (Figure 3.2). Indeed, only 5% of orthologous groups overall were found to have diverged in specificity (i.e. the average posterior probability across orthologues was less than 0.5). For example, the Wee2 protein in human features a hydrophobic -5 binding pocket, but retrieval of kinase orthologues reveals this to be the case mainly for vertebrate sequences only, due to the presence of an E at position 189 in most other sequences. Likewise, the CAMK kinase TSSK3 is predicted to be an R-3 kinase in human but not for its non-vertebrate orthologues due to non-conservative substitutions at positions 82, 86, and 162.

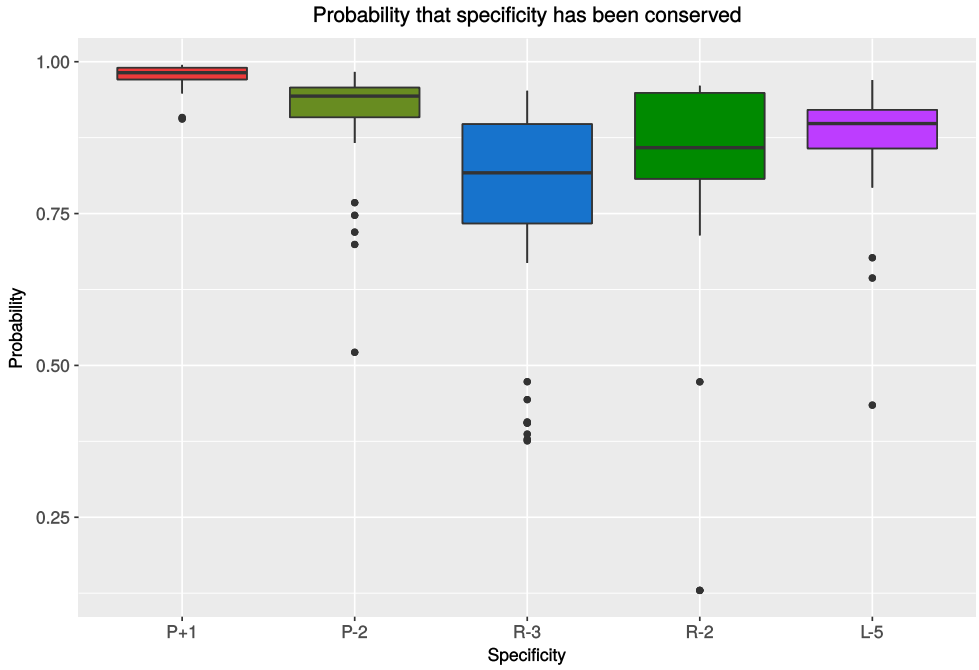


Fig. 3.2 For each of the five specificities represented, each data point represents the average posterior probability (PP) among orthologues of a human kinase of corresponding specificity (P+1, R-3, R-2, etc), according to the naive Bayes models presented in *Chapter 2*. This is the posterior probability that the kinase sequence has the same specificity as its human orthologue (P+1, R-3, R-2, etc).

Finally, I investigated the extent to which kinase specificity is predicted to differ between one-to-one orthologues and one-to-many orthologues. One-to-many orthologues differ from one-to-one orthologues in the sense that the former represents orthologues that have undergone at least one round of gene duplication subsequent to the speciation event that generated them. One-to-one orthologues however are present only as single copies for the two species being considered. I therefore hypothesised that one-to-many kinase orthologues would be more like to have diverged in their specificity than one-to-one orthologues of the same age,

assuming that gene duplication will have led to a relaxation of selective constraint upon gene function (Altenhoff et al., 2012; Koonin, 2005; Rogozin et al., 2014). In Figure 3.3, the divergence of one-to-many orthologues is compared with one-to-one orthologues of the same species, again using the naive Bayes predictive models presented in *Chapter 2*. For each data point, the mean posterior probability among one-to-many orthologues of a given species was subtracted from the mean posterior probability among one-to-one orthologues of the same species. Therefore, if one-to-many orthologues were more divergent on average then this net difference would be greater than 0. The results provide no evidence for the hypothesis that one-to-many orthologues will have been more likely to have diverged than one-to-one orthologues (Figure 3.3). This finding is discussed further in the *Discussion* section.

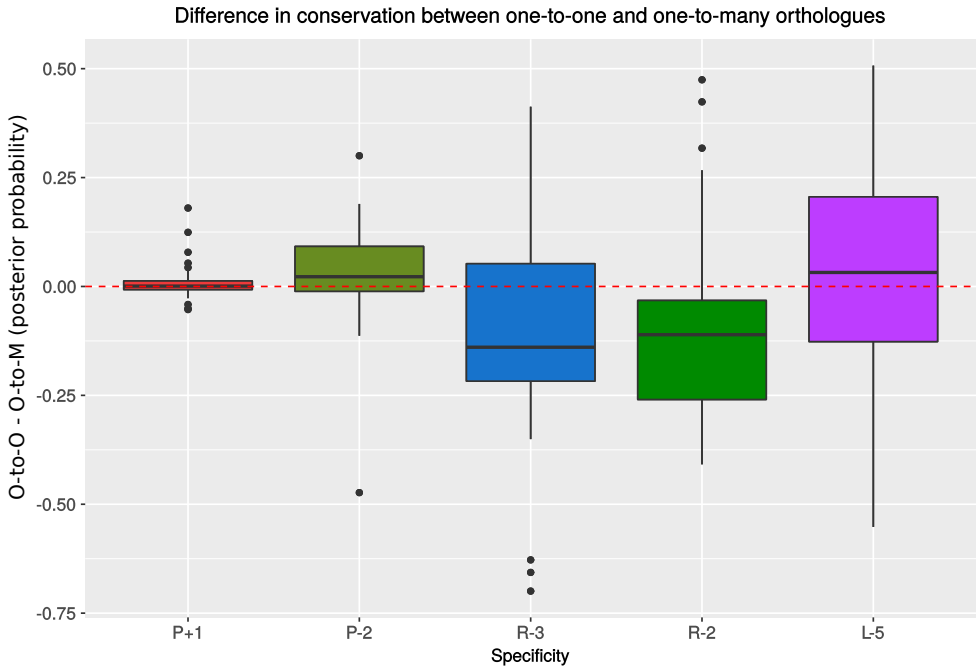


Fig. 3.3 Divergence in kinase peptide specificity for one-to-many orthologues and one-to-one orthologues. Each data point represents the difference in average posterior probability between one-to-one orthologues and one-to-many orthologues (O-to-O - O-to-M) for orthologues of the same species

3.2.2 Conservation of empirical specificities

I next used empirical models of specificity to assess the conservation of kinase specificity. As described in *Chapter 2*, substantive data on kinase specificity currently only exists for human, mouse, and *S. cerevisiae*. The objective of this analysis was therefore to determine the extent to which the specificity of yeast kinases differs from that of their orthologues in

mouse and human. The advantage of this approach relative to the analysis conducted above is that a broader range of kinase specificities can be considered, and that all positions flanking the phosphoacceptor can be taken into account rather than just a single position at a time.

For human and mouse kinases, I use the 101 phosphosite-based specificity models that were described in *Chapter 2*. For the yeast specificity models, I use the 18 phosphosite models described in *Chapter 2* in addition to 61 specificity models derived from the peptide library screen presented in (Mok et al., 2010). The Mok *et al* PWMs were required to extend the sample size of *S. cerevisiae* PWMs, although the accuracy of these 61 models is not clear as the cross-validation procedure described in *Chapter 2* can not be applied to them. The human and yeast orthologous kinases often share similar features, as represented by two examples of orthologous groups given in Figure 3.4.

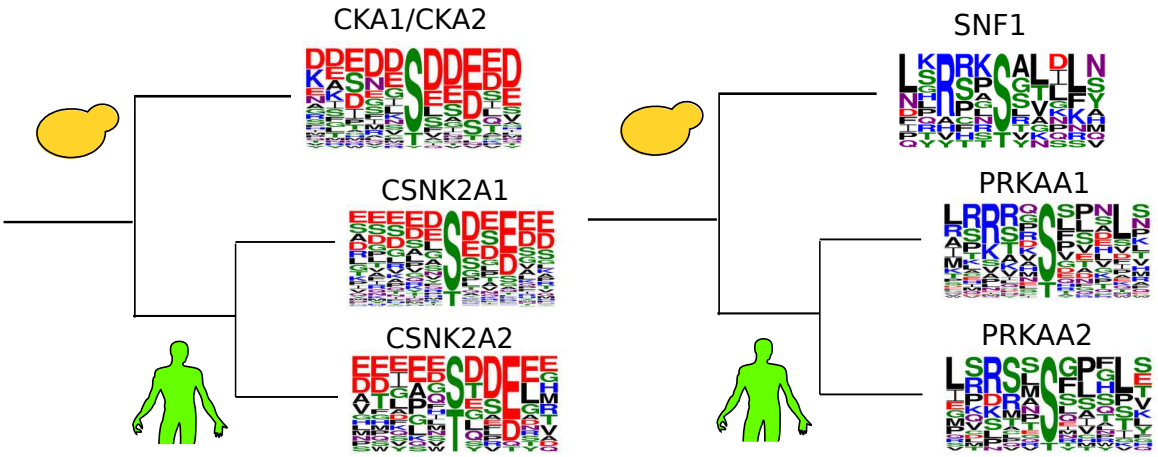


Fig. 3.4 Specificity logos for the casein kinase II orthologous group (left) and the SNF1/PRKAA orthologous group (right). In both examples, key features of specificity have been conserved between the human and yeast orthologues

Overall, 14 orthologous groups were identified that contained specificity models from both human/mouse and *S. cerevisiae*. The divergence in specificity between orthologues was computed by calculating the Frobenius distance between the two specificity matrices (human/mouse and *S. cerevisiae*). This metric represents the sum of the squared element-wise distances between models (followed by square-rooting) and so increases when specificity diverges between kinases. To determine whether the differences are significant, they are compared to a null distribution of Frobenius distances calculated for kinases within the same *Family* and species (n=218). This assumes that kinase specificity is generally conserved at the *Family* level within a species. This assumption is discussed further in the *Divergence of kinase peptide specificity* subsection (subsection 3.3.4). I also calculated p-values from a

more stringent null distribution generated by comparing kinases within the same *Subfamily* and species (n=110).

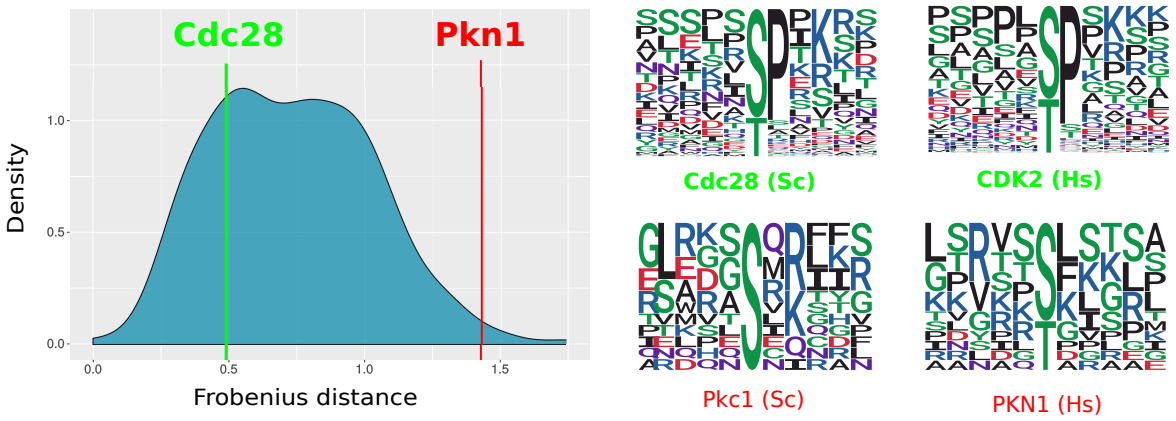


Fig. 3.5 The divergence in specificity between yeast kinases and their human/mouse orthologues is computed by calculating the Frobenius distance between the yeast and human/mouse specificity matrices. This is then compared to a null distribution of Frobenius distances from kinases belonging to the same *Family* and species. An example of specificity conservation (Cdc28/CDK2: top) and divergence (Pkc1/PKN1: bottom) is given on the right hand side. Hs = *Homo sapiens*, Sc = *Saccharomyces cerevisiae*.

As expected, the results of this investigation suggest that kinase specificity is highly conserved between human/mouse and *S. cerevisiae* despite an estimated divergence time of 1.3 billion years (Vlastaridis et al., 2017). Overall, I identify only 3 orthologous groups with evidence of orthologue divergence at a significance level of 0.05 (Table 3.1). Very similar results were found when using a null distribution constructed from within-*Subfamily* Frobenius distances. In Figure 3.5, an example is given of orthologue conservation and divergence for the yeast kinases Cdc28 and Pkc1, respectively. For the second example, the +2 R/K signature is present in the yeast kinase (PKC1) but seems to absent in one of its human orthologues (PKN1).

This analysis was then repeated but this time excluding positions from the PWMs without any evidence of substrate selectivity. The purpose of this was to exclude random variation (i.e. noise) from the distance calculations so that only positions contributing to kinase specificity are considered. The results of this analysis are consistent with those generated when considering the full matrix for comparison between orthologues, as I identify only two orthologous groups with evidence of orthologue divergence at a significance level of 0.05 (Table 3.1).

Finally, I repeated the analysis but instead grouped human/mouse and yeast kinases according to their *Family*/*Subfamily* assignations instead of by their orthology predictions. The purpose of this was to determine whether the conclusions stated above are robust to the

method used to group yeast kinases with their human/mouse counterparts, as orthology predictions themselves are not without error (Altenhoff et al., 2016; Vallender, 2009). For the results shown in Table 3.2, the yeast kinases are matched instead to human/mouse kinases belonging the same *Family* or *Subfamily* according to manual annotations (Manning et al., 2002b). The results of this analysis are largely concordant with the results given in Table 3.1. Specifically, the results reveal that 4/18 yeast kinases show evidence of divergence when comparing the full specificity matrices. The evidence overall therefore suggests that orthologous kinases between human/mouse and *S. cerevisiae* tend to be conserved in terms of their active site specificity.

Orthologous group	Family	Subfamily	Family (bit>0.75)	Subfamily (bit>0.75)
Sch9/Ypk1/Ypk2	0.078	0.081	0.083	0.054
Ipl1	0.22	0.22	0.13	0.081
Cmk1/Cmk2	0.0046	0.0	0.009	0.009
Ccd28	0.11	0.09	0.13	0.073
Cka1/Cka2	0.14	0.136	0.30	0.25
Yak1	0.06	0.055	0.13	0.072
Kss1/Fus3/Slt2	0.18	0.2	0.13	0.082
Hog1	0.06	0.072	0.046	0.045
Ste20	0.61	0.54	0.70	0.6
Pkh2	0.22	0.22	0.88	0.82
Cdc5	0.47	0.42	0.34	0.27
Snf1	0.35	0.35	0.64	0.53
Sky1	0.014	0.0	0.48	0.38
Pkc1	0.014	0.0	0.17	0.13

Table 3.1 P-values generated by comparing the maximum distances within an orthologous group against a null distribution of Frobenius distances. Separate null distribution were generated for within-*Family* and within-*Subfamily* comparisons (same species also; see the main text), and also for comparisons where noisy PWM positions were filtered out (bits<0.75 positions were filtered). p-values less than 0.05 are coloured in red.

Grouping	Family	Subfamily	Family (bit>0.75)	Subfamily (bit>0.75)
AKT (family)	0.09	0.091	0.06	0.045
AUR (family)	0.22	0.22	0.13	0.081
CAMKI (family)	0.0046	0.0	0.0018	0.009
CDK2/3 (subfamily)	0.11	0.091	0.13	0.072
CDK5 (subfamily)	0.36	0.38	0.06	0.045
CK2 (family)	0.14	0.136	0.3	0.25
DYRK (family)	0.06	0.055	0.13	0.072
ERK1 (subfamily)	0.18	0.2	0.13	0.082
Nek (family)	0.046	0.045	0.046	0.045
p38 (subfamily)	0.092	0.091	0.0505	0.045
PAKA (subfamily)	0.56	0.5	0.61	0.5
PDK1 (family)	0.22	0.22	0.89	0.82
PKA (family)	0.60	0.54	0.67	0.555
PLK (family)	0.092	0.091	0.13	0.072
Rad53 (family)	0.52	0.45	0.24	0.21
AMPK (subfamily)	0.35	0.35	0.64	0.53
SRPK (family)	0.014	0.0	0.48	0.38
PKC (family)	0.014	0.0	0.14	0.09

Table 3.2 P-values generated by comparing the maximum distances within a *Family* or *Subfamily* against a null distribution of Frobenius distances. Separate null distribution were generated for within-*Family* and within-*Subfamily* comparisons (same species also; see the main text), and also for comparisons where noisy PWM positions were filtered out (bits<0.75 positions were filtered). p-values less than 0.05 are coloured in red.

3.3 The evolution of kinase *Families* and *Subfamilies*

The research presented above concerns kinases separated originally by a speciation event (i.e. kinase orthologues). This following analysis instead focuses on the evolution of the kinase domain following a gene duplication event. This is achievable by studying the evolution of kinase *Families* and *Subfamilies*, which are normally generated by gene duplications. The objective of the research was therefore to identify residue substitutions following gene duplication that are responsible for the functional divergence of kinase *Families* and *Subfamilies*.

Towards this end, I have adapted a previously published method for the identification of functionally divergent residues using ancestral sequence predictions. The *BADASP* (Burst After Duplication with Ancestral Sequence Predictions) tool is a phylogenetic method that aims to identify divergent residues that arise following gene duplication (Edwards and Shields, 2005). Residues are predicted to be divergent if they are conserved within the clade of interest but differ between sister clades, as inferred using ancestral sequence reconstructions. This method is typically applied towards a given family of interest but I have automated its implementation to predict divergent residues for all kinase *Families* in a global kinase phylogeny. The adapted method used for this analysis is presented in Figure 3.6.

A score is generated for each position in the alignment. RC (recent conservation) represents the sequence conservation for the clade of interest (clade A). *AC* represents the conservation of ancestral nodes for the clade of interest (clade A) and the nearest sister clade (clade B); this is given as a 1 if the most likely residues are identical to each other and a -1 otherwise. As an innovation here I also weight the predicted scores to account for the uncertainty in ancestral sequence predictions. This is represented by the *p(AC)* term in the equation. For matching residues (AC=1), this is the posterior probability of the predicted residue for clade B; for differing residues (AC=-1) this is the summed posterior probability of all residues in clade B besides from the predicted residue for clade A. Therefore, scores for suspected divergence would be down-weighted if there is ambiguity concerning the nature (matching or mismatching) of the clade B ancestor.

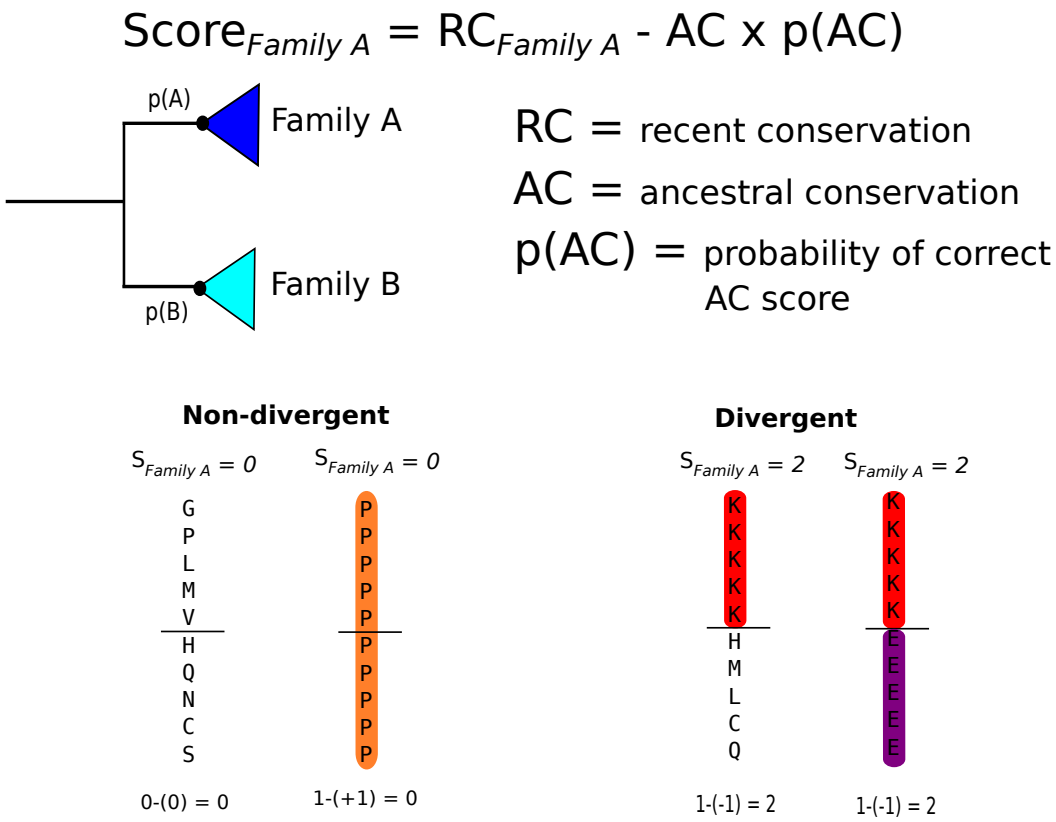


Fig. 3.6 Explanation of the scoring method for the systematic identification of kinase functionally divergent residues. A score is generated for each position in the alignment. RC: sequence conservation within the clade of interest (Family A). AC: ancestral conservation between predicted ancestral residue for the clade of interest (Family A) and predicted ancestral residue for the sister clade (Family B). p(AC): probability that the ancestral sequence predictions are correct

3.3.1 Functional divergence at the *Family* level

To begin the evolutionary analysis, a global kinase phylogeny was constructed from the kinase domain sequences of 9 different opisthokont species (*H. sapiens*, *M. musculus*, *S. purpuratus*, *D. melanogaster*, *C. elegans*, *A. queenslandica*, *M. brevicollis*, *S. cerevisiae*, *C. cinerea*). All kinomes had been manually annotated previously and so the computational prediction of *Group*/*Family*/*Subfamily* classifications was not necessary (Manning et al., 2002b). Functionally divergent residues across all *Families* were then predicted using the scoring approach described above. In Figure 3.7, the results of this analysis have been aggregated by counting the total number of observed 'switches' at each site across all *Families* considered. I consider a residue to have switched at a particular *Family* if its score is above the 95th percentile of scores for all residues compared at the *Family* level, which in this case is 1.79. Domain positions with more than 8 observed switches (90th percentile) are considered to be 'frequent switchers'. This plot reveals that the distribution of switch events is non-uniform. In particular, I find an especially high number of switches within or close to the kinase activation segment, but also at the αC helix, the β5-αD region, and the αF-αG regions, all of which had been discussed in *Chapter 2* in the context of kinase specificity.

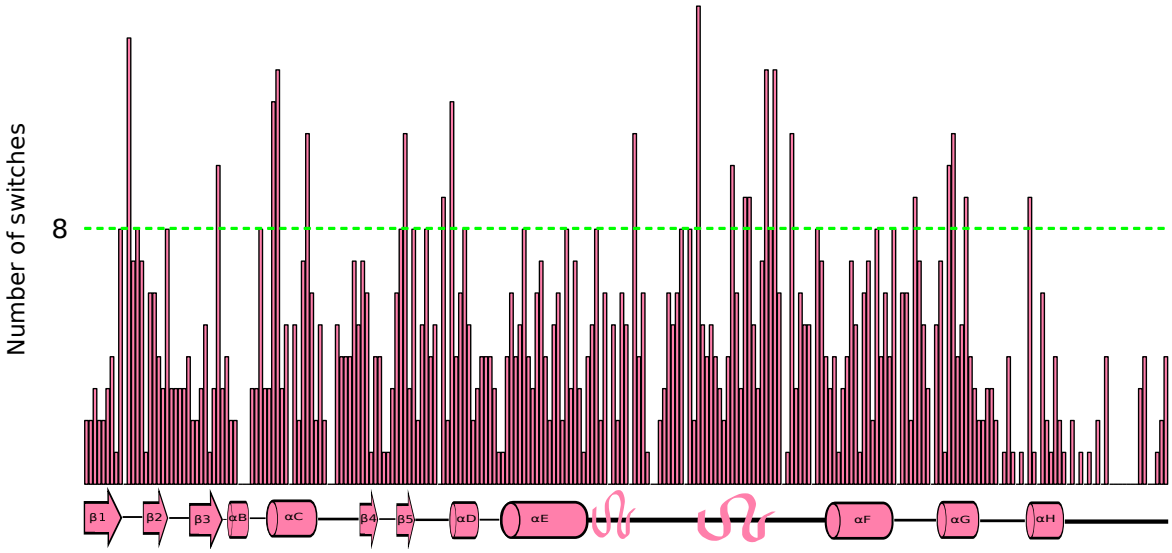


Fig. 3.7 Number of switches for each residue in the protein kinase domain when kinases are compared at the *Family* level

To aid in the structural interpretation of the results, functions have been assigned to each residue of the kinase domain where possible. The *Catalytic* and *Regulatory* categories are self-explanatory and are defined as they were for the analysis of cancer mutations in *Chapter 2*. I define as a *Proximal* residue any residue within 4 Å of the substrate peptide; substitutions of these residues may affect kinase specificity. *Distal* residues refer to putative SDRs listed

in Figure 2.4 that are not within 4 Å of the substrate peptide. Finally, *Interactions* residues refers to those sites that are often found to be in contact with other protein domains in co-crystal structures (*see Methods* chapter Section 6.2.3). *Other* represents the complement of the kinase domain against these 5 previous sets.

In Figure 3.8, the total number of switches per residue is grouped according to the functional annotation of the residues. The majority (14/21) of frequently-switching residues can be assigned to a functional category (*Catalytic*, *Proximal*, *Regulatory*, *etc.*), which is more than would be expected by chance (p_{family}=0.0083 ; Fisher's Exact Test, one-sided). This suggests that the approach here successfully predicts residues that are of functional relevance for the kinase domain. I also find that the median number of substitutions for *Proximal* residues is significantly higher than for residues with no assigned function (*Other* residues) (Mann-Whitney, one-tailed, p = 1.1×10^{-5}). This analysis therefore suggests that kinase *Family* evolution is dominated by changes that are likely to affect peptide specificity at the active site.

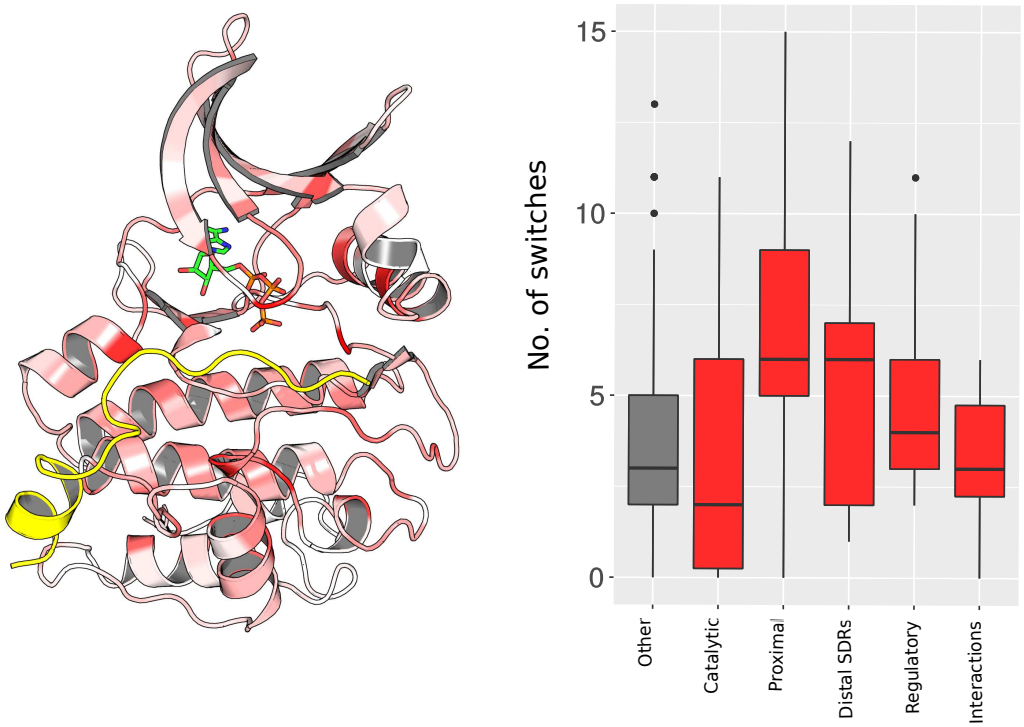


Fig. 3.8 Functional analysis of divergent residues at the *Family* level. *Left:* mapping of the number of switches per residue to the protein kinase domain (*PDB: 1ATP*). Deeper shades of red denote a higher number of switches. *Right:* The number of switches per residue is grouped according to the functional annotation of residues. Descriptions of the categories are given in the main text

3.3.2 Functional divergence at the *Subfamily* level

The above analysis was then repeated at the *Subfamily* level. This gives a more recent overview of kinase functional evolution considering that *Subfamilies* generally emerged later in evolution than kinase *Families*. As before, for a given comparison between *Subfamilies*, a residue is considered to have 'switched' if its score is greater than the 95th percentile of all scores for comparisons at the *Subfamily* level, which in this case is 1.90.

In Figure 3.9, the total number of observed switches at the *Subfamily* level is plotted for every residue in the protein kinase domain. I consider domain positions with more than 7 observed switches (90th percentile) to be 'frequent switchers'. Overall, the results are similar to what was observed for the *Family*-level analysis, as the switch events are non-randomly distributed and cluster towards particular domain regions (activation segment, αC helix, and the αF-αG regions).

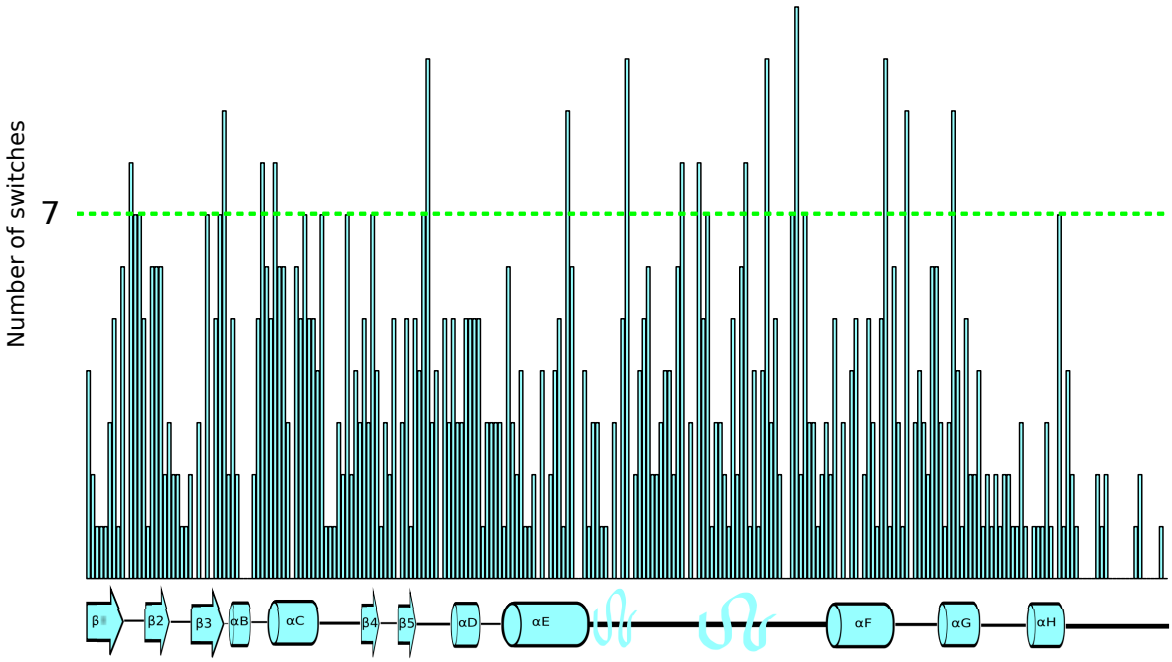


Fig. 3.9 Number of switches for each residue in the protein kinase domain when kinases are compared at the *Subfamily* level

In Figure 3.10, the total number of switches per residue at the *Subfamily* level is grouped according to the functional annotation of the residues. As with the analysis at the *Family* level, the majority (11/15) of frequently switching residues can be assigned to a functional category ($p_{subfamily}$=0.0068 ; Fisher's Exact Test, one-sided). Moreover, *Proximal* residues on average switch significantly more often than is the case for residues without an assigned function (Mann-Whitney, one-tailed, p = 1.0×10^{-4}). I therefore predict that changes in ki-

nase peptide specificity are more common than any other functional change at the *Subfamily* level, although the overlap here with 'Other' residues is greater than for *Families*. This is consistent with the detailed analysis of the GRK *Family* described below, where several substitutions between GRK *Subfamilies* were observed that would be likely to affect kinase specificity.

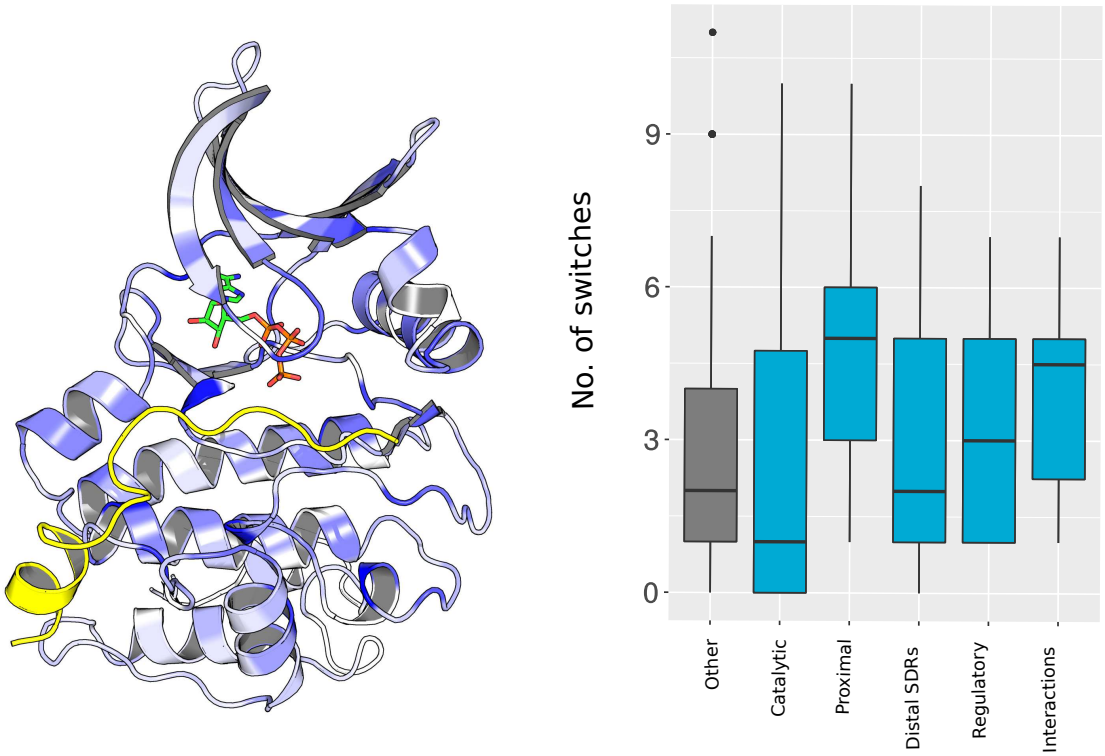


Fig. 3.10 Functional analysis of divergent residues at the *Subfamily* level. *Left:* mapping of the number of switches per residue to the protein kinase domain (*PDB: 1ATP*). Deeper shades of blue denote a higher number of switches. *Right:* The number of switches per residue is grouped according to the functional annotation of residues. Descriptions of the categories are given in the main text

3.3.3 Examples of functional divergence

From this analysis, residue divergence scores have been calculated for 99 kinase *Families* and 88 kinase *Subfamilies*. I expect the results generated for any given *Family* or *Subfamily* to be of interest to specialist kinase researchers. Examples are given below to illustrate the different mechanisms by which residue divergence may affect kinase function. The targeting of these regions with drugs could enable the selective inhibition of a kinase *Family* or *Subfamily* of interest.

Substrate docking

Three divergent residues identified for the SRPK (SR-protein kinases) *Family* map to the αF-αG region. The SRPKs are known primarily for their involvement in the control of mRNA splicing, but they also have roles in the cell cycle, metabolic regulation, and the organisation of chromatin (Giannakouros et al., 2011). They are members of the CMGC *Group* but are generally basophilic rather than proline-directed (Wang et al., 1998). This is in agreement with the results generated here, as 2 of the 3 switches mapping to the αF-αG region generate additional negative charges. Analysis of a SRPK-peptide structure reveals that the 3 residues identified bind to arginine residues in the substrate peptide (Ngo et al., 2005), suggesting that the substitutions presented in Figure 3.11 (top) increase the affinity of the kinase for the substrate (*PDB: 1WBP*). Since the kinase residues are located outside the active site, then the kinase-substrate contacts can be considered as docking interactions.

Regulation of activity

For the analysis of CDC7 *Family* (CMGC), I identify two divergent residues that are likely to affect the function of kinases within this *Family* (Figure 3.11 bottom). Both map to the N-terminal lobe of the kinase and interact with the DBF4 protein (*PDB: 4F9A*). Specifically, arginine at kinase position 68 interacts with D328 on DBF4, and an aspartate at kinase position 71 interacts with H315 on DBF4 (Hughes et al., 2012). Both interactions are likely to increase the affinity of the kinase for DBF4. This is relevant for the regulation of CDC7 activity as DBF4 is required for the activation of CDC7 (Matthews and Guarné, 2013), which is necessary for the initiation of DNA replication.

Peptide specificity

The Polo-like Kinases (PLKs) belong to the 'Other' *Group* and are generally important for mitotic entry and mitotic exit (Glover et al., 1998). They are acidophilic kinases but form a sister clade to the Aurora kinases (*kinase.com*), which are themselves basophilic (Brown et al., 2004; Johnson, 2011; Salvi et al., 2012). SDR divergence was therefore likely necessary for the functional specialisation of the Aurora and PLK *Families*. Indeed, I identify five high-scoring residues that are in locations likely to affect kinase specificity when mutated (Figure 3.12 bottom). Four of these substitutions would serve to increase the positive charge in the PLK active site. Notably, 3 of the 5 positions described (126, 127, and 162) have also undergone mutation during the evolution of the GRK *Family* (described below). The two *Families* are analogous in the sense that they are acidophilic kinases that have evolved from a basophilic ancestor.

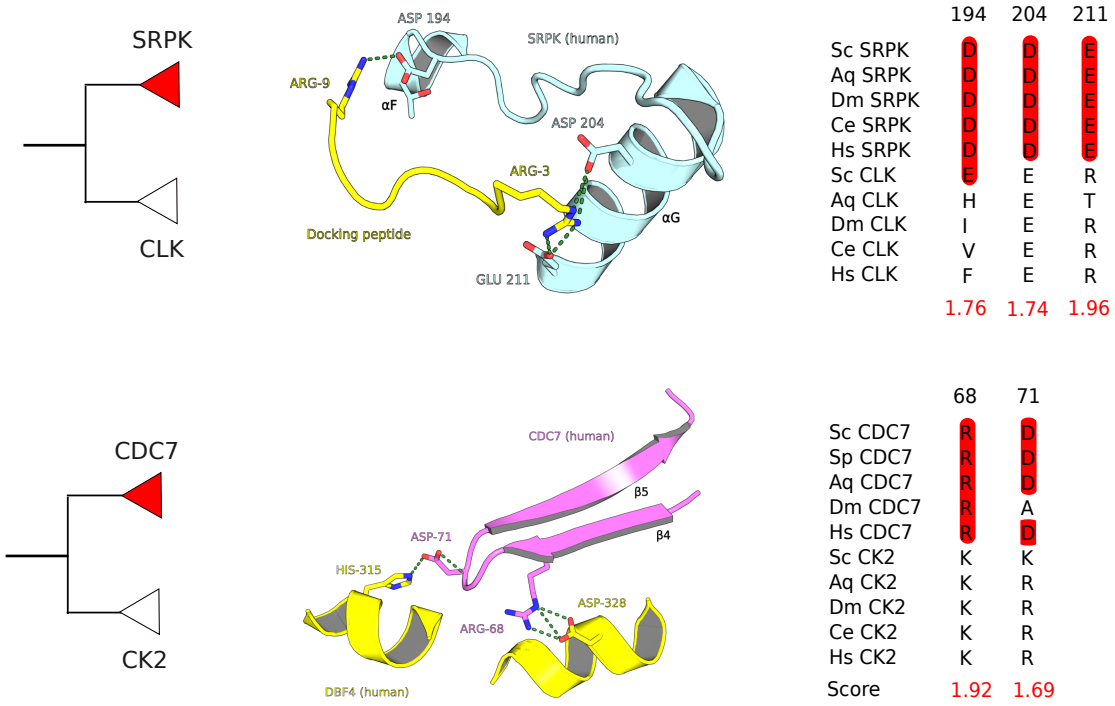


Fig. 3.11 Functionally divergent residues in the SRPK and CDC7 *Families*. Top: functionally divergent residues identified for the SRPK *Family* may influence substrate docking. Bottom: divergent residues identified for the CDC7 *Family* are likely to affect kinase activity via their proximity to the activator DBF4 protein.

Kinase activity and specificity

For the CAMK2 (CAMK) *Group*, I detected changes in the kinase domain that are likely to affect both the specificity of the kinase and its regulation. As discussed in *Chapter 2*, CAMK2 kinases have a +2 D/E preferences that is likely mediated by positively charged residues in the αC helix that bind at the +2 position (*PDB: 5H9B*). Both residues (positions 41 and 44) score highly using the approach described above and are therefore implicated as residues responsible for the functional divergence of CAMKII (Figure 3.12 top).

Unlike most other kinases, CAMKII kinases do not require phosphorylation at their activation loops to be activated. In CAMKII kinases, the HRD arginine that is usually stabilised by a phosphoserine/phosphothreonine instead binds to the carbonyl group of a phenylalanine at position 155. The residue directly C-terminal to this phenylalanine is a glycine (GLY 156) and is strongly implicated here as a functional determinant. This glycine replaces the threonine residue that is normally phosphorylated. I suggest here that the increased flexibility of the glycine backbone enables the observed contact between R122 and F155, and is therefore critical for the constitutive stability of the activation loop.

The importance of this glycine for CAMK2 has been demonstrated previously (LeBoeuf et al., 2007), and therefore validates the prediction here. However, the underlying mechanism is contested as it has been suggested that the glycine is instead necessary to prevent unfavourable interactions between CAMKII kinase domains in the holoenzyme (Chao et al., 2011).

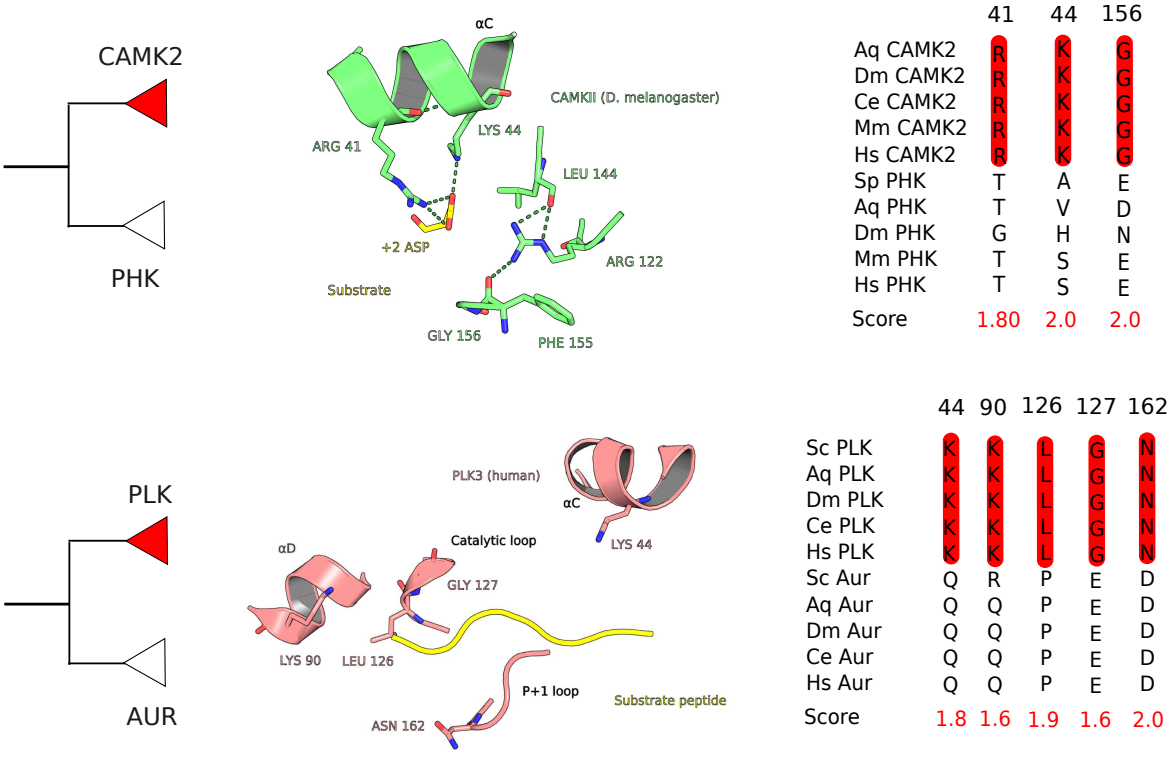


Fig. 3.12 Functionally divergent residues in the CAMK2 and PLK families. Some of the divergent residues identified for the CAMK2 *Family* map to substrate binding regions and may affect kinase specificity and regulation. Likewise, some divergent residues identified for the PLK *Family* map to the substrate-binding region and likely confer a preference for D/E residues.

3.3.4 Divergence of kinase peptide specificity

The results described above suggest that the emergence of new kinase *Families* and *Subfamilies* often coincides with the divergence of kinase SDRs. These predictions would imply that kinase specificity will diverge also at the level of the *Family* and *Subfamily*. To test these prediction, the kinase specificity models described in *Chapter 2* were compared systematically at the level of the kinase *Family* and *Subfamily*. At the *Family* level, each kinase was compared to the kinases in the same *Family* and also to kinases of different *Families* but belonging to the same *Group*. Likewise at the *Subfamily* level, each kinase was compared to the kinases in the same *Subfamily* and also to kinases belonging to different *Subfamilies* but to the same *Family*. In each case, the Frobenius distance is used to calculate the distance between the specificity matrices being compared. The Frobenius distance represents the sum of the squared distances between matrix elements (followed by square-rooting) and so would be relatively small when the two kinases being compared are similar in specificity.

The results of this analysis are presented in Figure 3.13. As expected, I find that kinase specificity often diverges at the level of the *Family* as kinase specificity models within a *Family* are more similar on average than they are between kinases belonging to different *Families* but the same *Group*. I find this to be the case also when comparing kinases at the *Group* level. However, this is difficult to relate to patterns of kinase functional divergence because there are too few kinase *Groups* to attempt a systematic analysis of residue divergence. At the *Subfamily* level, however, the distribution of matrix distances within a *Subfamily* overlaps strongly with the distribution of distances for kinases between *Subfamilies* (but belonging to the same *Family*). Contrary to expectation, there is therefore little evidence for the divergence of kinase specificity at the *Subfamily* level, suggesting that kinase specificities tend to be conserved within *Families*. Possible reasons for this are given in the *Discussion* section.

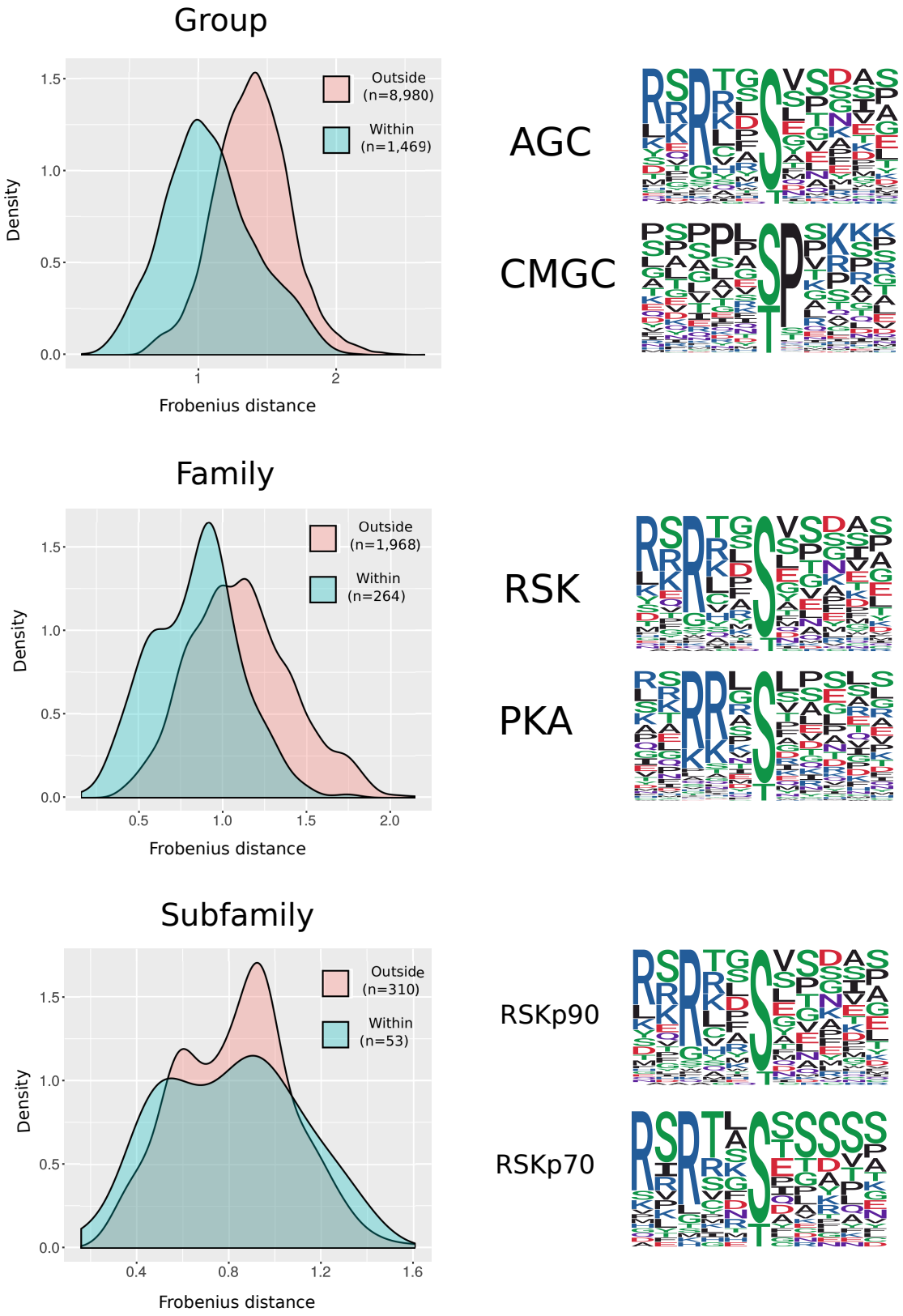


Fig. 3.13 The Frobenius distance represents the sum of the squared element-wise differences between two kinase specificity matrices (followed by square-rooting). Distributions of these distances within and between classifications were calculated at the *Group*, *Family*, and *Subfamily* level. On the right hand side, example are given of two kinases belonging to different *Groups*, *Families*, and *Subfamilies*, respectively

3.4 Evolution of the G-protein coupled receptor kinases

3.4.1 The GRK *Family* of protein kinases

In the previous sections, it was demonstrated that kinase specificities often differ between *Families*, and that kinase evolution at the *Family* level is dominated by changes to protein kinase SDRs. This process is characterised here in detail for a single *Family* using the ancestral sequence reconstruction of kinase domain sequences. Sequence changes between the associated *Subfamilies* are also characterised. The G protein-coupled receptor kinase (GRK) family was chosen for this purpose as it is an acidophilic (i.e. D/E-preferring) *Family* located within a generally basophilic *Group* (Pearce et al., 2010), and so its emergence represented a dramatic shift in kinase specificity at the active site. The knowledge of kinase SDRs acquired in *Chapter 2* is used for the interpretation of the observed kinase SDR substitutions.

The GRKs are one of 15 *Families* belonging to the AGC *Group* (Manning et al., 2002b). They are named for their ability to target G-protein coupled receptors (GPCRs) as substrates, although GRKs can also phosphorylate non-GPCR proteins (Gurevich et al., 2012). The *Family* itself is divided into two *Subfamilies*: the first called the BARK (β-adrenergic receptor kinase) *Subfamily* and the second *Subfamily* also called GRK. The BARK *Subfamily* comprises GRK2 (ADRBK1) and GRK3 (ADRBK2) in human, whereas the GRK *Subfamily* comprises GRK1 (rhodopsin kinase), GRK4, GRK5, GRK6, and GRK7 (Manning et al., 2002b). The GRK *Subfamily* can be divided further into two clades on the basis of sequence similarity, one consisting of GRKs 1 and 7 and the other consisting of GRKs 4, 5, and 6 (Mushegian et al., 2012).

Although a constituent *Family* of the AGC *Group*, the GRKs do not exhibit the characteristic -2/-3/-5 basic residue preference often found in other kinases of this *Group* (Pearce et al., 2010). The GRKs are instead characterised by varying levels of D/E preference between members of the *Family* (Asai et al., 2014; Lodowski et al., 2006; Onorato et al., 1991). Indeed, an inspection of a GRK sequence alignment by other researchers has revealed that many of the D/E residues thought to be determinant for R-3/R-2 preferences have undergone non-conservative substitutions in these kinases (Lodowski et al., 2006).

Here I have taken a taxonomically broad sample of kinase sequences to generate a comprehensive phylogeny of the GRK *Family*. From this, a maximum-likelihood reconstruction of all ancestral sequence states has been performed. This has enabled the diversity of substrate preferences among extant GRKs to be rationalised with respect to the set of 'most probable' SDR substitutions occurring along the phylogeny branches, on the basis of the current understanding of how kinase structure relates to kinase specificity (as discussed in *Chapter 2*).

3.4.2 Phylogeny of the GRK domain

A maximum-likelihood representation of the GRK phylogeny is represented in Figure 3.14. Protein sequences were first retrieved from a taxonomically broad set of non-redundant proteomes (representative proteomes) (Chen et al., 2011), and then each proteome was queried with a hidden Markov model (HMM) of the GRK domain to retrieve GRK sequences. The *Subfamily* classifications of each GRK were then predicted using *Kinannote* (Goldberg et al., 2013). Sequences of the basophilic RSK *Family* kinases – the *Family* most similar in sequence to the GRKs – were also included as an expected outgroup in the phylogeny, as were two kinases of the basophilic PKA *Family*. The kinase domain sequences (GRK kinases plus outgroups) were then aligned and filtered to remove pseudokinases and redundant sequences (97% threshold), resulting in 163 sequences to be used for phylogenetic reconstruction.

The phylogeny presented here is generally concordant with a GRK phylogeny published in a previous study (Mushegian et al., 2012). The BARK *Subfamily* emerge as a single clade with relatively high confidence (57 out of 100 bootstrap trails), with the exception of a single sequence (A0A0L0CKW0 in *Lucilia cuprina*) classified as a BARK by Kinannote. The GRK *Subfamily* is also recapitulated in the most probable tree, and in 99 out of 100 bootstrap trials. A single GRK representative is present in the choanoflagellate *Monosiga brevicollis*, and also in the closely-related unicellular filasterean *Capsaspora owczarzaki*. However, both sequences are found in separate clades (*Monosiga* kinase in the BARK clade and the *Capsaspora* kinase in the GRK clade), suggesting that gene duplication probably occurred before the emergence of the metazoa over 600 million years ago but one copy was lost in each respective lineage (choanoflagellate and filasterea) (Mushegian et al., 2012). A duplication of the ancestral GRK gene in vertebrates likely generated the partition between the GRK1/7 and GRK4/5/6 clades (supported here by 99 bootstrap replicates), with subsequent duplications contributing to the gene diversity within clades. Both gene duplications – separating GRK 1/7 from GRK 4/5/6 and generating new copies within the clades – are speculated to be a result of two rounds of whole-genome duplication events during the evolution of the vertebrates (Mushegian et al., 2012).

A small clade of heterokont sequences is also reconstituted in this analysis, as in the previous phylogenetic study of GRK (Mushegian et al., 2012). However, I also discovered a small clade of GRK sequences belonging to organisms of the eukaryotic super-group rhizaria. The heterokont and rhizaria clades are both strongly supported as monophyletic groupings (99 and 100 supporting bootstrap replicates, respectively) and both show greater sequence similarity to BARK sequences than to GRK *Subfamily* sequences with respect to the kinase domain. Notably, the heterokont and rhizarian sequences are predicted to be orthologous to the sequences found in human, suggesting that they were separated originally

by a speciation event in a very early eukaryotic ancestor. The GRK and BARK *Subfamilies* however are present in most animal species and therefore must have been generated by a gene duplication event, as is confirmed by *ensembl Compara* orthology/paralogy predictions. These findings suggest that the ancestral gene duplication leading to the emergence of the GRKs occurred in a common ancestor of the filozoa, heterokonts, metazoa, and rhizaria early during eukaryote evolution, but was subsequently lost in multiple lineages such as in green plants and in fungi. This vertical form of descent is implied by the phylogeny presented in Figure 3.14. The GRK ancestral gene may alternatively have been transmitted to the metazoa (or *vice versa*) via a process of horizontal gene transfer (HGT).

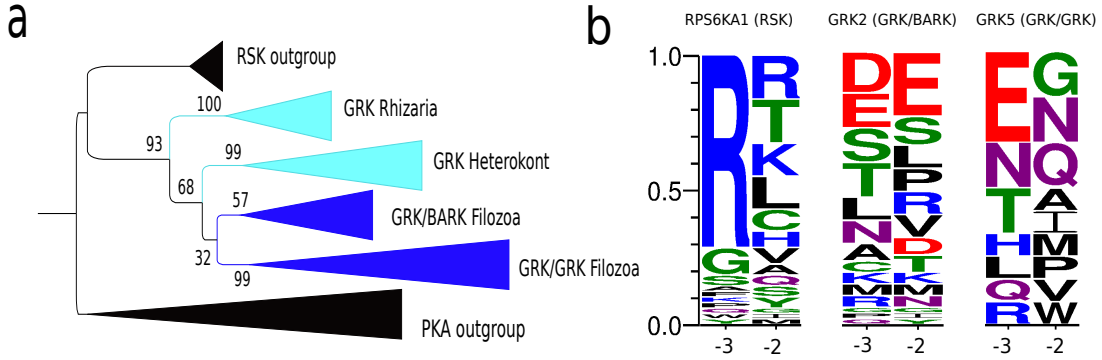


Fig. 3.14 a) Maximum likelihood phylogeny of the GRK kinase domain, including RSK and PKA kinases as outgroups. Numbers on the internal nodes refer to the number of bootstrap replicates (out of 100) that support the partitions presented here. B) Specificity logos at the -2 and -3 positions for GRK2, GRK5, and an outgroup kinase belonging to the RSK family

3.4.3 Ancestral probabilities

Posterior probabilities were determined for each site of each ancestral node in the phylogeny. As expected, more ancient sites were reconstructed with less confidence than recent sites (Howard et al., 2014). For example, the mean posterior probability for all sites of the universal GRK+RSK ancestor is 0.69, for the universal GRK ancestor is 0.77, whereas it is 0.92 and 0.88 for the metazoan GRK/BARK and GRK/GRK ancestral nodes, respectively (*Appendix* Figure A.2). The posterior probabilities for the substitutions described below in Figures 3.15 and 3.16 and the main text are provided in Figures A.3 and A.4 in the *Appendix* section for reference.

3.4.4 SDR evolution N-terminal to the phosphoacceptor

Ancestral sequence reconstruction was performed for all nodes of the phylogeny, in which the probability of all 20 amino acids is calculated for each alignment position using a maximum likelihood algorithm (Ashkenazy et al., 2012). On the basis of the *Chapter 2* results, I suggest two non-conservative substitutions occurring between the ancestor of RSK and GRK kinases and the ancestor of all GRK kinases that are likely to have altered the peptide specificity of the GRK ancestor (Figure 3.15). The substitution of glutamate at 162 (R-3 and R-2 SDR) for glycine would be expected to weaken substrate binding to basic peptides on the basis of the *Chapter 2* results and on previous biochemical evidence (Moore et al., 2003). The non-conservative substitution of phenylalanine at position 86 – most likely either to histidine or to lysine – would also be expected to weaken R-3 binding on the basis of the *Chapter 2* results and one previous study (Mok et al., 2010).

In the Rhizarian lineage there is an additional substitution of glutamate at 189 for arginine, suggesting the complete loss of the R-2/R-3 preference and probably the emergence of a novel aspartate/glutamate preference at position -2 given the presence of the 86K/189R pair (c.f. *Chapter 2 Section 2.3.6* aspartate/glutamate), potentially analogous to the 127E/189E pair found in basophilic kinases. Proline at position 127 is substituted for leucine in Rhizarian kinases also. The loss of proline at 127 has previous been linked tenuously to a reduced R-2 preference, although the mechanism is unclear (Ben-Shimon and Niv, 2011).

In the heterokont lineages, the histidine/lysine at position 86 is substituted for serine. Therefore, while the heterokont kinases retain the aforementioned 127E/189E pair, the R-2 and R-3 specificities are likely to be attenuated or eliminated completely given the non-conservative substitutions at positions 86 and 162.

BARK kinases are characterised by the loss of a negative charge at 127 (E → A), and the gain of a lysine at position 189 (E → K). The ancestral sequence reconstructions suggest that the former substitution preceded the divergence of the metazoa and choanoflagellates, whereas the mutation at 189 occurred in metazoa after their divergence. The changes correspond well to the observed preference for aspartate/glutamate at -2 and -3 in empirical specificity models of GRK2. The glutamate preferences in GRK2 more N-terminal to the phosphoacceptor – -4, -5, -6, etc. – are more difficult to understand given a lack of understanding of how specificity is determined at these positions.

In the GRK *Subfamily*, a lysine residue is usually found at position 86, although ancestral sequence reconstruction suggests that this may have occurred via a glutamine intermediate, as glutamine is found in the extant *Capsaspora owczarzaki* GRK sequence. Notably, no R-2/R-3/R-5 preference is evident in the GRK5 specificity model, suggesting that the described substitutions (E162 → G162 and F86 → K86) are sufficient to eliminate this specificity.

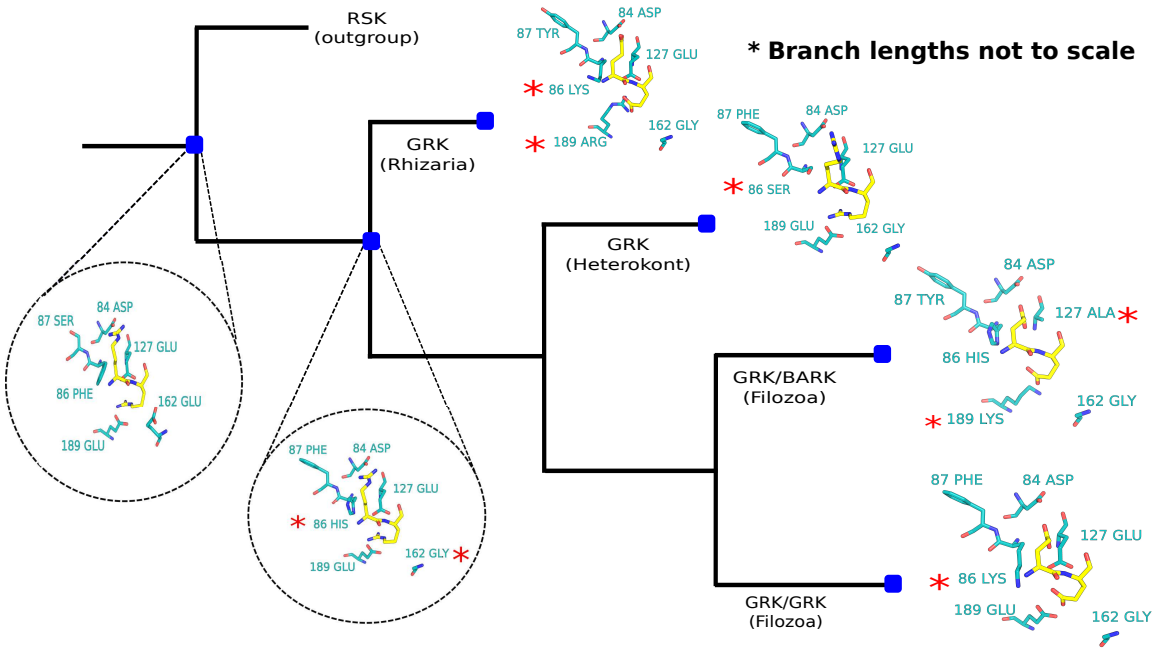


Fig. 3.15 Structures of GRK homology models at the -2 and -3 binding pockets. The identity of the ancestral kinases residues (blue) was predicted from the ancestral sequence reconstructions. Favoured residues at position -2 and -3 (yellow) have been predicted qualitatively by the author. The red asterisks represent positions where an amino acid substitution has been predicted along the branch from the ancestor to the extant sequence.

3.4.5 SDR evolution in the P+1 pocket

No changes in the P+1 pocket are observed between the ancestor of GRK+RSK and the ancestor of all GRKs according the reconstruction here. At the critical 164 position, methionine is substituted for threonine in the Rhizarian GRKs (Figure 3.16). While this is expected to affect P+1 binding, the resulting P+1 specificity is unclear as none of the 119 Ser/Thr kinase models used in the *Chapter 2* analysis feature a threonine at this position, although the *Predikin* webserver predicts a resulting A/G+1 preference (Saunders et al., 2008).

There is a substitution of tyrosine for alanine at position 157 in the branch leading to the heterokont and filozoan GRKs (Figure 3.16). The substitution to a smaller side chain may be expected to accommodate bulkier substrate residues in the P+1 pocket. I also find that 6 of the 8 heterokont sequences form a clade that features tryptophan at 164 and arginine at 161, suggesting selectivity for D/E at +1. Four of these six sequences also feature an arginine at position 157, which may strengthen selectivity for +1 D/E even further.

In the BARK kinases, proline at position 161 is substituted for histidine, which I hypothesise is – in combination with lysine at position 202 on the αG-helix – responsible for a moderate D/E preference at this position, as discussed in *Chapter 2*. The ancestral sequence reconstruction suggests that this occurred in the metazoa after their divergence from the choanoflagellates.

In the GRK *Subfamily*, the ancestral sequence reconstruction suggests a substitution of alanine for arginine at 157 in the filozoan GRK ancestor (Figure 3.16). Again, I hypothesise here that this substitution is responsible for a moderate D/E preference at +1. Indeed, a previous biochemical analysis has established a C-terminal glutamate preference for GRK1 (GRK *Subfamily*) and an N-terminal glutamate preference for GRK2 (BARK *Subfamily*), although the experimental design did not allow for the differentiation of +1/+2/+3 glutamate preferences (Onorato et al., 1991).

In the branch leading to GRK7, position 161 is substituted from proline to asparagine, polarising the P+1 pocket further (Figure 3.16). The ancestral sequence reconstruction performed here also suggests that the arginine at position 157 is replaced by phenylalanine or tyrosine in the branch leading to the vertebrate GRK1 clade, which is likely to reconstitute the hydrophobic P+1 binding pocket found in the kinase ancestral to all extant GRKs.

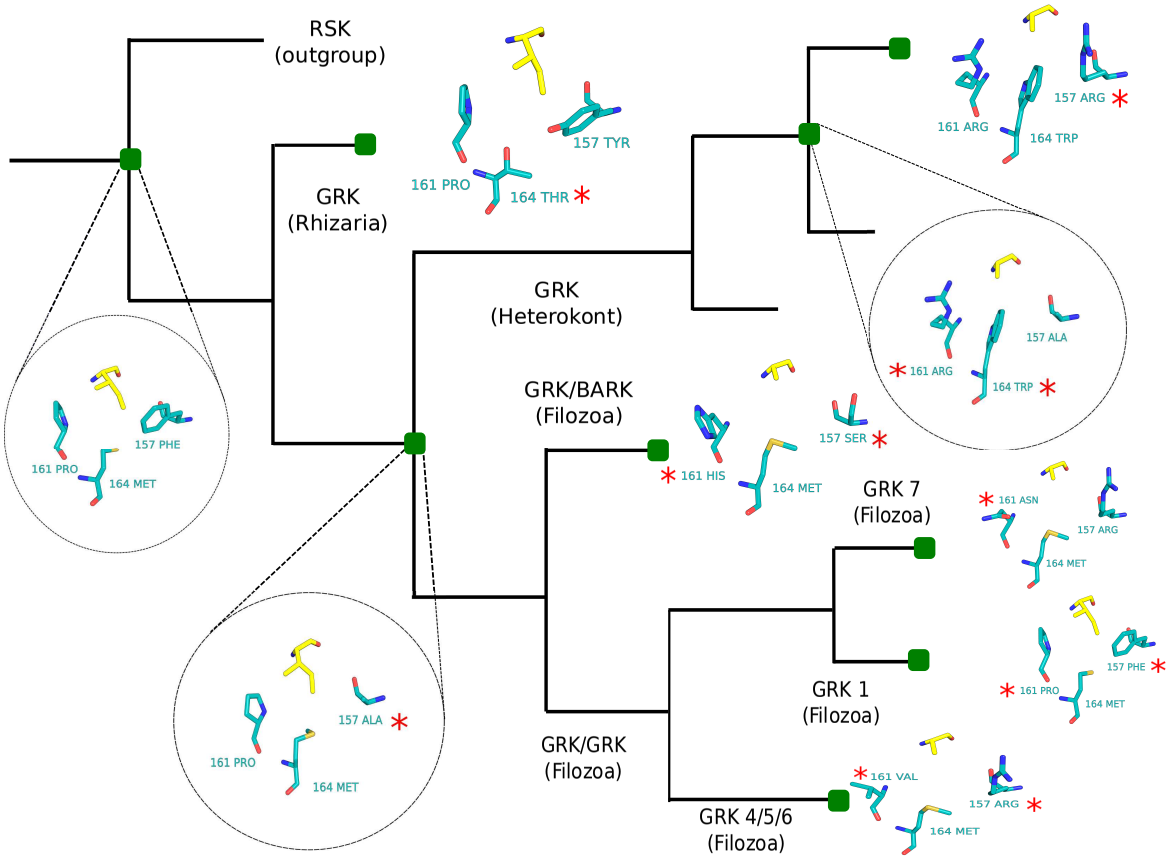


Fig. 3.16 Structures of GRK homology models at the P+1 pocket. The identity of the ancestral kinases residues (blue) was predicted from the ancestral sequence reconstructions. Favoured residues at position +1 (yellow) have been predicted qualitatively by the author. The red asterisks represent positions where an amino acid substitution has been predicted along the branch from the ancestor to the extant sequence.

3.4.6 Evolution of the αF-αG loop

The loop of amino acids linking the αF and αG helices can be observed to contact N-terminal substrate amino acids in some co-crystal structures, such as for PKA (PDB: 1ATP) and for MARK2 (PDB: 3iec). In GRK kinases, the αF-αG loop is generally longer and has a higher net positive charge than for kinases of the RSK sister clade. Specifically, I find from the GRKs used to construct the phylogeny presented here, that the median length and charge of the αF-αG loop is 14 and +3 respectively, compared to 10 and 0 in RSK kinases, and 10 and -1 in PKA kinases.

The extension of the αF-αG loop is represented in all major GRK clades (not necessarily however in all GRK sequences), suggesting that the kinase ancestral to all extant GRKs featured an extended αF-αG loop. The structural role of the flexible loop is inherently difficult to predict for sequences without associated structures. However, if considered in analogy to the aforementioned PKA and MARK2 structures, they could potentially contribute to the moderate -4/-5/-6 aspartate/glutamate preferences suggested by the empirical GRK models. The basic loop extension may alternatively/additionally contribute to the -2/-3 D/E selectivity or general substrate recruitment.

3.5 Discussion

An in-depth evolutionary analysis of protein kinase specificity was performed in this chapter. Of particular interest was the substitution of SDR residues following speciation or the duplication of kinase genes. For the former, the analysis performed here strongly confirms the hypothesis that kinase peptide specificities tend to be conserved even between distantly related orthologues. This finding is suggested both from the computational prediction of kinase specificities and from the comparison of empirical specificity matrices from human and *S. cerevisiae*. This finding aligns well with similar studies of transcription factor specificity across species (Brandt et al., 2009; Jolma et al., 2013; Nitta et al., 2015). Taken together, these studies support the conjecture that orthologues tend to be conserved in function, although this in part may reflect the roles of kinases and transcription factors as functional hubs in their respective regulatory networks (Carlson et al., 2006).

The extent of specificity conservation was quantified for the first time here by generating a null distribution of kinase matrix distances from kinases within the same *Family* and same species. This assumes that kinase within the same *Family* will tend to have similar specificities, a finding which is strongly supported by the data used for the analysis (subsection: *Divergence of kinase peptide specificity*). An obvious caveat of this analysis is that only kinases from three different species (*H. sapiens*, *M. musculus*, *S. cerevisiae*) can be compared

in this manner; this reflects a basic limitation in the current availability of phosphorylation data. It will be important to repeat this analysis in the future as data from more species becomes accessible.

Similarly, the predictive models described in *Chapter 2* strongly suggest that kinase specificities tend to be conserved between species, but were trained on data from human, mouse, and *S. cerevisiae* sequences only, which presents concerns for the reliable prediction of kinase specificities from other species. This is less likely to be a problem for animal or fungal sequences than it is for species distantly related from either human or *S. cerevisiae* (plant sequences, for example). Moreover, the limited availability of phosphorylation data for model training currently prevents rarer specificities (e.g. L+4, R+2, D/E-2) from being represented in this analysis. Both issues could be remedied by the acquisition of more data on kinase target sites or on kinase specificity more directly (e.g. from peptide libraries).

Surprisingly, the results of the analysis suggest that one-to-one kinase orthologues are no more highly conserved than one-to-many orthologues. Therefore, kinase orthologues in general seem to be highly conserved even after gene duplication. It will be important in future analyses to use phylogenetic methods to determine if this finding would still apply for one-to-many orthologues arising from ancient duplications. Moreover, it should be noted that the current analysis includes strong preferences such as P+1 or R-3 that are critical for substrate binding, but weaker substrate preferences not examined here may be more labile during evolution.

A phylogenetic approach was also adopted in this chapter to predict divergent residues across multiple *Families* and *Subfamilies* in the fungi and metazoa. These results supplement a published study performed concurrently that employed a BLAST-based approach to identify divergent residues across multiple *Families* (Kalaivani et al., 2018). The results generated from both studies are expected to be of interest to specialists for the prediction and experimental validation of drug target sites or protein-protein interactions particular to a *Family* or *Subfamily* of interest. Detailed examples for four different *Families* (CDC7, CAMK, PLK, SRPK) are given in the main text, and for PKC and PKG in (Kalaivani et al., 2018).

For the analysis performed here I adopt an evolutionary perspective by aggregating results between *Families* and grouping residues with a common predicted function (*SDR*, *catalytic*, *regulatory*, etc). These results suggests that kinase divergence at the *Family* level is dominated by changes that are likely to affect kinase specificity. At the *Subfamily* level, divergence is distributed more evenly between different functions although a significant enrichment of SDR substitutions is still observed. Overall, these results suggests that the divergence of older paralogues (i.e. *Family* level) in particular following duplication is often

driven by changes in peptide specificity, but that this is less common for newer paralogues (i.e. *Subfamily* level). Currently however, kinase substitutions occurring between the *Family* and *Subfamily* levels, in addition to substitutions before *Family* duplications and after *Subfamily* duplications, are hidden from the analysis. Sequence divergence occurring at all levels of the kinase phylogeny could be considered in the future to further resolve the relationship between specificity divergence and the time since the last common kinase ancestor.

Predictions of domain residue function could also be improved for future analyses. The *Interaction* residues in particular are predicted across kinase domains from only a limited set of structural data (Mosca et al., 2014). The acquisition of more kinase specificity data could also improve the annotation of distal SDRs. Other functional categories however are inherently more difficult to predict. Docking sites for examples have been discovered in several different kinase domain regions (Biondi and Nebreda, 2003), and are therefore not amenable to the aggregated analysis performed here. Kinase specificity divergence is therefore likely to be underestimated in general as most kinase docking sites will be assigned to the 'Other' category. For the *Regulatory* category, I assume that any substitutions in the activation loop or the alpha-C helix N-terminus are likely to affect kinase regulation. However, kinase-specific modes of regulation are known that fall outside these regions (Sang et al., 2018; Simon et al., 2016), and likely exist for uncharacterised kinases also. Therefore, as with docking regions, the extent of divergence for *Regulatory* regions may also be underestimated. Such kinase-specific examples of residue function may account for many of the switching events currently placed in the 'Other' function. Finally, for multi-domain kinase *Families* or *Subfamilies*, a certain degree of functional specialisation may be driven by changes in the protein domain composition of from sequence changes in the non-kinase domains (Pearce et al., 2010), which are not accounted for here .

To some extent the above analysis parallels the analysis of specificity models, which suggests that kinase specificity diverges at the *Family* level but not the *Subfamily* level. However, from the analysis of *Subfamily* sequences, some divergence of *Subfamily* specificity would be expected given that known SDRs are often substituted between *Subfamilies*. It should be noted that the sample size for the *Subfamily*-level comparisons between specificity models is low, and that there are examples (GRK vs. BARK, PLK1 vs. PLK2) where specificity is known to diverge between *Subfamilies* (Franchin et al., 2014; Onorato et al., 1991), suggesting some level (albeit modest) of *Subfamily* divergence. Taken together, it is therefore likely that specificity divergence can occur between *Subfamilies* but is more likely to occur at the *Family* level. However, a more complete understanding of specificity divergence awaits the full characterisation of kinome specificity.

Finally, ancestral sequence reconstructions were used for detailed evolutionary reconstruction of the GRK protein *Family*. This represents the second study of this kind, following from the 2014 evolutionary analysis of the CMGC *Group* (Howard et al., 2014). In contrast to that analysis, where a single determinant was predicted for the P+1 -> R+1 transition, I predict several different kinase domain substitutions that are likely to impact upon kinase peptide specificity. Notably, divergence in specificity was predicted between both GRK orthologues and paralogues. These include likely effects on substrate residues N-terminal and C-terminal to the phosphoacceptor. The actual effect of the substitutions observed however awaits experimental validation. This is also required to reveal whether the derived specificities emerged *de novo* or via a process of subfunctionalisation (as shown in 2014). Of note also is whether any of the predicted changes in kinase specificity will also affect the activity of the enzyme. For many protein superfamilies, changes in enzyme specificity are often accompanied by a loss of enzyme activity (the 'stability-activity' trade-off) (Miller, 2017; Tokuriki et al., 2008). Compensatory or permissive mutations that stabilise the protein are therefore often required also for the evolution of new functions (Bloom et al., 2010; McKeown et al., 2014). However, no loss of activity was found for ancestral kinases of the CMGC *Group* (Howard et al., 2014). Whether this robustness to the evolution of new specificities constitutes a universal feature of all kinase clades however awaits the experimental characterisation of the ancestral GRK enzymes reconstructed here and other ancestral kinases. If this robustness is a general feature of the kinase domain, then it may help to explain why this superfamily has been able to successfully evolve such a wide range of specificities (Bloom et al., 2006).

Chapter 4

The evolution of phosphorylation motifs

In this chapter, I investigated the evolution of phosphorylation motifs in over 50 different species. All of the analysis was performed by the author (David Bradley) under the supervision of Pedro Beltrao.

4.1 Introduction

Sequence patterns surrounding the phosphoacceptor that occur more often than would be expected by chance are referred to as 'phosphorylation motifs'. They are typically identified by comparing a large sample of phosphorylation sites against a background sample of sequences from the species proteome. Advances in phosphoproteomic technologies have therefore proved pivotal for the identification of new motifs (Ritz et al., 2009; Schwartz and Gygi, 2005; Wang et al., 2012b). Notably, many motifs identified in such a way overlap with the kinase substrate motifs revealed from more reductionist approaches (Schwartz and Gygi, 2005). The accumulation of phosphorylation data can therefore yield insights into kinase specificity even without knowledge of kinase-substrate relationships. This is important because available phosphorylation data at the time of writing far exceeds kinase-substrate relationship data in terms of both its depth and taxonomic range (Peri et al., 2003; Ritz et al., 2009). For this reason, the analysis of raw phosphorylation data across species provides an avenue into the research of kinase specificity and its evolution.

Phosphoproteomics studies in many species have now been conducted. Their results reveal that the classic acidophilic (S/T-D/E-x-D/E), basophilic (R-x-x-S/T), and proline-directed (S/T-P) motifs are present in all species so far examined and were likely universal (Al-Momani et al., 2018; Resjö et al., 2014; Tian et al., 2014; Zhai et al., 2008). However, attempts to extend this analysis to multiple species at a time have so far been limited. In 2014, Yoshizaki and Okuda collated motifs from publicly available phosphorylation data

and then calculated their conservation across 9 different opisthokont species (Yoshizaki and Okuda, 2014). As expected, the phosphomotifs identified in human were highly conserved, but this conservation decreases with evolutionary distance. In 2015, a related analysis was performed across species but with the aim of identify all over-represented substrings ('N-grams') among phosphorylation sites instead of motifs *per se* (Frades et al., 2015). The authors of this study were able to identify several N-grams that could discriminate different species from each other. However, there has been no systematic attempt as of yet to relate the detected motifs to known kinase preferences, or to quantify enrichment across several species.

More recently, phosphoproteomics efforts have been extended to cover S/T phosphorylation sites in prokaryotes (Kennelly, 2014; Lin et al., 2015b; Potel et al., 2018; Wu et al., 2016). As discussed in the *Introduction* chapter, many prokaryotes encode kinases homologous to eukaryotic protein kinases. A motif analysis of prokaryotic phosphorylation data could therefore reveal insights into the specificity of such kinases. While the number of identified ST-phosphosites from such studies is typically low (Pan et al., 2015), more recent advances have increased the number of sites detected and therefore made a motif analysis tractable (Lin et al., 2015b). Analyses of the few species studied so far however has failed to recover any of the standard phosphorylation motifs often found in eukaryotes (Lin et al., 2015b). Moreover, for the set of 11 S/T protein kinases present in *Mycobacterium Tuberculosis*, none of the experimentally characterised kinase substrate motifs correspond to a known eukaryotic motif (Prisic et al., 2010). These are not surprising findings if it is to be believed that prokaryotic S/T-kinases were not horizontally transferred but were in fact present in an early ancestor of prokaryotes and eukaryotes (Stancik et al., 2018).

In general, the extent to which a species kinome can be related to its phosphoproteome is not clear. The latter represents the summation of kinase-substrate interactions across the proteome, and so some evidence of co-evolution between the two is expected. This was first demonstrated when it was found that genomic tyrosine content in the metazoa correlates negatively with the number of predicted tyrosine kinases in the genome (Tan et al., 2009b). A follow-up study later demonstrated that the loss of tyrosine was unlikely to have been driven by selection against deleterious phosphorylation, as previously suggested (Pandya et al., 2015). However, since then there have been limited efforts to correlate other kinase-motif combinations. A 2016 study by Studer and colleagues correlating the S/T-P motif with predicted proline-directed kinases represents – to the author's knowledge – the only other investigation of this topic to date (Studer et al., 2016). Despite this, some qualitative observations have been made on this subject. For examples, the plant species *A. thaliana* is depleted for the R-R-x-S/T, which may be explained by absence of the cognate effector PKA

in plants (Frades et al., 2015). Moreover, the strong enrichment of the L-x-x-S/T motif in *Tetrahymena* correlates with the expansion of cognate Nek kinases in the genome (Tian et al., 2014). Finally, for the CK1d kinase, known substrates were found to be conserved across vertebrate orthologues for the cognate target motif (SR motif), but for a related protein that is not a substrate for CK1d kinase, the substrate motif was not conserved, implying co-evolution between the kinase and target motif (Xing et al., 2017). In the proceeding analysis, I attempt to systematically quantify the relationship between protein kinase *Families* and their cognate target motifs where this information is known.

4.2 Results

4.2.1 Motif prevalence across the Tree of Life

The primary objective of this analysis was to study the evolution of phosphorylation motifs in the eukaryotes. Towards this end, phosphorylation data was collected from 48 different eukaryotic species. This dataset covers a taxonomically broad set of organisms, including species from the alveolates (4), amoebozoa (1), excavates (3), fungi (19), heterokonts (1), metazoa (12), and plants (8). A global eukaryotic phylogeny is provided in Figure 4.1 for reference. Phosphorylation motifs were extracted automatically from each species using the *motif-x* tool (Schwartz and Gygi, 2005). Motifs found in less than a third of all species within a clade (metazoa, fungi, plants, etc) were considered as 'low-confidence' and therefore excluded from any further analysis (see *Methods* chapter Section 6.3.1).

Ten of the motifs identified had been characterised previously in the literature and assigned as substrate motifs for particular kinase *Families* or *Subfamilies* (Amanchy et al., 2007; Miller and Turk, 2018). Notably, nine of the ten motifs feature either the P+1 (proline-directed), R-2 (basophilic), or D/E+3 (acidophilic) signatures. Enrichment values for the three signatures across the 48 species is presented in Figure 4.2a. These are calculated by dividing the foreground motif fraction (phosphosite motif count normalised by total number of phosphosites) by the background motif fraction (motif count among random S/T sites in the proteome normalised by the total number of such sites). Ratios greater than 1.0 suggest motif enrichment among phosphorylation sites whereas values less than 1.0 imply motif depletion among phosphosites. Significantly, all three signatures (P+1, R-3, and D/E+3) are enriched across the majority of the species and therefore suggest that these signatures were present in the universal eukaryotic ancestor.

Enrichment values are calculated for the full motifs in Figure 4.2b. Enrichment values were calculated as described previously but are normalised also by enrichment values for

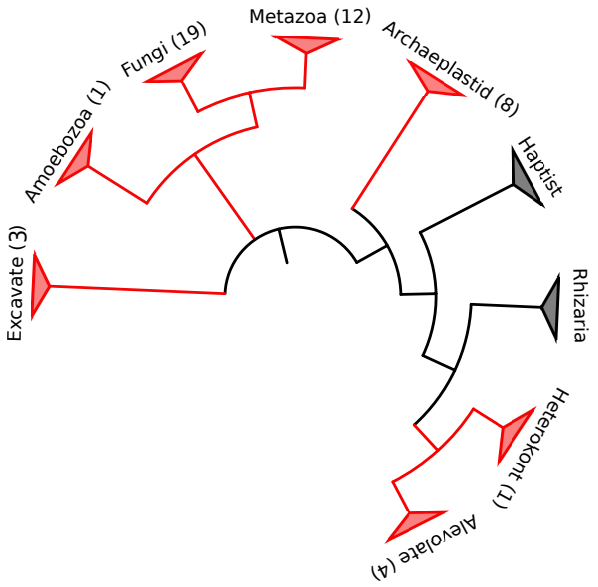


Fig. 4.1 Simplified phylogeny of the 48 species from which phosphorylation data was sampled. The number of species contributing to each clade for this analysis is given in parentheses. The clades drawn in black are not represented for this analysis

the relevant subset motif (i.e. P+1, R-3, or D/E+3). The purpose of this normalisation is to quantify motif enrichment over and above that which can be accounted for by enrichment due to the P+1, R-3, and D/E+3 signatures. Therefore, random residue additions to the aforementioned signatures (e.g. S/T-P-N, R-W-x-S/T, S/T-C-x-D/E) would not be expected to yield enrichment values greater than 1.0 following this normalisation. The results suggest that the majority of motifs are broadly enriched across the Tree of Life and were therefore probably present in the universal eukaryotic ancestor. In Figure 4.2c, the statistical significance of each enrichment is calculated by using the motif probability in the background set to calculate binomial p-values for the foreground motif prevalence (see *Methods* chapter Section 6.3.1). The results of this analysis support the above conclusion that most motifs are universal or near universal. The depletion of two basic motifs (R-R-x-S/T and R-x-R-x-x-S/T), in plants are exceptions to this rule, as has been observed previously (Resjö et al., 2014). I also observe that the L-x-R-x-x-S/T motif is highly enriched in plant species but rarely in other species. The results also reveal that the R-x-x-S/T-x-D/E motif seems to be depleted in the fungal species.

This analysis was then repeated for 'new' motifs that had not previously been assigned to an upstream effector kinase *Family* or *Subfamily*. Multiple constraints were imposed to ensure that the phosphorylation motifs are likely to represent *bona fide* kinase target motifs. For example, motifs with simple S/T additions to a classic motif were filtered from the analysis, as they likely result from the clustering of phosphorylation sites in the substrate primary

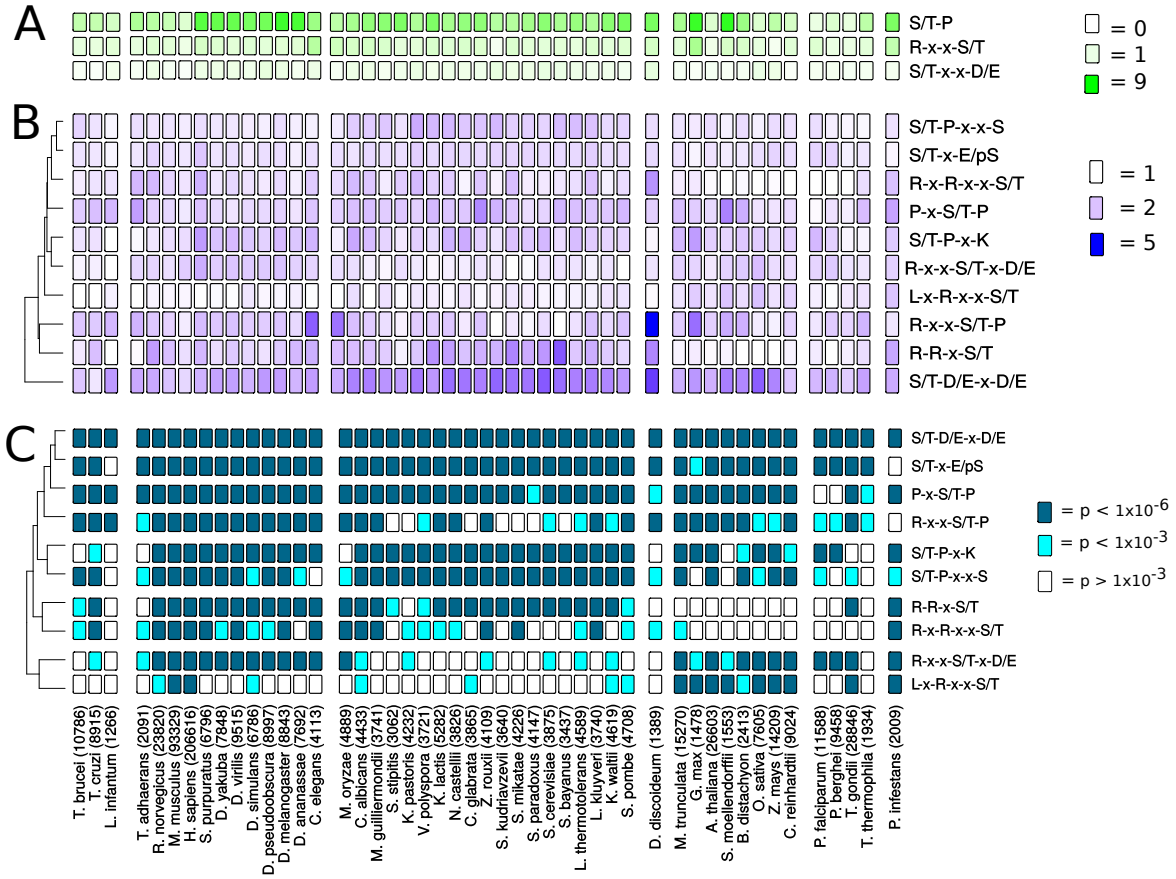


Fig. 4.2 Enrichment of phosphorylation motifs where the upstream kinase is known. Rows represent different motifs and columns represent different species. A: enrichment values for three different motif signatures. B: enrichment values relative to the proteome for 10 different phosphorylation motifs. C: p-values for the motifs, which are calculated by using the motif probability in the background set to calculate binomial p-values for the foreground motif prevalence

sequence (Moses et al., 2007; Schweiger and Linial, 2010). Moreover, I do not consider motifs that may result from the 'looping' of a known linear motif (Duarte et al., 2014). The putative linear motif 'S/T-P-x-x-x-K' for example could represent the classical 'S/T-P-x-K' motif when considered in three dimensions. Overall, 24 such motifs were identified using the strict criteria outlined in the *Methods* section (see *Methods* chapter Section 6.3.1). Notably, the list comprises motifs with determinants such as asparagine and glycine in addition to more standard motifs with R/K (basic), D/E (acidic), or proline (Figure 4.3). As above, some motifs were identified (e.g. S/T-P-x-x-x-P, S/T-P-x-P) that were likely present in the universal eukaryotic ancestor. However, for the majority of motifs identified the distribution is either intermittent or confined to a single clade or species. There are multiple asparagine-containing motifs for example that are confined either to fungal or apicocomplexan clades.

Overall, the established and 'new' motifs account for a significant proportion of the phosphoproteome across all species (Figure 4.4). Specifically, the established motifs account for 31.6% of phosphorylation sites on average (averaging across species) while the newly identified motifs account for 23.7% of motifs on average. The combined average across species is 55.3%.

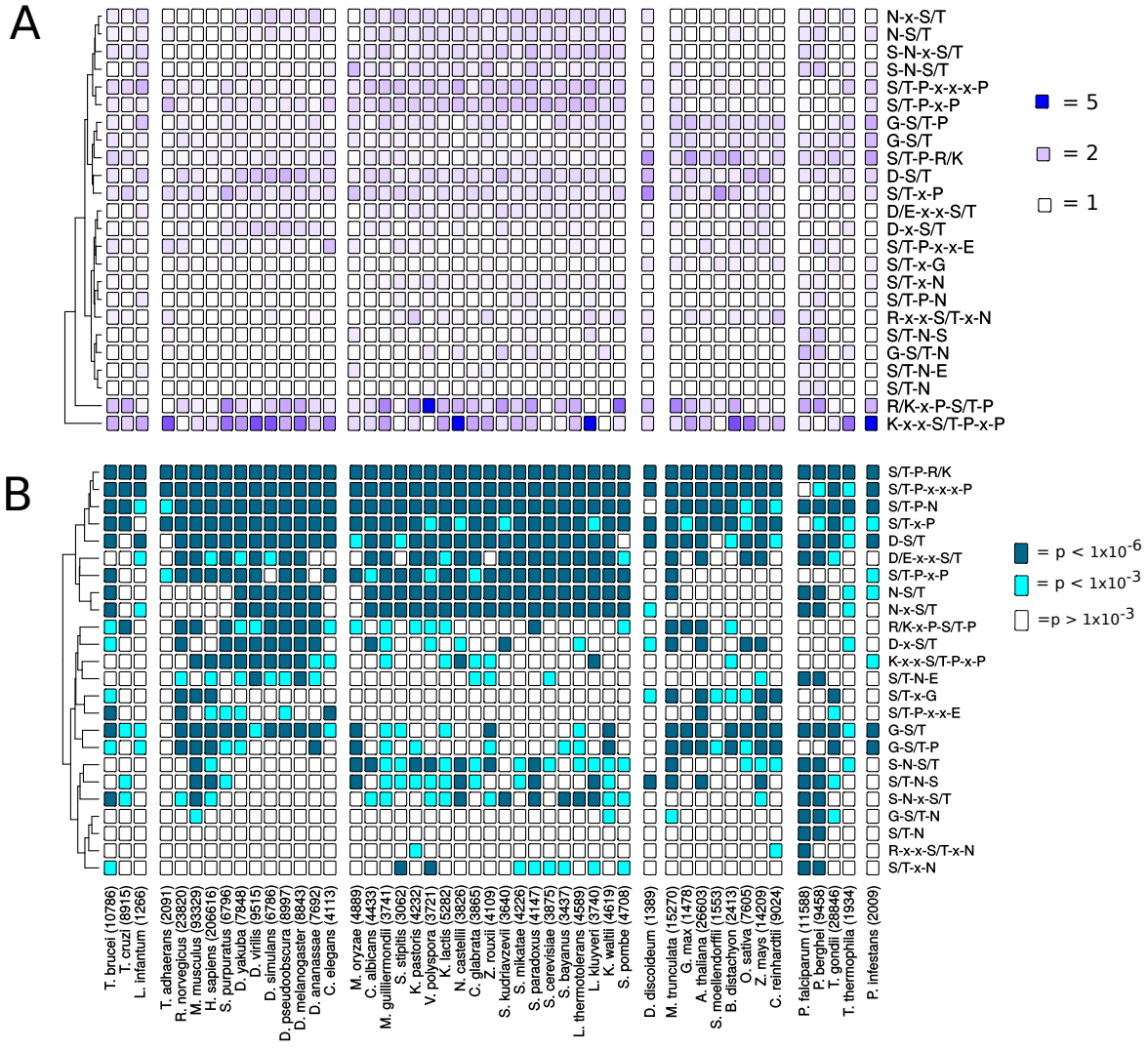


Fig. 4.3 Analysis of phosphorylation motifs for which the upstream kinase is unknown. A) Enrichment values relative to the background proteome are calculated across all 48 eukaryotic species. B) p-values for the motifs, which are calculated by using the motif probability in the background set to calculate binomial p-values for the foreground motif prevalence

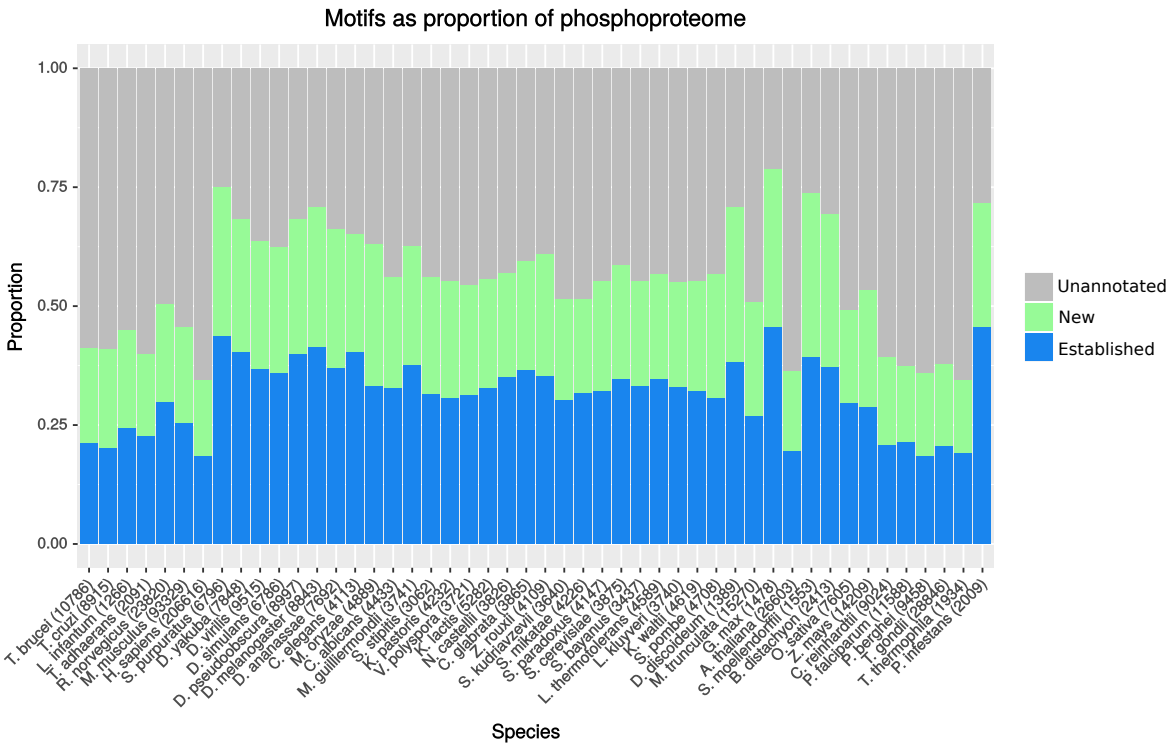


Fig. 4.4 Proportion of phosphorylation sites that match the motifs examined for this analysis. 'Established' refer to phosphosites matching motifs previously described in the literature and for which the upstream kinase is known. 'New' refers to phosphosites matching motifs for which the upstream kinase (if any) is not known. 'Unannotated' motifs are those not matching any of the motifs discussed for this analysis

4.2.2 Kinase motif enrichment in prokaryotes

As discussed in the *Introduction* section, many kinases encoded in the genomes of prokaryotes share homology to eukaryotic protein kinases. The specificity of such kinases would be expected to influence the S/T phosphoproteomes in these species. However, the S/T phosphoproteomes of archaean and bacterial species have remained poorly characterised until recently. It is only with the advent of new experimental methods that a motif-based analysis of prokaryotic S/T phosphosites has become tractable (Lin et al., 2015a).

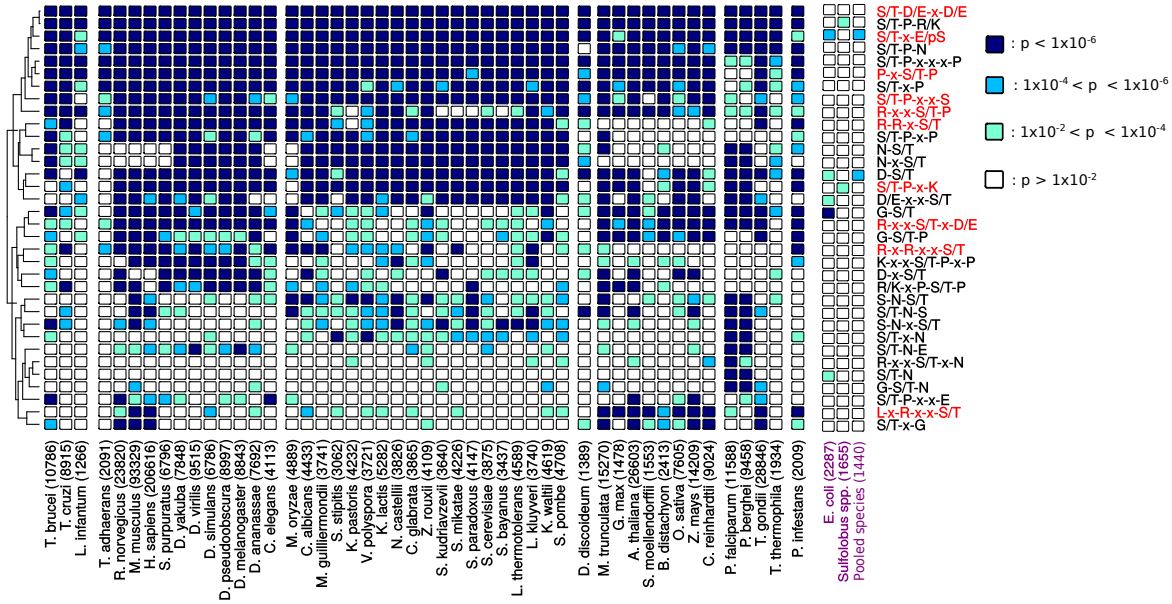


Fig. 4.5 Binomial p-values were calculated for every eukaryotic phosphorylation motif across all 48 eukaryotic species. Binomial p-values were calculated for two prokaryotic organisms also, and for a pooled set of prokaryotic phosphorylation data (right hand side). Motifs previously described in the literature and for which the upstream kinase is known are coloured in red.

In Figure 4.5, the 34 motifs described above were queried in the S/T phosphoproteome of *E. coli* (bacteria) and *Sulfolobus spp* (archaea), which are the only two organisms to the author's knowledge with more than 1,000 known S/T phosphorylation sites. The same process was applied also to a pooled sample of phosphorylation sites from 8 different prokaryotic species. While the sample sizes (E. coli: 2287, Sulfolobus spp: 1655, pooled sample: 1440) are relatively small, it is comparable to that of some eukaryotic species that show evidence of motif enrichment for motifs such as S/T-D/E-x-D/E and P-x-S/T-P (c.f. *L. infantum*, sample size 1,266). However, the analysis presented in Figure 4.5 suggests that the majority of eukaryotic phosphorylation motifs are not significantly enriched in these two species, or in the pooled sample of phosphorylation sites. I observe also that the S/T-P and R-x-x-S/T

motifs that are highly prevalent in all eukaryotic species so far examined show no evidence of strong enrichment across the 4 prokaryotic species tested (Table 4.1). The enrichment of R-3 for *Sulfolobus* is an exception to this rule, but this specificity is part of a uniform R/K preference from -5 to +5 and therefore is unlikely to be determined by a mechanism homologous to eukaryotic R-3 determination.

Species	P+1	R-3
E.coli (2287)	1.00	0.03
Synechococcus sp. (448)	0.28	0.02
M. tuberculosis (610)	1.00	0.84
Sulfolobus (1655)	1.00	<0.01

Table 4.1 Binomial p-values for the R-3 and P+1 signatures across four different prokaryotic species. The number of phosphorylation sites used for the analysis are given in brackets

Next, phosphorylation motifs were identified *de novo* in *E.coli* and *Sulfolobus* using the *motif-x* tool that was described previously. As previously shown (Lin et al., 2015a), I find that S/T phosphosites are over-represented towards the N-terminal end of proteins, hence the over-representation of methionine in known phosphosites. For both species, a fairly uniform preference for K and/or R in the S/T flanking positions is found (Table 4.2). None of the motifs identified correspond to the eukaryotic motifs presented above.

Finally, the kinomes of both species was analysed using *Kinannote* to help account for the phosphorylation patterns observed in the eukaryotes. Neither species was found to contain a protein kinase that could be placed into a canonical eukaryotic kinase *Group* (AGC, CMGC, CAMK, etc.). This is consistent with the lack of observed eukaryotic motifs, and with the argument that eukaryotic and prokaryotic kinases have been diverging vertically for billions of years and were not transferred horizontally (Stancik et al., 2018). It should be stated however that these species may still contain highly divergent kinases of the ELK (ePK-like kinase) class, which can not easily be detected using sequence-based methods but may still phosphorylate serine or threonine.

Motif (E. coli)	Score
...K...[ST].....	10.81
K.....[ST].....	8.892
.....K[ST].....	8.861
.....S[ST].....	8.761
.....M[ST].....	8.665
.....K.[ST].....	7.493
.....[ST]H.....	7.692
.....G[ST].....	6.729
.....[ST]....K..	6.169
....K..[ST].....	6.423
Motif (Sulfolobus spp.)	**Score**
....M..[ST]K.....	314.4
.....[ST].K.....	307.7
K.....[ST]R.....	314.6
.....[ST]K.....	307.7
.....[ST]R.....	307.7
.....K.[ST].....	307.7
.....[ST].R.....	307.7
....K..[ST].....	15.65
....R..[ST].....	307.7
.....[ST]..K....	12.71
.....R.[ST].....	12.43
...K...[ST].....	13.09
.....K[ST].....	10.91
.....[ST]..R....	12.64
.....M[ST].....	10.92
.....R[ST].....	9.496
...R...[ST].....	7.383
.....[ST]...K...	6.274

Table 4.2 Significant motifs identified from *E. coli* and *Sulfolobus* phosphorylation data using the *motif-x* tool (minimum of 20 sites and p-value threshold of 1×10^{-6})

4.2.3 Co-evolution between the kinome and phosphoproteome

The results presented in Figure 4.5 reveal that many phosphorylation motifs are broadly distributed across the eukaryotic Tree of Life (Figure 4.1). This is especially the case for established motifs where the upstream kinase is known from experimental analyses in human and budding yeast. These results imply that the effector kinases will also be broadly distributed across the Tree of Life. To test this hypothesis, the *Kinannote* tool was used to predict the presence of upstream kinases across a taxonomically broad set of species in the eukaryotic Tree of Life (Figure 4.1). Indeed, the results reveal that almost all of the effector kinase *Families* are universal or near-universal in the eukaryotes, with the exception of the CAMK2 *Family*, which was present in only a small number of the species tested (*Table 4.3*). In particular, it should be emphasised that these kinases seem to be absent in plants but their cognate substrate motif (R-x-x-S/T-x-D/E) is still enriched in this species, suggesting that a different kinase *Family*/*Subfamily* has convergently evolved this specificity. Overall, the results suggest that, for many motifs, both the phosphorylation motif and the upstream kinase emerged in an early ancestor of all eukaryotes.

Family	Al. (Pv)	Al. (Tt)	Am. (Dd)	Am. (Eh)	Ar. (At)	Ar. (Cr)	Ar. (Cm)	Ex. (Gi)	Ex. (Ng)	Ha. (Cc)	Ha. (Eh)	He. (Aa)	He. (Tp)	Op. (Mb)	Op. (Sc)	Rh. (Pb)	Rh. (Rf)
CK2 (S/T-D/E-x-D/E)	1	3	2	3	4	2	1	1	2	1	1	2	2	1	2	1	1
CDK (S/T-P-x-K)	6	15	10	12	47	13	5	5	13	9	12	12	9	12	7	11	11
GSK (S/T-P-x-x-S)	3	6	2	4	18	1	2	2	1	2	2	1	1	2	4	1	2
CAMK2 (R-x-x-S/T-x-D/E)	0	0	0	0	0	0	0	0	0	0	1	2	0	1	0	0	2
DYRK+RCK (R-x-x-S/T-P)	1	16	5	18	24	6	1	7	7	16	8	9	5	8	2	8	2
CMGC* (P-x-S/T-P)	13	48	19	37	118	47	12	14	35	43	34	39	18	29	19	30	46
PKA+PKG (R-R-x-S/T)	1	4	2	0	2	6	0	1	1	11	6	23	12	3	3	5	19
AKT+RSK+SGK (R-x-R-x-x-S/T)	0	18	1	6	19	1	2	1	3	1	0	8	0	7	6	2	0
CAMK (L-x-R-x-x-S/T)	16	57	21	25	120	43	5	7	16	64	78	45	38	30	22	28	83

Table 4.3 For the phosphorylation motifs with known upstream kinase, the predicted frequency of these kinases in several different species is represented. The associated substrate motif of the kinase *Family* or *Group* is given in parentheses. Each species is given by a two-letter abbreviation of a superclade to which it belongs, followed by a two-letter abrreviation of its species name. Al. (Pv): Alveolate (*Plasmodium vivax*), Al. (Tt): Alveolate (*Tetrahymena thermophila*), Am. (DD): Amoebozoa (*Dictyostelium discoideum*), Am. (Eh): Amoebozoa (*Entamoeba histolytica*), Ar. (At): Archaeplastida (*Arabidopsis thaliana*), Ar. (Cr): Archaeplastida (*Chlamydomonas reinhardtii*), Ex. (Gi): Excavate (*Giardia intestinalis*), Ex. (Ng): Excavate (*Naegleria gluberi*), Ha. (Cc): Haptophyte (*Chrysocromulina CCMP291*), Ha. (Eh): Haptophyte (*Emiliana huxleyi*), He. (Aa): Heterokont (*Aureococcus anophagefferans*), He. (Tp): Heterokont (*Thalassiosire pseudonanna*, Op. (Mb): Opisthokont (*Monosiga brevicollis*, Op. (Sc): Opisthokont (*Saccharomyces cerevisiae*, Rh. (Pb): Rhizaria (*Plasmodiophora brassicae*, Rh. (Rf): Rhizaria (*Reticulomyxa filosa*). CMGC*: CDK + MAPK + CDKL + CLK + DYRK + GSK + HIPK *Families*

.

On the other hand, the motifs described in Figure 4.3 have no known upstream effector kinase *Family* or *Subfamily*. However, it may be possible to predict upstream kinases by identifying kinase *Families* or *Subfamilies* that correlate most strongly with the motif of interest. A strong correlation would imply a putative upstream effector that could be tested experimentally.

As a positive control, kinase *Families* of known specificity could be correlated with their target motifs across the 48 eukaryotic species for which data exists. In Figure 4.6, the relationship between kinase *Family* frequencies and kinase motif enrichments across 48 species is explored. Phylogenetic independence contrasts are used for this analysis to account for the phylogenetic non-independence of data points (Felsenstein, 1985). However, for all 6 *Families* tested, no significant relationship was found between the kinase *Family* frequency and the target motif enrichments. I therefore conclude that kinase *Family* frequencies cannot reliably be used for the quantitative prediction of target motif enrichments. Potential reasons for this are given in the *Discussion* section of this chapter.

Next, the kinase *Family* frequencies and motif enrichments were mapped onto a species phylogeny to determine if there was any local evidence of co-evolution between the kinome and phosphoproteome. Such effects may be obscured when the two variables are compared globally, as performed above. Formal tests were therefore used to demonstrate that the phosphomotif enrichments and kinase *Family* frequencies possess phylogenetic signal, meaning that both variables are non-randomly distributed with respect to the species phylogeny, implying the possibility of co-evolution. P-values were below 0.01 for all kinase *Families* tested (see *Appendix*), and for the phosphorylation motifs are given in Table 4.4. Figure 4.7 reveals some examples of local co-evolution between the kinome and phosphoproteome. In the plants, for example, the lack of enrichment of the basophilic R-R-x-S/T motifs can likely be explained by the depletion of their cognate effector kinases (PKA and PKG). Also for the yeast kinases, relative expansions in the GSK *Family* correspond to stronger enrichments for the S/T-P-x-x-S motif. However, many other patterns cannot be similarly accounted for, which suggests that there are multiple factors that can affect the fold enrichment values calculated. Plots for the remaining *Families* are provided in the *Appendix* (Figure A.5 and Figure A.6).

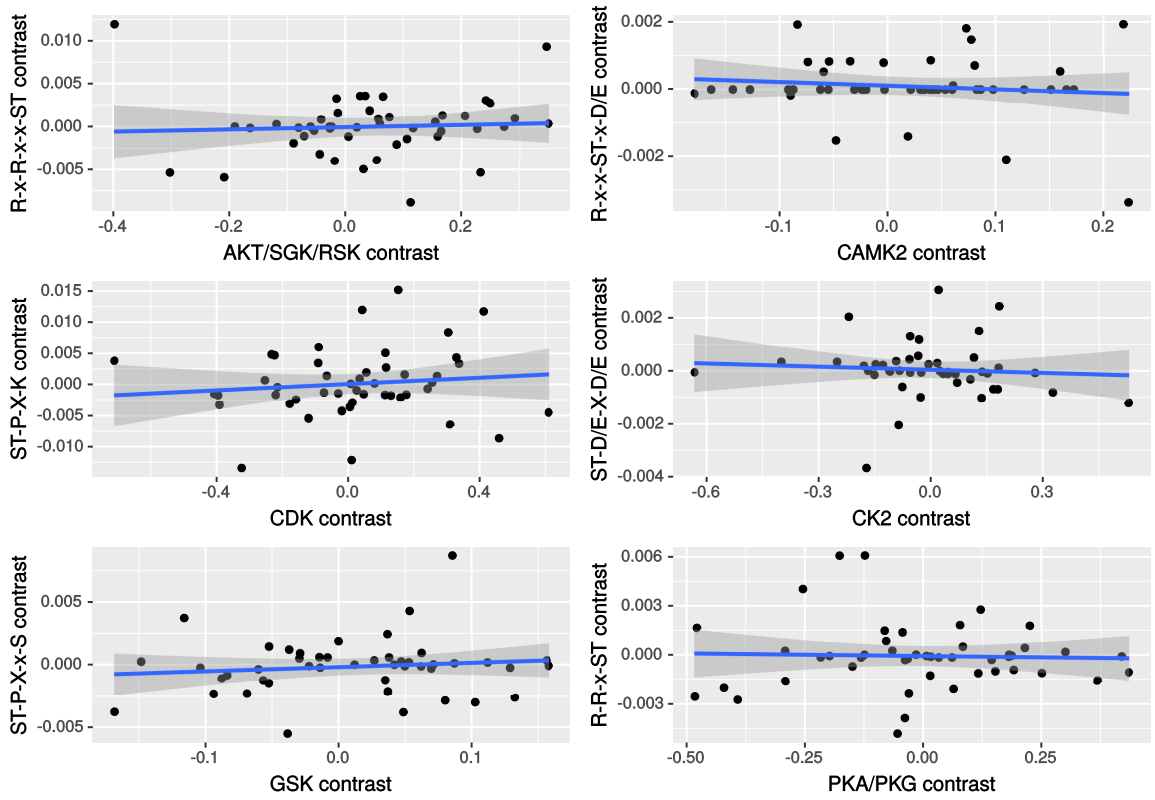


Fig. 4.6 Phylogenetic independence contrasts for kinase *Family* frequency (independent variable) and the phosphomotif enrichment value (dependent variable). No significant relationship was found for any of the 6 kinase *Families* tested.

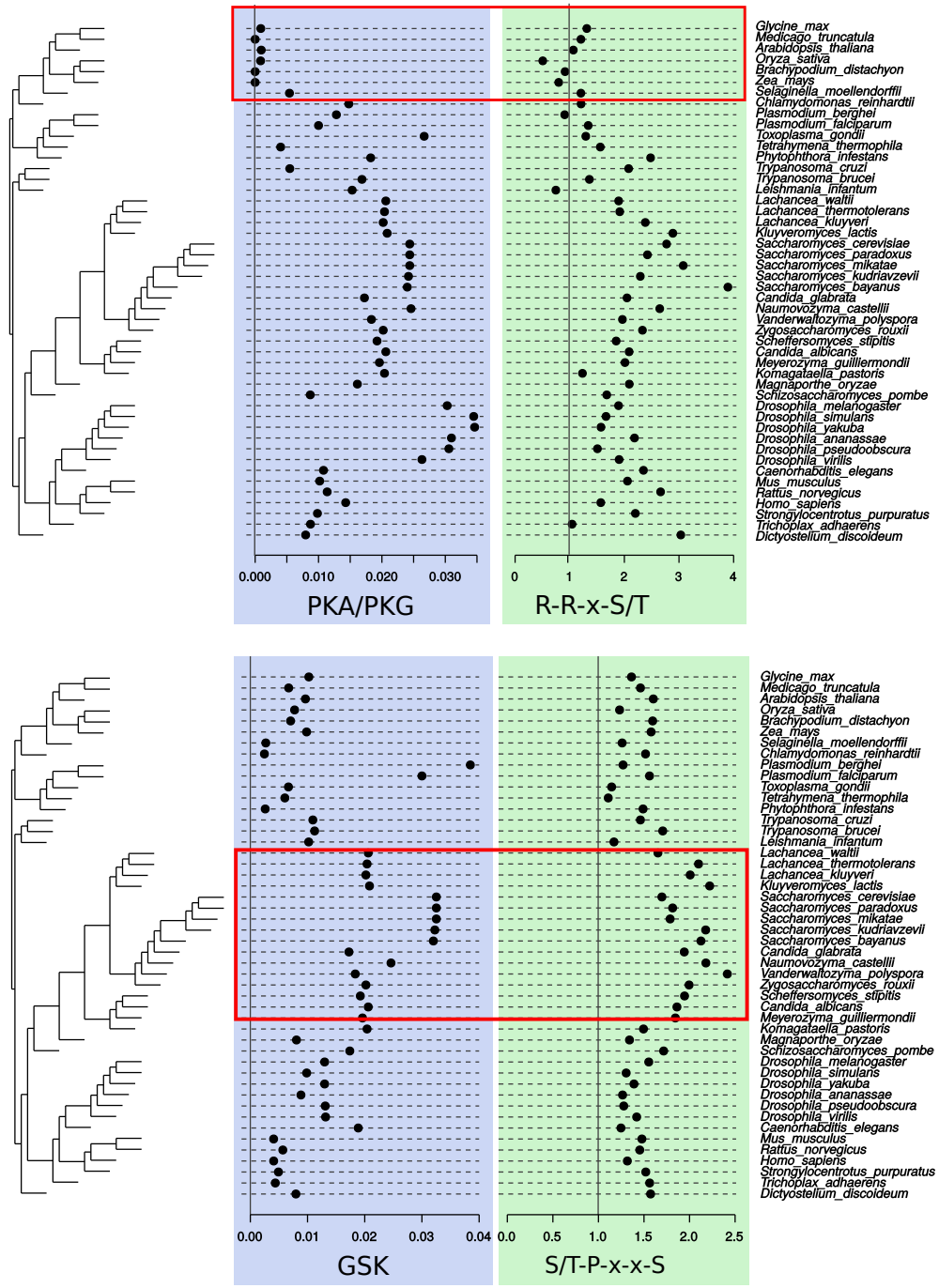


Fig. 4.7 For two kinase clades (PKA/PKG and GSK), the kinase frequency (divided by total number of kinases) and the cognate substrate motif enrichment were mapped onto a species phylogeny of the 48 eukaryotes used for this analysis. Similar plots for four other kinase clades are provided in the *Appendix*. Red boxes are used to guide the reader to relevant features of the plot. Top: Absence of PKA/PKG in plant species correlates with a lack of enrichment for the R-R-x-S/T motif. Bottom: An expansion of the GSK *Family* in yeast species correlates with a stronger enrichment for the S/T-P-x-x-S motif

Motif	Cmean	I	K	K.star	Lambda
S/T-D/E-x-D/E	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*
R-x-x-S/T-P	0.05	0.02	0.02	0.04	*0.01*
P-x-S/T-P	0.16	0.12	0.27	0.15	0.15
S/T-P-x-K	0.04	0.1	*0.01*	0.04	0.01
R-R-x-S/T	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*
R-x-x-S/T-x-D/E	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*
S/T-P-x-x-S	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*
R-x-R-x-x-S/T	*0.01*	0.01	0.04	*0.01*	*0.01*
L-x-R-x-x-S/T	*0.01*	*0.01*	0.01	*0.01*	*0.01*
S/T-x-E/pS	0.01	*0.01*	*0.01*	*0.01*	*0.01*

Table 4.4 Statistical tests for the phylogenetic signal of 10 different phosphorylation motifs. p-values less than 0.01 are represented by the number 0.01 in italics

4.3 Discussion

The analysis above represents a comprehensive investigation of phosphorylation motifs and their distribution across species. I find that many phosphorylation motifs are universal across species, although there is some evidence of clade-specific motif evolution. This suggests a 'burst' of motif generation in early eukaryotic evolution followed by a much lower rate of motif emergence. However, while informative, there are a number of caveats inherent to any comparative phosphoproteomics analysis that apply also here. The first is that the list of motifs presented above is unlikely to be exhaustive, as there may be insufficient power to detect all kinase target motifs. This is especially likely for species with poorly sampled phosphoproteomes (A. Solari et al., 2015; Riley and Coon, 2016), or kinase *Families* with few target sites. To some extent therefore the analysis above is biased towards strongly sampled clades and kinases that are particularly active or have a high number of target sites, or target sites with a high copy number. Kinases of the ATM/ATR *Family* for example target the S/T-Q motif (Traven and Heierhorst, 2005), but this motif is not represented here, perhaps because this *Subfamily* is represented by only two kinases in most species. More obviously, the Tree of Life presented in Figure 4.1 is unevenly sampled for this analysis, with a bias towards opisthokont species (animals and fungi), and with some protist superkingdoms (such as the Haptists and Rhizaria) not represented at all. Ongoing projects for the more even representation of model organisms among the protists superkingdoms may help to remedy this issue in the future (Waller et al., 2018).

Other caveats to the analysis are more subtle. The experimental workflow of high-throughput phosphoproteomics contains several steps, each of which can introduce biases for the phosphopeptides that are detected (Boekhorst et al., 2011). Slight differences in experi-

mental protocol between species may therefore produce artefactual variation. The approach used for phosphoenrichment in particular (antibodies, IMAC or TiO2) has been shown to affect phosphosite sampling (Boekhorst et al., 2011). In some cases there can be systematic experimental biases towards particular amino acids in the phosphopeptide, such as for TiO2 and acidic amino acids (Pinkse et al., 2008). Another issue is that strategies used for the growth and sampling of the organisms are likely to strongly affect the *in vivo* phosphorylation sites present at the time of sampling (Boekhorst et al., 2011). This is one of the reasons why the extent of inter-species overlap in phosphorylation sites is typically low (Boekhorst et al., 2011). However, this is less likely to be a limitation for the aggregated phosphomotif analysis here, assuming that artefacts 'average out' between motif classes.

As part of the analysis presented above, 'low confidence' motifs found in one or a few species were excluded from the analysis. This precludes the possibility of species-specific phosphomotif evolution in the analysis. However, from Figure 4.5 it is apparent that clade-specific motif evolution is quite rare. The probability of motif evolution in a single species or pair of species – thus occurring on a shorter evolutionary timescale – I therefore judge to be low, suggesting that many more false positive than false negative (if any) phosphomotifs were excluded.

More generally, the extent to which the results presented here are affected by the background set used for motif detection remains to be seen. As outlined above, an unbiased sample of sequences flanking random S/T positions in the proteome was used to generate the background set. However, given that most phosphorylation sites are found in disordered regions (Landry et al., 2009; Levy et al., 2012), a proteome sample of S/T sites biased towards disordered regions may provide a more suitable background, assuming different amino acid distributions between ordered and disordered regions (Brown et al., 2010). More recently, it has been suggested that the use of random S/T sites in the proteome may bias motif detection towards amino acids that are over-represented in the foreground set (phosphoproteome) relative to the background set (proteome) (Cheng et al., 2018). A potential solution could be to randomly shuffle foreground peptides to generate a background set that is identical in amino acid composition to the foreground set (Cheng et al., 2018).

The analysis of prokaryotic phosphorylation data performed suggests that the motifs identified originated in the universal common ancestor of eukaryotes and not in the universal common ancestor of all life. However, a phosphoproteomic analysis of more prokaryotic species will be required to strengthen this conclusion. In this case, it would be particularly useful to profile prokaryotic species with a high number of ePK kinases. Myxobacteria for example have been found to contain several species with S/T protein kinase densities rivalling those of some eukaryotic species (Pérez et al., 2008). It would be especially inter-

esting in these cases to determine how the diversity of phosphomotifs compares with that of eukaryotic species with a similar number of kinases. This could be supplemented by analysis of sequence diversity for kinase domain positions that are likely proximal to the peptide substrate. Such an analysis would be enabled by the finding that prokaryotic S/T kinases are structurally similar to canonical S/T kinases in eukaryotes (Pereira et al., 2011).

Finally, I conclude from the above analysis that kinase frequency is not generally predictive of motif enrichment for the cognate phosphorylation motif. In other words, the number of kinases encoded in a genome does not correlate positively with the number of phosphorylation sites matching the target motif, as might have been expected. There are a number of reasons why this may be the case. If kinase evolution proceeds via a process of subfunctionalisation, for example, then the target sites of the ancestral kinase would be subdivided among the duplicate kinases rather than increasing in proportion to the number of duplicate kinases. Similarly, paralogous kinases may be functionally redundant or just spatially or temporally separated from closely related kinases, in which case there is no particular reason why the pool of target sites would be significantly larger than it was for the ancestral kinase. At this point it should be emphasised that, in Figure 4.6, the independent variable is an imperfect proxy for the total number of kinases that were active at the time of phosphosite sampling. This proxy is likely to be especially poor for multicellular species, or for kinase *Families* with a broad range of activity levels between kinases. Finally, I remind the reader that specificity data currently exists only for a minority of human kinases, and therefore that kinase *Families* of unknown specificity may have peptide specificities overlapping with the kinase *Family* of interest used to produce the plots in Figure 4.6.

Some qualitative relationship was found however between the kinome and phosphoproteome, even in the absence of a clear quantitative relationship. For example, for all phosphorylation motifs found to be universal or near-universal across the eukaryotes, their cognate effector kinases were also found to be broadly distributed across the eukaryotic Tree of Life (Figure 4.1 and Table 4.3), in line with expectation. Therefore, the origin of these motifs can likely be traced to the origin of their upstream effector kinases. Also, as previously described, the data presented here shows that the depletion of basic motifs in plant species broadly correlates with the absence/depletion of the AKT and PKA kinases in these gene *Families* (Resjö et al., 2014). A simple binary prediction of the enrichment/non-enrichment of a motif of interest is still therefore possible just from an inspection of the kinome. Whether this could ever be extended to quantitative models however awaits the acquisition of more phosphorylation data and an improved annotation of kinase specificities.

Chapter 5

Summary and future directions

The objective of this thesis was to explore the evolution of protein kinase specificity. The results of all analyses performed are divided into three chapters. In *Chapter 2*, specificity determining residues (SDRs) of the eukaryotic protein kinase (ePK) domain were examined in detail. In *Chapter 3*, the evolution of kinase specificity following speciation and gene duplication was explored. In *Chapter 4*, phosphorylation data from over 50 species was collected and used to evaluate the evolution of phosphorylation motifs.

The results presented in *Chapter 2* represent a comprehensive analysis of protein kinase specificity at the active site. In it, kinase SDRs are predicted and rationalised structurally for several different kinase specificities. The results overlap with previous studies but also yield many new predictions. In two cases, the predicted SDRs have been validated experimentally. The full extent of available structural data was also leveraged to generate a binding profile for substrate positions -5 to +4. Sequence-based specificity predictors were constructed also and benchmarked for the five most common position-based preferences. Finally, the analysis of cancer tissue data demonstrated again that kinase SDRs are heavily targeted during cancer (Creixell et al., 2015b; Dixit et al., 2009), and for the first time suggested that kinase SDRs are differentially mutated between kinases of different specificity.

This work represents a significant advance in the mechanistic understanding of kinase peptide specificity. A number of obstacles however prevent a more complete understanding of specificity. The most fundamental of these is the lack of phosphorylation data both within and between species. As described, accurate specificity models can only be constructed for ~25% of human kinases and these are unequally distributed within the human kinome phylogeny. Between species, only a small portion (mammals and *S. cerevisiae*) of the Tree of Life is significantly represented. In the author's view, this currently represents the biggest bottleneck for further advances in this field. While it can be argued that sequence-based predictors can be used to address this problem, it should be stated also that none of these

methods predicts specificity completely *de novo*, which is to say that all depend to varying degrees on current kinase-substrate annotations for human, mouse, and *S. cerevisiae* for model training. It is for this reason difficult to envisage how kinase specificities could reliably be predicted for the kinases that lack close homology to previously characterised kinases. More experimental data will be required in the future to train these experimental models. An important first step will be the full kinome annotation of a model organism in a similar vein to the (Mok et al., 2010) characterisation of 61 *S. cerevisiae* kinases (50 % of kinome). This would give a sense of the range of possible kinase specificities, and also the extent to which the kinome is covered by current sequence-based predictors.

An alternative would be the complete *de novo* prediction of protein kinase specificity. Such an aim could be conceivably realised by the biophysical modelling of the kinase-substrate interface to identify high-affinity substrate sequences. A previous attempt however did not produce promising results (Kumar and Mohanty, 2010). As before, such efforts are likely to be hampered by the lack of kinase-substrate models in the PDB. This is especially likely to be the case for the 'Other' *Group* of kinases, which are poorly represented currently but have been shown previously to bind the substrate in non-standard conformations (Eswaran et al., 2008; Maiolica et al., 2014). This is a polyphyletic grouping, and so the binding conformation is likely to differ between kinases belonging to different clades. Sequence-based predictors that assume homology of binding conformation are therefore likely to perform poorly for these kinases. In view of a full kinome characterisation, it would be especially important to target all kinase specificities not currently represented in Tables 2.1, 2.2, and 2.3 for kinase-substrate crystallography analyses. Such projects are likely to significantly expand the current set of residues known to act as proximal SDRs.

Changes to kinase specificity in cancers will likely be subject to further research. In a previous study and in here, it has been found that kinase SDRs are often targeted for mutation in cancerous cells (Creixell et al., 2015b). However, the precise changes in substrate phosphorylation or signalling flux following SDR mutation have yet to be determined by phosphoproteomics experiments, nor have SDR mutations been linked to a particular stage of cancer progression. It is assumed that SDR mutations promote carcinogenesis by rewiring signalling networks, thereby contributing towards the dysregulation of cell cycle control. This implies that other changes to specificity – to docking motifs, to short linear motifs, to adaptor proteins, to scaffold proteins, and to subcellular localisations – may occur at a similar frequency to what is observed at the kinase active site, although this has not yet been demonstrated. This avenue of research will be particularly important for the tyrosine kinases, which are often mutated in cancers but tend to rely less heavily on active site interactions

to determine their specificity (Ubersax and Ferrell, 2007). More generally, the mechanism linking SDR mutation to cancer progression requires further attention.

The results from *Chapter 3* suggest that kinase peptide specificity tends to be highly conserved between orthologues. As previously discussed, this analysis could be strengthened by the experimental characterisation of a taxonomically broader set of kinases. Beyond this, it will be important to determine if specificity outside the active site – docking interactions, adaptor interactions, scaffold interactions – is also strongly conserved. This could involve a relatively simple conservation analysis of the short linear motifs that determine these interactions. Such analyses could conceivably conclude that overall kinase specificity is highly divergent even if active site specificity is conserved. It would also be useful to extend the analysis to kinase-substrate interactions rather than kinase specificity *per se*, as has been attempted previously (Tan et al., 2009a). This analysis suggested that kinase-substrate relationships tend to be conserved between species. However, for this research the specificity of non-human (*D. melanogaster*, *C. elegans*, and *S. cerevisiae*) kinases was assumed to be the same as their human orthologues. More work will be needed to generate predictors that take into account the kinase and substrate sequence simultaneously when predicting kinase-substrate interactions.

The *Chapter 3* results also suggest a simplified model of kinase evolution in which specificities diverged rapidly during early kinase evolution but to a much lower extent late in kinase evolution. This is based on the analysis of kinase clades at *Family* and *Subfamily* levels. As discussed previously, this conclusion could be strengthened by considering the kinase phylogeny at all depths. The finding also raises the question of whether the result is an exclusive feature of kinase evolution or will generalise to other enzymes and/or signalling modules. It may be the case, for example, that the emergence of eukaryotic cells necessitated a 'burst' of communication potential in early evolution, but that in the later stages of evolution there was little selective pressure for the evolution of new specificities. It would therefore be interesting to repeat the analysis for other protein domains such as the SH2 domain, which is an important signalling protein and is also characterised by a range of binding motifs (Tinti et al., 2013).

Finally from *Chapter 3*, a detailed evolutionary reconstruction was performed for the GRK *Family* of kinases responsible for the phosphorylation and regulation of G-protein coupled receptors. As discussed previously, the results of this analysis yield important insights into the evolution of specificity at the active site. However, a full interpretation for this analysis awaits an experimental characterisation of the reconstructed kinases. Perhaps the most important question is whether the predicted changes in specificity lead to any corresponding changes in protein kinase activity in the reconstructed kinase; if not, then the apparent toler-

ance of the kinase domain to the 'stability-activity tradeoff' would have implications for the evolvability of the protein kinase domain.

In *Chapter 4*, phosphorylation data was sampled from over 50 species and used to study the evolution of phosphorylation motifs. This analysis revealed that many eukaryotic phosphorylation motifs were likely present in the universal eukaryotic ancestor. While insufficient species were sampled to determine whether the motifs are universal across the eukaryotic Tree of Life – heterokont and haptist species are missing, for example – the broad distribution of species sampled strongly support the conclusion that many of the motifs emerged in a common ancestor of all eukaryotes. It is unlikely however that the list of motifs provided in *Chapter 4* is exhaustive, as weakly enriched motifs are unlikely to be detected in poorly sampled phosphoproteomes. A more comprehensive evolutionary analysis of phosphorylation motifs therefore awaits advances in phosphoproteome protocols that will increase coverage. For many of the motifs detected, the upstream kinase is unknown. *In silico* docking experiments could conceivably be used to predict upstream kinases, but would require experimental validation assays also. The analysis performed also suggests that the phosphorylation motifs observed in eukaryotes post-date their divergence from the prokaryotes. However, this conclusion is currently supported by relatively scant phosphoproteome data from the prokaryotes – a taxonomically broad sample of phosphoproteomes from the archaea and bacteria in the future will be required to strengthen this conclusion.

More generally, for the phosphorylation data used in *Chapters 2*, *3*, and *4*, each phosphosite was assigned an equal weight during the analyses. However, some phosphorylations may represent 'noisy' target sites, either in the sense that the mass spectrometry analysis can not assign the phosphoacceptor serine, threonine, or tyrosine with high confidence (Olsen et al., 2006), or that the target site represents a low-affinity substrate for the kinase (Levy et al., 2010). Measures taken to downweight spurious or low confidence phosphorylations may help to improve the quality of kinase specificity models and motif predictions from phosphoproteome data. For example, phosphosites identified from MS/MS data can be assigned to one of four categories (Class I, Class II, Class III, or Class IV) depending upon the localisation probability of the phosphoacceptor, and matches to known phosphorylation motifs (Olsen et al., 2006). Moreover, previous analyses have suggested an enrichment of non-functional phosphorylation sites at low stoichiometry on proteins of high abundance, suggesting that off-target phosphorylations could occur by chance in crowded molecular environments (Levy et al., 2012). Finally, recent studies have suggested that high-affinity kinase substrates tend to be phosphorylated more early in phosphoproteomic time courses than poor substrates (Godfrey et al., 2017; Kamenz and Ferrell, 2017; Swaffer et al., 2016). Future studies could therefore take into account these factors – substrate abundance, phos-

phorylation stoichiometry, and time-point of phosphorylation – to identify model kinase substrates among the complete set of target sites.

In conclusion, a number of questions in the field remain open, even in spite of the advances presented in *Chapters 2, 3, and 4*. It is the view of the author that further progress in the field will hinge upon the acquisition of experimental data primarily rather than the development of new computational methods. Significant advances may therefore require sizable investments of money and resources. To secure funding, it will be incumbent upon biochemists and bioinformaticians to demonstrate the importance of this field. This objective is abetted by a number of recent high-profile studies relating to protein kinase specificity and its evolution (Creixell et al., 2015a,b; Howard et al., 2014; Studer et al., 2016). In particular, the finding that phosphorylation site turnover in fungal species can act as a driver of phenotypic diversity highlights the central role of kinases in cellular evolution. The role of kinases also as hubs in signalling networks makes them prime candidates for studies within the burgeoning field of evolutionary systems biology (Albert, 2005; Zhu et al., 2007). Moreover, the ever detailed knowledge of kinase SDRs at the active site presents applications within the domain of synthetic biology, as has been demonstrated recently (Lubner et al., 2016). As this field expands, opportunities for further kinase-driven research are likely to increase also. Finally, the clinical relevance of protein kinases has now been apparent for many years, but recent research demonstrating the importance of kinase specificity rather than activity *per se* is likely to consolidate interest in kinase SDRs. It is for these reasons that research efforts directed towards kinase specificity and kinase SDRs are likely to remain robust. Advances in bioinformatic analyses – as has been presented here – would then be expected to arise secondarily from the influx of new data.

Chapter 6

Materials and Methods

The methodology underlying the results presented in chapters *2*, *3*, and *4* are described in detail here.

6.1 Methods for Chapter 2

6.1.1 Generating kinase specificity models

Phosphorylation site data were retrieved from the databases HPRD (human), Phospho.ELM (human), PhosphoGRID (*S. cerevisiae*), and PhosphoSitePlus (human and mouse). Phosphorylation sites without an annotated upstream kinase or literature reference were removed from the dataset. Phosphorylation sites in PhosphoGRID supported exclusively by the (Bodenmiller et al., 2010) or (Holt et al., 2009) studies were excluded from further analysis as these studies provide only indirect evidence for kinase-substrate interactions. Target sites that are likely to be homologous were removed with the CD-HIT program using an 85% sequence identity cut-off (Li and Godzik, 2006). Kinases of the *Atypical* class were also excluded, as they share little to no sequence homology with canonical eukaryotic protein kinases (Manning et al., 2002b).

The dataset was further filtered to remove phosphorylation sites mapping to the activation segment of kinase substrates. The justification for this is twofold. First, it has been observed that kinase autophosphorylation sites at the activation segment often conform poorly to kinase consensus motifs derived from peptide library experiments and/or trans-phosphorylation site data (Miller et al., 2008; Pike et al., 2008). Second, from the preliminary analysis I observed a small number of kinases (CAMKK1, PDK1, and LKB1/STK11) with strong substrate motifs corresponding to the CG[S/T]P motifs found in non-CMGC kinase activation segments. However, for the kinases CAMK11 and PDK1, experimental

evidence suggests that substrate specificity is determined predominantly by allosteric factors, with only a weak reported affinity between the kinase and consensus substrate peptide (Biondi et al., 2000; Okuno et al., 1997). For LKB1/STK11, while the kinase is able to efficiently phosphorylate substrate activation loop sequences in vitro (Lizcano et al., 2004), peptide library results fail to recapitulate any residues from the C-terminal CG[S/T]P motif, instead implicating leucine at the -2 position as a substrate determinant (Shaw et al., 2004). These results suggest that the strong CG[S/T]P consensus motifs observed are more likely to be artefacts of the functional constraints upon this activation segment motif rather than substrate determinants of specificity.

Specificity matrices for each kinase with at least ten phosphorylation sites were then constructed in the form of a position probability matrix (PPM). In this study, the PPMs constructed are 20 x 11 matrices with the columns representing substrate positions -5 to +5; each value in the matrix represents the relative residue frequencies (from 0 to 1) for the given substrate position. Cross-validation was used to assess kinase model performance. Briefly, a 10-fold cross-validation procedure was implemented to determine the extent to which each kinase model could successfully discriminate between true positive and true negative phosphorylation sites using a matrix-based scoring function, using the protocol described in Wagih et al., 2015. For the purpose of scoring only, the PPMs were converted into PWMs by accounting for background amino acid frequencies in the proteome. Kinase PWMs with an average AUC (area under curve) value < 0.60 were excluded from further analysis (Wagih et al., 2015).

For all kinase *Group*/*Family*/*Subfamily* classifications, the KinBase data resource was used unless otherwise specified (Manning et al., 2002b).

6.1.2 Position-based clustering of specificity models

Clustering of the PPMs was performed in a position-wise manner for the sites N- and C-terminal to the phosphoacceptor (-5, -4, -3, -2, -1; +1, +2, +3, +4, +5) using the affinity propagation (AP) algorithm (Frey and Dueck, 2007), which is a graph-based clustering method. For the application here, single column vectors (20 x 1) from each kinase PPM constitute nodes in the network, and the negative Euclidean distance between vectors represent edges upon initialisation. AP considers all nodes as potential exemplars upon initialisation, and then uses an iterative procedure to automatically identify the optimal number of clusters and cluster exemplar nodes (Frey and Dueck, 2007). AP was implemented in *R* using the *APCluster* package with default parameters for the *apcluster()* clustering function (Bodenhofer et al., 2011).

The position-based clusters generated were subject to further refinement before any further analysis. Non-specific clusters, which I define here as any cluster where the summed average probability of the two most preferred residues is <0.30, were filtered from the analysis. Clusters with fewer than 6 constituent kinases were also excluded. I also merged clusters where the dominant cluster preference was for the same amino acid or for physicochemically similar residues, as such fine-grained analysis of specificity – for example, comparisons between kinases with moderate +1 proline specificity and strong +1 proline specificity, or between arginine preference and lysine preference – are beyond the scope of this investigation. For each remaining specificity cluster I retrieved possible 'false negative' kinases by incorporating kinases in clusters for which the maximum vector weight is greater than the lower quartile of the dominant cluster preference. I suggest such false negative cluster placement to result from noisy weights for non-preferred residues and/or the presence of non-linear phosphorylation sites in the training data. Finally, potential 'false positive' cluster members were designated as those kinases where the most preferred residue(s) differs from that of the top three average preferred residues of the cluster, and were subsequently removed from the cluster.

6.1.3 Sequence alignment-based detection of putative specificity determining residues (SDRs)

Three alignment-based methods (GroupSim, Multi-Relief 3D, SPEER) were used for the detection of putative specificity-determining residues (SDRs). The use of more than a single method was motivated by the finding that ensemble approaches that incorporate predictions from three high-performing methods achieve higher specificity values than either two-method predictions or the best-performing single-method predictions when benchmarked (Chakrabarti and Panchenko, 2009). The three methods employed here represent the three algorithms with the highest single AUC values when benchmarked against a set of 20 protein family alignments with known specificity determinants (Chakraborty and Chakrabarti, 2015).

6.1.4 Procedure for sequence alignment-based inference of SDRs

I implemented an automated pipeline for the MSA-based inference of SDRs in an *R* environment. The inputs to the pipeline are the kinase specificity models and an MSA of all corresponding kinase protein sequences. The MAFFT L-INS-i method was used to generate MSAs for this analysis (Katoh et al., 2005); this was the highest-performing method in two independent benchmarks of popular alignment tools (Ahola et al., 2006; Nuin et al., 2006).

The trimAl tool was also used to remove MSA positions containing more than 20% 'gap' sites (Capella-Gutiérrez et al., 2009).

The pipeline clusters the kinase specificity models in a position-wise manner (discussed above), and then iteratively predicts SDRs for each cluster identified (e. g. +1 proline preference). This is achieved for each cluster by generating a binary partition of the MSA on the basis of cluster membership, and then using the GroupSim, Mutli-Relief 3D, and SPEER methods to predict the most likely SDRs from the MSA partition.

The GroupSim, Multi-Relief 3D, and SPEER methods use distinct schemes for position scoring; I therefore follow the precedent of the Chakrabarti and Panchenko 2009 study and identify as putative SDRs those residues among the top 15 ranked sites across all three methods. Standalone versions of GroupSim and SPEER were employed in the pipeline (Capra and Singh, 2008; Chakrabarti et al., 2007); for MultiRelief-3D, a custom R script for the method was generated on the basis of the algorithm description in (Ye et al., 2008).

This whole process described above was implemented for serine/threonine and not tyrosine kinases. I base this on the observation from structural data that SDR ordering with respect to the substrate differs at some positions; for example, the -1 residue in tyrosine kinases binds to residues that usually bind at position -3 in serine/threonine kinases (Brinkworth et al., 2003). Combining serine/threonine and tyrosine specificity models could therefore confound the analysis, and there were too few (n=16) tyrosine kinase specificity models for the reliable detection of kinase SDRs.

6.1.5 Identification of kinase-substrate cocrystal structures

Multiple steps were employed in an automated procedure to identify with confidence all cocrystal structures in the protein data bank (PDB) featuring an active site kinase-substrate/inhibitor interface. I suggest three different types of structure that are relevant to this investigation: kinase-peptide complexes at the active site, kinases in complex with long-chain substrates or inhibitors at the active site, and trans-autophosphorylation complexes.

For the detection of kinase-peptide complexes, I first used the *hmmsearch* command in HMMER (default parameters) to identify all PDB structures containing a eukaryotic protein kinase domain (PFAM: PF00069) sequence (Finn et al., 2016). All PDB files with at least one peptide chain comprising fewer than 35 amino acids were then selected. To distinguish between active site and allosteric short-chain binders, I selected all PDB files with at least one residue in contact with either the HRD catalytic aspartate of the kinase domain (P0 binding) or with the position 159 residue of the kinase activation loop (position +1 binding). I used a lenient cut-off of 6 Å to determine inter-chain contacts; the retrieved PDB files were then filtered manually to retain kinase-substrate complexes at the active site only.

The above protocol was adapted slightly to account for long-chain substrates and autophosphorylation complexes. For long-chain substrates, the procedure above was repeated to select for protein chains longer than 35 amino acids. For autophosphorylation complexes, the PDB biological assemblies were also screened for kinase-substrate contacts. All processing was performed in R with use of the Bio3D package (Skjærven et al., 2016). SIFTS XML files were also used for residue-level structure-sequence mappings (Velankar et al., 2013). The results of these searches are given in Tables 2.1, 2.2 and 2.3.

6.1.6 Structural analysis of the kinase-substrate interface

For all of the retrieved kinase-substrate structures, an automated procedure was implemented to identify the kinase substrate-binding residues for the substrate positions -5 to +4 (excluding P0). I used the PDBSum tool to predict all substrate-binding residues (de Beer et al., 2014). The substrate residue in closest proximity to the catalytic aspartate of the kinase HRD motif was identified as P0, and the flanking positions (-1,+1,-3, etc) were designated accordingly. For the binding profile shown in Fig 2.2a, Tyr kinases were excluded from the analysis. Kinase domain residues that bind infrequently to the substrate peptide (<10% of structures) were also excluded.

6.1.7 Construction of kinase-substrate models

Kinase-substrate models were constructed using existing X-ray cocrystal structures as templates. Superposition of the kinase of interest (query) with a template cocrystal structure is used to achieve a plausible positioning of the substrate peptide with reference to the query kinase. The template kinase is then removed and the template peptide mutated *in Silico* to the sequence of a known phosphorylation site of the query kinase. After resolving steric clashes between kinase and substrate, the resulting complex is then subject to energy minimisation (EM), followed by molecular dynamics (MD) equilibration and production runs.

For all models constructed, the template kinase was chosen to be similar in sequence to the query of the kinases listed in 2.1, 2.2 and 2.3. Structural superposition was performed in PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC). Steric clashes were resolved manually by rotating side chains in PyMOL. Where applicable, side-chain positioning was guided by simple physicochemical considerations and/or the residue interaction data discussed above.

All necessary input files for EM and MD were prepared using the web-based CHARMM graphical user interface (CHARMM-GUI) with default parameters (Jo et al., 2008). EM and MD runs were executed with the CHARMM36 force field using the NAMD molecu-

lar dynamics tool (Phillips et al., 2005). I imposed a harmonic restraint (force constant 90 kcal/mol/Å^2) on the catalytic aspartate of the HRD motif and on the substrate P0 residue to ensure correct positioning of the phosphoacceptor residue.

In each case, the final model used for analysis was generated by finding a representative set of co-ordinates from the protein trajectory. I used the Bio3D package to generate a Principal Components Analysis (PCA) plot of the substrate peptide trajectory co-ordinates (Skjærven et al., 2016). Partition around medoids (PAM) was then used to cluster *n* PCA component scores, where *n* is the lowest number of components that can account for 70% of the variation. I selected as the kinase-substrate model the set of peptide co-ordinates that served as the medoid to the terminal cluster i.e. the cluster of co-ordinates corresponding to the trajectory before the end of simulation.

6.1.8 Construction of predictive models and cross-validation

Naive Bayes (NB) algorithms were used to predict the specificity of protein kinases on the basis of sequence alone. Five separate classifiers were generated, corresponding to the five preferences – P+1, P-2, R-2, R-3, and L-5 – supported by at least 20 kinases.

Each classifier was trained on the 119 Ser/Thr kinase sequences of known specificity, and each kinase was labelled ('positive' or 'negative') according to the clustering of kinase specificity models described above. In each case, the prior probability of classification was set to 0.5 so that positive or negative classifications would be equally likely *a priori*. I also set a Laplace correction factor of 0.5 during training to account for the absence of particular amino acids in either positive or negative sets of the training data for a given alignment position. The R libraries *klaR* and *cvTools* were used for model generation and cross-validation, respectively (Weihs et al., 2005).

Each naive Bayes classifier was initialised with the putative specificity-determining alignment positions listed in Figure 2.4. Leave one-out cross-validation (LOOCV) was then used for each classifier to identify the subset of input SDRs that would optimise the performance of the model on the training data with respect to the AUC. For R-2, positions 127 and 189 were not implicated here as the methods used for SDR detection considers each alignment position independently of other positions. Both positions however are strongly supported as co-operative SDRs in the literature (Ben-Shimon and Niv, 2011; Zhu et al., 2005b), and are included here for specificity prediction. For R-3, separate models were trained for CMGC and non-CMGC kinases separately as this preference seems to be determined by an independent mechanism in CMGC kinases (see *Chapter 2*).

The performance of the naive Bayes models were also compared to that of the *Predikin* method (Saunders et al., 2008), which was described in the *Introduction* chapter. For this

purpose, a stand-alone version of the *Predikin* approach was implemented using a custom R script written by the author. The naive Bayes model and *Predikin* method were both trained on the same phosphorylation data to make their performance comparable. For the benchmarking of *Predikin*, each of the 119 models was predicted using a leave-one-out approach, where a single kinase at a time was exlcuded from the training set and then had its specificity predicted on the basis of the 118 other kinases of known specificity. All 119 predicted PPMs were then clustered using the procedure outlined above, and then were compared to cluster membership of the 119 empirical PPMs to generate the 'true positive'/'true negative'/'false positive'/'false negative' assignations, thus enabling the calculation of an AUC score for different predictions (P+1, P-2, R-2, R-3). Predictions could not be made for the L-5 preference as no *Predikin* SDRs exist for positions outside the -3 to +3 window.

6.1.9 Analysis of kinase mutations in cancer

Mutation data for primary tumour samples was obtained from The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/). Each kinase mutation was assigned to the correct protein isoform and then mapped to the corresponding kinase domain position.

All kinase domain positions were categorised as 'SDR', 'Catalytic', 'Regulatory', and 'Other'. Catalytic and regulatory sites were inferred on the basis of literature evidence. 'SDR' sites refers to residues that are both potential SDRs (Fig 2.4) and found within 4 Å the substrate peptide. 'Other' refers to the complement of these three sets relative to the kinase domain. The list of relevant residues is given in Table 2.5.

Most of the 'Catalytic' residues were described in the *Introduction* chapter and are found in the highly conserved motifs and loops of the protein kinase domain. Specifically, domain positions 8, 10, and 13 represent the critical glycine residues in the glycine-rich loop; positions 30 and 48 represent the highly conserved lysine and glutamate residues (respectively) in the N-terminal lobe that form an ionic bond in active kinases; residues 123, 125, and 128 are found in the 'xR**D**x**K**xx**N**' motif of the catalytic loop; residue 140 forms side chain contacts with ATP, and residue 141 represents the aspartate residue of the '**D**FG' motif that co-ordinates the Mg^{2+} ion between the ATP β- and γ- phosphates. All other positions listed – 15, 28, 85, 129, 130, 131, 186, and 190 – constitute the catalytic spine (Kornev et al., 2008).

The frequency of mutations for each functional category is given in Fig 2.20b, and a comparison of mutation recurrence per site for predicted P+1 and R-3 kinases is represented in Fig 2.20c. Per site, I used the proportion of mutations mapping to that site for a given kinase, and then took the average of this value across all kinases of the same specificity. This was preferred to the use of raw mutation frequencies, which would bias the analysis

towards highly frequent kinase-specific mutations (e.g. BRAF V600E). Tyrosine kinases were excluded from all of the cancer-based analysis, as the predicted SDRs relate to Ser/Thr kinases only.

6.1.10 SNF1 mutant *in vitro* kinase activity assay

This experimental analysis was performed by Cristina Viéitez at the EMBL research campus. The description of the methods used below was also written by Cristina Viéitez

The SNF1 plasmid from the Yeast Gal ORF collection was used as a template for directed mutagenesis to create the mutants A218L and V244R. Wild type and mutant plasmids were transformed into a BY4741 SNF1 KO strain. Cells were grown to exponential phase in SD media lacking uracil, and Snf1 expression was induced with 2% galactose for 8h. Cell pellets were collected, lysed using protease and phosphates inhibitors (Sigma) and stored O/N at -80C. Snf1 immunoprecipitation was performed using Protein A agarose beads (Sigma) with rotation for 2h at 4C. Kinase assay was performed using AQUA synthetic peptides (Sigma). Each of the 3 kinases was incubated with equal concentration of the 3 synthetic peptides (VQLKRPASVLALNDL, VQDKRPASVLALNDL and VQLKRPASVLAANDL), ATP mix (ATP 300 µM, 15 mM MgCl2, 0.5 mM EGTA 15mM β- glycerol phosphate, 0.2 mM sodium orthovanadate, 0.3 mM DTT) and allowed to react for 0, 2, 7 and 20 minutes. The reactions were quenched by transferring the reaction mixture onto dry ice at the corresponding times.

6.1.11 Mass spectrometry identification and quantification

This experimental analysis was performed by Vinothini Rajeeve at the Barts Cancer Institute, Queen Mary University of London, under the supervision of Pedro Cutillas. The description of the methods used below was also written by Vinothini Rajeeve

Kinase reaction products were diluted with 0.1% formic acid in LC-MS grade water and 5 µl of solution (containing 10 pmol of the unmodified peptide substrates) were loaded LC-MS/MS system consisting of a nanoflow ultimate 3000 RSL nano instrument coupled on-line to a Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific). Gradient elution was from 3% to 35% buffer B in 15 min at a flow rate 250 nL/min with buffer A being used to balance the mobile phase (buffer A was 0.1% formic acid in LC-MS grade water and B was 0.1% formic acid in LC-MS grade acetonitrile). The mass spectrometer was controlled by Xcalibur software (version 4.0) and operated in the positive ion mode. The spray voltage was 2 kV and the capillary temperature was set to 255 ºC. The Q-Exactive Plus was operated

in data dependent mode with one survey MS scan followed by 15 MS/MS scans. The full scans were acquired in the mass analyser at 375- 1500m/z with the resolution of 70,000 , and the MS/MS scans were obtained with a resolution of 17, 500. For quantification of each phosphopeptide and its respective unmodified form, the extracted ion chromatograms were integrated using the theoretical masses of ions using a mass tolerance of 5 ppm. Values of area-under-the-curve were obtained manually in Qual browser of Xcalibur software (version 4.0).

6.2 Methods for Chapter 3

6.2.1 Conservation of predicted specificities

Kinases with the P+1, P-2, R-2, R-3, and L-5 classes were first predicted for each kinase in human by identifying each sequence with a posterior probability greater than 0.9 for the corresponding naive Bayes model. Pan-taxonomic orthologues for each kinase were then identified using the REST API for *ensemblGenomes* Compara (Herrero et al., 2016; Kersey et al., 2018; Yates et al., 2015). Each orthologous set of sequences was aligned using the MAFFT L-INS-i algorithm (Katoh et al., 2005), and then pseudokinases were filtered from the orthologous set by identifying substitutions at the catalytic domain positions 30, 123, and 141. For an MSA of an orthologous group, the conservation of all alignment positions was assessed on the basis of the *bio3d* substitution matrix similarity (Skjærven et al., 2016). For each MSA, these values were then averaged across the groups 'Kinase domain', 'SDR', and 'Catalytic'. The 'SDR' group in Fig 3.1 represents the predicted SDRs present in Fig 2.4a, with the addition of domain positions 127 and 189 for the R-2 specificity. The 'Catalytic' group is the same as what is listed in Table 2.5. 'Kinase domain' represents the complement of the kinase domain to the other two groups.

Posterior probabilities corresponding to the human kinase specificity were also predicted for every sequence in an orthologous MSA. These values were averaged across all sequences within an MSA to quantify the extent of specificity divergence among a set of orthologues. A value of 1.0 would indicate complete conservation of specificity among orthologues and *vice versa*. Therefore, each data point in Fig 3.2 represent the average posterior probability across all sequences in the MSA of having the same specificity as the human kinase orthologue ('R-2','P+1','P-2', etc).

For Fig 3.3 however, each data point represents the mean difference in posterior probabilities between one-to-one orthologues and one-to-many orthologues for a given species (O-to-O - O-to-M). One-to-one and one-to-many predictions were again made using the

REST API for *ensemblGenomes* Compara (Herrero et al., 2016; Kersey et al., 2018; Yates et al., 2015).

6.2.2 Conservation of empirical specificity models

Human, mouse, and budding yeast (*S. cerevisiae*) specificity models were used for this analysis. All human and mouse models were generated from target site phosphorylation data as described in the *Kinase specificity models* subsection. Only 18 of 81 models however were constructed this way for the yeast kinases, with the remaining models deriving from the peptide screening data presented in (Mok et al., 2010). Before further analysis, the pT and pY sites were removed from each of the peptide screening models and then the matrices were normalised so that all columns (i.e. site positions) sum to 1.

To generate a null distribution of model distances, all possible pairwise comparisons were made for kinases existing within both the same *Family* and same species (n=218). This procedure was repeated also at the *Subfamily* level (n=110). In both cases, the Frobenius distance ('FD') was calculated for each pairwise comparison using the *norm()* function in R. This is equal to the sum of the squared element-wise distance between matrices, and then square-rooted:

$$A = matrix_1 - matrix_2$$
$$FD = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2}$$

Human/mouse orthologues of yeast kinases were identified using the Rest API for Ensembl Compara (Yates et al., 2015). From this dataset, only 14 orthologous groups were detected containing specificity from both *S. cerevisiae* and human/mouse. Each of the *S. cerevisiae* kinases was compared to all of its human/mouse orthologues, and then the maximum Frobenius distance was extracted and compared to the null distribution (Table 3.1.) This procedure was repeated in Table 3.2 for *S. cerevisiae* and mouse/human kinases belonging to either the same *Subfamily* or *Family*, as defined by the *kinase.com* resource (Manning et al., 2002b).

This analysis was also repeated using truncated matrices where 'noisy' positions in the matrix had been filtered. Unconstrained positions were identified by calculating the information content (in bits) for each matrix position, and then filtering out matrix positions at a threshold of 0.75 bits. The calculation for the number of bits (R_i) is as follows:

$$R_i = \log_2(20) - (H_i + e_n)$$

where H_i and e_n are as follows:

$$H_i = -\sum_{b=a} f_{b,i} \times \log_2 f_{b,i}$$
$$e_n = \frac{1}{\ln 2} \times \frac{(20-1)}{2n}$$

where $f_{b,i}$ is the relative frequency of amino acid b at position i, and n is equal to the number of sequences in the alignment

When comparing any two matrices, all constrained positions from *either* of the two matrices are considered together (i.e. if $matrix_1$ is constrained for 1 position, and $matrix_2$ for 2 positions, then 3 positions from each of the two matrices will be compared). The Frobenius distance calculated is then divided by total number of constrained positions to normalise the distance metric calculated.

6.2.3 Analysis of the evolution of kinase function

Kinase domain sequences were collected for all 9 opisthokont species in *KinBase* with an annotated kinome (*H. sapiens*, *M. musculus*, *S. purpuratus*, *D. melanogaster*, *C. elegans*, *A. queenslandica*, *M. brevicollis*, *S. cerevisiae*, and *C. cinerea*). The kinase domain sequences were then aligned using the MAFFT L-INS-i method (Katoh et al., 2005), and filtered to remove pseudokinases (kinases without expected residues at positions 30, 48, 123, 128, and 141). A manual correction was then made to the multiple sequences alignment (MSA), and the trimAl tool employed to remove positions with 80% or more of 'gap' characters among the sequences (Capella-Gutiérrez et al., 2009). Finally, a further filter was applied to remove truncated sequences with fewer than 190 kinase domain positions.

The resulting MSA (2094 sequences) was used to generate a maximum-likelihood kinase domain phylogeny with the RaxML tool (Stamatakis, 2014). Amino acid substitutions were modelled using the LG matrix, and a gamma model was employed to account for the heterogeneity of rates between sites. A neighbour-joining phylogeny generated with the R *ape* package was used as the starting tree (Paradis et al., 2004).

Ancestral sequence reconstructions were performed with the *CodeML* program (part of the *PAML* package) using an LG substitution matrix (Yang, 2007). No molecular clock was assumed ($clock = 0$), and a gamma model was employed again to account for rate heterogeneity between sites. The alpha parameter of the gamma distribution was estimated (fix-alpha $= 0$) with a staring value of 0.5 ($alpha = 0.5$), and four categories of the gamma distribution were specified ($ncatG = 4$). The physicochemical properties of the amino acids were not taken into account when performing the ancestral sequence reconstructions ($aaDist = 0$).

For the analysis of kinase evolution, each *Family* and *Subfamily* was assessed iteratively and a divergence score (s) was assigned to each position of the MSA. The divergence scores are calculated by comparing the *Family*/*Subfamily* of interest (clade A) with the closest sister clade (clade B) in the phylogeny. The score calculated is adapted from the BADX score of a previous publication (Edwards and Shields, 2005), specifically:

$$S = RC_A - AC.p(AC)$$

RC (recent conservation) represents the sequence conservation for the clade of interest (clade A) and is calculated here on the basis of substitution matrix similarity in the R package *bio3d* (Skjærven et al., 2016). AC represents the conservation of ancestral nodes for the clade of interest (clade A) and the ancestral node for the nearest sister clade (clade B); this is given as a 1 if the predicted residues are identical to each other and a -1 otherwise. Finally, as an innovation here, the score is weighted by the value $p(AC)$, which represent the the probability that the AC value was correctly assigned. For matching residues ($AC = 1$), this is the posterior probability of the predicted residue for clade B; for differing residues ($AC = -1$) this is the summed posterior probability of all residues in clade B besides from the predicted residue for clade A. Therefore, scores for suspected divergence would be down-weighted if there is ambiguity concerning the nature (matching or mismatching) of the clade B ancestor.

Where the sequences of interest were divided into two or more clades in the phylogeny, only the largest clade was considered for further analysis. In some cases also the clade of interest contained spurious sequences from the wrong *Group*, *Family*, or *Subfamily*. Spurious sequences were tolerated only if they comprised less than 15% of the clade sequences, otherwise the largest 'pure' subclade (with the sequences of interest only) was selected for further analysis. For the calculation of divergence scores, the nearest sister clade to the clade of interest was selected. However, scores were only calculated if the nearest sister clade contained 5 or more sequences and belonged to the correct category (e.g. two *Subfamilies* that are being compared must belong to the same *Family*). All searching/manipulation of the phylogeny was performed using a custom script in R with the aid of the *ape* package.

For the global analysis represented in Fig 3.7 and 3.9, the number of switches was calculated at the *Family* and *Subfamily* level. A substitution is considered a switch if it is above the 95th percentile for all *Subfamily* ($s_{subfamily(95)} = 1.904$) or *Family* ($s_{family(95)} = 1.793$) scores. For the one-sided Fisher tests described in the *Results* subsection of *Chapter 3*, a site is considered to be frequently switching if the number of switches is above the 90th percentile of switch frequencies for the 246 alignment positions. This was calculated separately at the *Family* (90th percentile = 8) and *Subfamily* (90th percentile = 7) level. For the analysis,

each kinase domain position was assigned to a functional category. The *Catalytic* and *Regulatory* categories are defined as they were for the analysis of cancer mutations in *Chapter 2* (*Table 2.5*). I define as a *Proximal* residue any residue within 4 Å of the substrate peptide, as determined from the *PDB: 1ATP* structure. *Distal* residues refer to putative SDRs listed in Figure 2.4a that are not within 4 Å of the substrate peptide. *Interactions* residues refers to those sites that are often found to be in contact with other protein domains in co-crystal structures, as determined using data from the 3DID database, which is a database of protein-protein interactions for which residue-level structural data exists (Mosca et al., 2014). The *Interactions* category refers to the 17 residues found to be in contact with at least 10 other protein domains from structural data. *Other* represents the complement of the kinase domain against these 5 previous sets.

6.2.4 Divergence of kinase specificity at the *Group*, *Family*, and *Subfamily* levels

For the analysis of kinase specificity, 101 high-confidence specificity models of human and mouse kinases were collected as described in the *Kinase specificity models* subsection. Each kinase was annotated at the *Group*, *Family*, and *Subfamily* level (as required) using the manual annotations given in the *kinase.com* website. The analysis of specificity divergence was performed separately at each of the three levels. For each level, all pairwise Frobenius distances within a grouping is computed and then all possible pairwise distances are calculated between groupings. Importantly, the higher-level categorisation is retained for all pairwise comparisons. For example, at the *Family*-level, all between-*Family* distance comparisons would occur for kinases belonging to the same *Group* only. For each pairwise comparison, the Frobenius distance between specificity models was calculated using the 'norm()' function in R.

6.2.5 Evolution of the GRK *Family*

Protein sequences were first retrieved from a taxonomically broad set of non-redundant proteomes (representative proteomes) (Chen et al., 2011), and then each representative proteome (rp35) was queried with a hidden Markov model (HMM) of the GRK domain (KinBase) using HMMsearch (E = 1e-75) (Eddy, 1998). The *Subfamily* classifications of each GRK were then predicted using Kinannote (Goldberg et al., 2013). Sample sequences of the RSK *Family* kinases – the *Family* most similar in sequence to the GRKs – were also included as an expected outgroup in the phylogeny, as were two kinases of the basophilic PKA *Family*.

The kinase sequences (GRK kinases plus outgroups) were then aligned using the L-INS-i algorithm of MAFFT (Katoh et al., 2005), and filtered to remove pseudokinases and redundant sequences (97% threshold), resulting in 163 sequences to be used for phylogenetic reconstruction. A maximum likelihood phylogeny was generated with RAxML using a gamma model to account for the heterogeneity of rates between sites. The optimum substitution matrix (LG) for reconstruction was also determined with RAxML using a likelihood-based approach (Stamatakis, 2014). FastML was then used for the ML-based ancestral reconstruction of sequences for all nodes in the phylogeny (Ashkenazy et al., 2012). Sequence probabilities were calculated marginally using a gamma rate model and the LG substitution matrix.

Orthology and paralogy predictions for the GRK sequences were obtained from the *Compara* resource of *ensemblGenomes* (Herrero et al., 2016; Kersey et al., 2018), and were accessed using the *ensembl* REST-API (Yates et al., 2015).

6.3 Methods for Chapter 4

6.3.1 Kinase motif enrichment across many eukaryotic species

The phosphorylation site data was collected from a range of sources. They are as follows: *Trypanosoma brucei* (Nett et al., 2009; Urbaniak et al., 2013), *Trypanosoma cruzi* (Amorim et al., 2017; Marchini et al., 2011), *Leishmania infantum* (Tsigankov et al., 2013), *Trichoplax adhaerens* (Ringrose et al., 2013), *Homo sapiens*/*Mus musculus*/*Rattus norvegicus* (Hornbeck et al., 2015), *Strongylocentrotus purpuratus* (Guo et al., 2015), *Drosophila spp.* (Hu et al., 2018), *Caenorhabditis elegans* (Rhoads et al., 2015), *Magnaporthe oryzae* (Franck et al., 2015), 18 fungal species (Studer et al., 2016), *Dictyostelium discoideum* (Charest et al., 2010), *Medicago truncatula* (Rose et al., 2012; Yao et al., 2014), *Glycine max* (Nguyen et al., 2012; Yao et al., 2014), *Arabidopsis thaliana* (Lin et al., 2015a; Yao et al., 2014), *Selaginella moellendorffii* (Chen et al., 2014b), *Brachypodium distachyon* (Lv et al., 2014), *Oryza sativa* (Hou et al., 2015; Yao et al., 2014), *Zea mays* (Marcon et al., 2015; Yao et al., 2014), *Chlamydomonas reinhardtii* (Wang et al., 2014), *Plasmodium falciparum*/*Plasmodium berghei*/*Toxoplasma gondii* (Invergo et al., 2017; Treeck et al., 2011), *Tetrahymena thermophila* (Tian et al., 2014), and *Phytophthora infestans* (Resjö et al., 2014). For each species, redundant phosphosite 15mers (centred on S or T) were filtered from the analysis.

Phosphorylation motifs for each of the 48 species were obtained by running *r-motif-x* using its default parameters (p-value of 1×10^{-06} and a minimum of 20 motif occurrences). This tool takes as its input a 'foreground' set of known target sites and a 'background' set of sites known not to be target sites (Wagih et al., 2015). In this case, sites not currently

thought to be phosphorylation sites were sampled randomly from the background proteome of each relevant species at 10x the number of observed phosphorylation sites. The frequency of motifs in the foreground relative to the background expectation is then used to calculate binomial p-values for each detected motif. The foreground and background sites were 15 residues long and with S/T present as the central residue.

For further analysis, I selected only those 'high-confidence' motifs appearing in at least a third of species within one or more superphyla (i.e. fungi, metazoa, and plants). For the poorly-sampled superphyla (alveolates and excavates), the motif had to be present in at least two of the examined species. Motifs exclusive to the amoebozoa or heterokonts were not considered as both superphyla are represented here by only a single species. Other constraints were imposed to filter out potentially spurious motifs. Serine or threonine additions to a classical motif were not considered, as they may result from the clustering of phosphorylation sites in the substrate primary sequence (Moses et al., 2007; Schweiger and Linial, 2010). I did not also consider motifs that may result from the 'looping' of a known linear motif (Duarte et al., 2014). The putative linear motif 'S/T-P-x-x-x-K' for example could represent the classical 'S/T-P-x-K' motif when considered in three dimensions. I also considered R/K and D/E to be synonymous when identifying new motifs. Finally, D/E additions to the classic casein kinase 2 motif 'S/T-D/E-x-D/E' were not considered as weak D/E preferences outside the +1 and +3 positions have already been described for this kinase (Sarno et al., 1997). Motifs detected here that do not match the list of motifs given in (Amanchy et al., 2007) or (Miller and Turk, 2018) are declared to be 'new' motifs with an unknown upstream regulator.

To calculate motif enrichment values in each species, the proportion of motif matches in the phosphorylation set (no. of matches to phosphosites / total number of phosphosites) is divided by the proportion of motif matches to the background set (no. of matches to background set / total number of sites in the background set). For motifs with more than one constrained position (i.e. $n > 1$), I normalised the enrichment value by the highest enrichment score of a subset motif with 1 fewer constrained position (n-1). This control ensures that all positions within the motif are enriched, and not just a favoured subset motif such as S/T-P or R-x-x-S/T. Binomial p values were calculated in an analogous sense as the motif null probability was taken to be equal to the total frequency of motif (e.g. P-x-S/T-P) matches to the background sites divided by the total number of matches for the subset motif (e.g. S/T-P). All binomial p values are represented in Fig 4.5.

6.3.2 Kinase motif enrichment for prokaryotic phosphorylation sites

Phosphorylation data for *E. coli* comes from the following sources: (Lin et al., 2015b; Pan et al., 2015; Potel et al., 2018). Phosphorylation data for *Sulfolobus acidocaldarius* and *Sulfolobus solfataricus* comes from the *dbPSP* database (Pan et al., 2015). Phosphorylation data for *Synechococcus sp.* and *M. tuberculosis* (Table 4.1) also derives from the dbPSP database (Pan et al., 2015). The 'pooled species' in Fig 4.5 comprises 180 unique phosphorylation sites from 8 different prokaryotic species – *Halobacterium salinarum*, *Bacilus subtilis*, *Mycobacterium tuberculosis*, *Streptomyces coelicolor*, *Escherichia coli*, *Synechococcus sp.*, *Sulfolobus solfataricus*, *Sulfolobus acidocaldarius* – all of which comes from the dbPSP database also (Pan et al., 2015).

Enrichment values and binomial p-values were calculated using the same methods described in the previous subsection. The *motif-x* tool was executed using its default parameters (p-value of 1×10^{-06} and a minimum of 20 motif occurrences), as described above.

6.3.3 Co-evolution between the kinome and phosphoproteome

A starting phylogeny for the 48 species was assembled using the NCBI taxonomy tool (Benson et al., 2009; Sayers et al., 2009). Unresolved branches (polytomies) for particular clades were then resolved manually after referring to previous phylogenetic studies in the literature (Cavalier-Smith et al., 2014; Consortium, 2007; Mathews et al., 2000; Shen et al., 2016; Telford et al., 2015). Kinome annotations for each species were generated automatically using the *KinAnnote* tool (Goldberg et al., 2013), which employs BLAST- and HMM-based searches to identify and classify ePKs.

The relationship between kinase motifs and their cognate kinases (e.g. S/T-P-x-K and CDKs) was modelled with phylogenetic independent contrasts (PIC) in R using the *ape* package (Felsenstein, 1985; Paradis et al., 2004). This method generates phylogenetic 'contrasts' between variables on a tree to account for the non-independence of data points (Felsenstein, 1985). The contrasts are first generated by finding the difference between trait values on the phylogeny for each pair of neighbouring tips (Freeman and Herron, 2003). Then, for each pair of neighbouring tips, the ancestral trait values are predicted by taking a weighted average of the two neighbour values, where the reciprocal of the branch lengths are used as weights. Each neighbouring pair of tips is then pruned from the tree, and the branch leading to their common ancestor is extended by taking into account the branch lengths from the common ancestor to its descendant tips. All possible contrasts for each pair of neighbouring tips is then calculated for the pruned tree, and the process is repeated until the phylogeny can not be pruned any further. Each contrast is then standardised by taking into account the

branch lengths of the tip pairs that were used to generate the contrast (Freeman and Herron, 2003). This process is applied for two continuous variables ('X' and 'Y'), with the objective of generating independent data points for each variable that can be used for a regression or correlation analysis.

In Fig 4.6, indepence contrasts were calculated for motif enrichment values on the x-axis and for relative kinase frequencies (number of kinases of interest divided by total number of kinases detected in the proteome) on the y-axis. The kinase frequencies for each species were calculated using the *Kinannote* tool (Goldberg et al., 2013). Kinase frequencies for species spanning the eukaryotic Tree of Life (as presented in Table 4.3) were also calculated using the *Kinannote* tool.

Tests for the phylogenetic signal of different motifs were conducted in R using the *PhyloSignal* package (Keck et al., 2016), as were tests for the phylogenetic signal of kinase family frequencies. The phylogenetic plots in Fig 4.7 were also generated using *PhyloSignal*.

References

Ahola, V., Aittokallio, T., Vihinen, M., and Uusipaikka, E. (2006). A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics*, 7:484.

Al-Momani, S., Qi, D., Ren, Z., and Jones, A. R. (2018). Comparative qualitative phosphoproteomics analysis identifies shared phosphorylation motifs and associated biological processes in evolutionary divergent plants. *Journal of Proteomics*, 181:152–159.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Alexander, J., Lim, D., Joughin, B. A., Hegemann, B., Hutchins, J. R. A., Ehrenberger, T., Ivins, F., Sessa, F., Hudecz, O., Nigg, E. A., Fry, A. M., Musacchio, A., Stukenberg, P. T., Mechtler, K., Peters, J.-M., Smerdon, S. J., and Yaffe, M. B. (2011). Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. *Science Signaling*, 4(179):ra42.

Allen, J. J., Lazerwith, S. E., and Shokat, K. M. (2005). Bio-orthogonal affinity purification of direct kinase substrates. *Journal of the American Chemical Society*, 127(15):5288–5289.

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., Huerta-Cepas, J., Linard, B., Pereira, C., Pryszcz, L. P., Schreiber, F., da Silva, A. S., Szklarczyk, D., Train, C.-M., Bork, P., Lecompte, O., von Mering, C., Xenarios, I., Sjölander, K., Jensen, L. J., Martin, M. J., Muffato, M., Quest for Orthologs consortium, Gabaldón, T., Lewis, S. E., Thomas, P. D., Sonnhammer, E., and Dessimoz, C. (2016). Standardized benchmarking in the quest for orthologs. *Nature Methods*, 13(5):425–430.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M., and Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS computational biology*, 8(5):e1002514.

Amanchy, R., Kandasamy, K., Mathivanan, S., Periaswamy, B., Reddy, R., Yoon, W.-H., Joore, J., Beer, M. A., Cope, L., and Pandey, A. (2011). Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome. *Journal of Proteomics & Bioinformatics*, 4(2):22–35.

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., and Pandey, A. (2007). A curated compendium of phosphorylation motifs. *Nature Biotechnology*, 25(3):285–286.

Amorim, J. C., Batista, M., Cunha, E. S. d., Lucena, A. C. R., Lima, C. V. d. P., Sousa, K., Krieger, M. A., and Marchini, F. K. (2017). Quantitative proteome and phosphoproteome analyses highlight the adherent population during Trypanosoma cruzi metacyclogenesis. *Scientific Reports*, 7(1):9899.

Anderson, D. P., Whitney, D. S., Hanson-Smith, V., Woznica, A., Campodonico-Burnett, W., Volkman, B. F., King, N., Thornton, J. W., and Prehoda, K. E. (2016). Evolution of an ancient protein function involved in organized multicellularity in animals. *eLife*, 5:e10147.

Archambault, V. and Glover, D. M. (2009). Polo-like kinases: conservation and divergence in their functions and regulation. *Nature Reviews Molecular Cell Biology*, 10(4):265–275.

Asai, D., Toita, R., Murata, M., Katayama, Y., Nakashima, H., and Kang, J.-H. (2014). Peptide substrates for G protein-coupled receptor kinase 2. *FEBS letters*, 588(13):2129–2132.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., and Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40(Web Server issue):W580–584.

A. Solari, F., Dell'Aica, M., Sickmann, A., and P. Zahedi, R. (2015). Why phosphoproteomics is still a challenge. *Molecular BioSystems*, 11(6):1487–1493.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291.

Babu, M. M., Teichmann, S. A., and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of molecular biology*, 358(2):614–633.

Bao, Z. Q., Jacobsen, D. M., and Young, M. A. (2011). Briefly bound to activate: transient binding of a second catalytic magnesium activates the structure and dynamics of CDK2 kinase for catalysis. *Structure (London, England: 1993)*, 19(5):675–690.

Barber, K. W., Miller, C. J., Jun, J. W., Lou, H. J., Turk, B. E., and Rinehart, J. (2018a). Kinase Substrate Profiling Using a Proteome-wide Serine-Oriented Human Peptide Library. *Biochemistry*.

Barber, K. W., Muir, P., Szeligowski, R. V., Rogulina, S., Gerstein, M., Sampson, J. R., Isaacs, F. J., and Rinehart, J. (2018b). Encoding human serine phosphopeptides in bacteria for proteome-wide identification of phosphorylation-dependent interactions. *Nature Biotechnology*, 36(7):638–644.

Batkin, M. and Shaltiel, S. (1999). The negative charge of Glu-127 in protein kinase A and its biorecognition. *FEBS letters*, 452(3):395–399.

Bax, B., Carter, P. S., Lewis, C., Guy, A. R., Bridges, A., Tanner, R., Pettman, G., Mannix, C., Culbert, A. A., Brown, M. J., Smith, D. G., and Reith, A. D. (2001). The structure of phosphorylated GSK-3beta complexed with a peptide, FRATtide, that inhibits beta-catenin phosphorylation. *Structure (London, England : 1993)*, 9(12):1143–1152.

Beenstock, J., Mooshayef, N., and Engelberg, D. (2016). How Do Protein Kinases Take a Selfie (Autophosphorylate)? *Trends in Biochemical Sciences*, 41(11):938–953.

Beltrao, P., Peer, B., J, K. N., and Vera, N. (2013). Evolution and functional cross-talk of protein post-translational modifications. *Molecular Systems Biology*, 9(1):714.

Beltrao, P., Trinidad, J. C., Fiedler, D., Roguev, A., Lim, W. A., Shokat, K. M., Burlingame, A. L., and Krogan, N. J. (2009). Evolution of Phosphoregulation: Comparison of Phosphorylation Patterns across Yeast Species. *PLOS Biology*, 7(6):e1000134.

Ben-Shimon, A. and Niv, M. Y. (2011). Deciphering the Arginine-binding preferences at the substrate-binding groove of Ser/Thr kinases by computational surface mapping. *PLoS computational biology*, 7(11):e1002288.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009). GenBank. *Nucleic Acids Research*, 37(Database issue):D26–31.

Berthon, A. S., Szarek, E., and Stratakis, C. A. (2015). PRKACA: the catalytic subunit of protein kinase A and adrenocortical tumors. *Frontiers in Cell and Developmental Biology*, 3:26.

Biondi, R. M., Cheung, P. C., Casamayor, A., Deak, M., Currie, R. A., and Alessi, D. R. (2000). Identification of a pocket in the PDK1 kinase domain that interacts with PIF and the C–terminal residues of PKA. *The EMBO Journal*, 19(5):979–988.

Biondi, R. M. and Nebreda, A. R. (2003). Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions. *The Biochemical Journal*, 372(Pt 1):1–13.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, 294(5):1351–1362.

Bloom, J. D., Gong, L. I., and Baltimore, D. (2010). Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science*, 328(5983):1272–1275.

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences*, 103(15):5869–5874.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. *Bioinformatics (Oxford, England)*, 27(17):2463–2464.

Bodenmiller, B., Wanka, S., Kraft, C., Urban, J., Campbell, D., Pedrioli, P. G., Gerrits, B., Picotti, P., Lam, H., Vitek, O., Brusniak, M.-Y., Roschitzki, B., Zhang, C., Shokat, K. M., Schlapbach, R., Colman-Lerner, A., Nolan, G. P., Nesvizhskii, A. I., Peter, M., Loewith, R., von Mering, C., and Aebersold, R. (2010). Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Science Signaling*, 3(153):rs4.

Boekhorst, J., Boersema, P. J., Tops, B. B. J., Breukelen, B. v., Heck, A. J. R., and Snel, B. (2011). Evaluating Experimental Bias and Completeness in Comparative Phosphoproteomics Analysis. *PLOS ONE*, 6(8):e23276.

Bossemeyer, D. (1995). Protein kinases–structure and function. *FEBS letters*, 369(1):57–61.

Bourque, G. and Pevzner, P. A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12(1):26–36.

Bradham, C. A., Foltz, K. R., Beane, W. S., Arnone, M. I., Rizzo, F., Coffman, J. A., Mushegian, A., Goel, M., Morales, J., Geneviere, A.-M., Lapraz, F., Robertson, A. J., Kelkar, H., Loza-Coll, M., Townley, I. K., Raisch, M., Roux, M. M., Lepage, T., Gache, C., McClay, D. R., and Manning, G. (2006). The sea urchin kinome: a first look. *Developmental Biology*, 300(1):180–193.

Brandt, T., Petrovich, M., Joerger, A. C., and Veprintsev, D. B. (2009). Conservation of DNA-binding specificity and oligomerisation properties within the p53 family. *BMC Genomics*, 10:628.

Brinkworth, R. I., Breinl, R. A., and Kobe, B. (2003). Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 100(1):74–79.

Brown, C. J., Johnson, A. K., and Daughdrill, G. W. (2010). Comparing Models of Evolution for Ordered and Disordered Proteins. *Molecular Biology and Evolution*, 27(3):609–621.

Brown, D., Krishnamurthy, N., Dale, J. M., Christopher, W., and Sjölander, K. (2005). Subfamily hmms in functional genomics. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 322–333.

Brown, J. R., Koretke, K. K., Birkeland, M. L., Sanseau, P., and Patrick, D. R. (2004). Evolutionary relationships of Aurora kinases: Implications for model organism studies and the development of anti-cancer drugs. *BMC Evolutionary Biology*, 4:39.

Brown, N. R., Noble, M. E., Endicott, J. A., and Johnson, L. N. (1999). The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nature Cell Biology*, 1(7):438–443.

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., and Manning, G. (2004). The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11707–11712.

Cao, Y., He, M., Gao, Z., Peng, Y., Li, Y., Li, L., Zhou, W., Li, X., Zhong, X., Lei, Y., Su, T., Wang, H., Jiang, Y., Yang, L., Wei, W., Yang, X., Jiang, X., Liu, L., He, J., Ye, J., Wei, Q., Li, Y., Wang, W., Wang, J., and Ning, G. (2014). Activating hotspot L205r mutation in PRKACA and adrenal Cushing's syndrome. *Science*, 344(6186):913–917.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, 25(15):1972–1973.

Capra, J. A. and Singh, M. (2008). Characterization and prediction of residues determining protein functional specificity. *Bioinformatics (Oxford, England)*, 24(13):1473–1480.

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, 7:40.

Cavalier-Smith, T., Chao, E. E., Snell, E. A., Berney, C., Fiore-Donno, A. M., and Lewis, R. (2014). Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Molecular Phylogenetics and Evolution*, 81:71–85.

Chagoyen, M., García-Martín, J. A., and Pazos, F. (2016). Practical analysis of specificity-determining residues in protein families. *Briefings in Bioinformatics*, 17(2):255–261.

Chaikuad, A., Keates, T., Vincke, C., Kaufholz, M., Zenn, M., Zimmermann, B., Gutiérrez, C., Zhang, R.-g., Hatzos-Skintges, C., Joachimiak, A., Muyldermans, S., Herberg, F. W., Knapp, S., and Müller, S. (2014). Structure of cyclin G-associated kinase (GAK) trapped in different conformations using nanobodies. *Biochemical Journal*, 459(1):59–69.

Chakrabarti, S., Bryant, S. H., and Panchenko, A. R. (2007). Functional specificity lies within the properties and evolutionary changes of amino acids. *Journal of Molecular Biology*, 373(3):801–810.

Chakrabarti, S. and Panchenko, A. R. (2009). Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC bioinformatics*, 10:207.

Chakraborty, A. and Chakrabarti, S. (2015). A survey on prediction of specificity-determining sites in proteins. *Briefings in Bioinformatics*, 16(1):71–88.

Chao, L. H., Stratton, M. M., Lee, I.-H., Rosenberg, O. S., Levitz, J., Mandell, D. J., Kortemme, T., Groves, J. T., Schulman, H., and Kuriyan, J. (2011). A mechanism for tunable autoinhibition in the structure of a human Ca2+/calmodulin- dependent kinase II holoenzyme. *Cell*, 146(5):732–745.

Charest, P. G., Shen, Z., Lakoduk, A., Sasaki, A. T., Briggs, S. P., and Firtel, R. A. (2010). A Ras signaling complex controls the RasC-TORC2 pathway and directed cell migration. *Developmental Cell*, 18(5):737–749.

Chen, C., Ha, B. H., Thévenin, A. F., Lou, H. J., Zhang, R., Yip, K. Y., Peterson, J. R., Gerstein, M., Kim, P. M., Filippakopoulos, P., Knapp, S., Boggon, T. J., and Turk, B. E. (2014a). Identification of a major determinant for serine-threonine kinase phosphoacceptor specificity. *Molecular Cell*, 53(1):140–147.

Chen, C., Natale, D. A., Finn, R. D., Huang, H., Zhang, J., Wu, C. H., and Mazumder, R. (2011). Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation. *PLOS ONE*, 6(4):e18910.

Chen, C., Nimlamool, W., Miller, C. J., Lou, H. J., and Turk, B. E. (2017). Rational Redesign of a Functional Protein Kinase-Substrate Interaction. *ACS Chemical Biology*, 12(5):1194–1198.

Chen, X., Chan, W. L., Zhu, F.-Y., and Lo, C. (2014b). Phosphoproteomic analysis of the non-seed vascular plant model Selaginella moellendorffii. *Proteome Science*, 12:16.

Cheng, A., Grant, C. E., Noble, W. S., and Bailey, T. L. (2018). MoMo: Discovery of statistically significant post-translational modification motifs. *Bioinformatics (Oxford, England)*.

Cheng, F., Jia, P., Wang, Q., and Zhao, Z. (2014). Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy. *Oncotarget*, 5(11):3697–3710.

Cheng, K. Y., Noble, M. E., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., Shen, K., Cole, P. A., Siligardi, G., and Johnson, L. N. (2006). The role of the phospho-CDK2/cyclin A recruitment site in substrate recognition. *The Journal of biological chemistry*, 281(32):23167–23179.

Cheung, J., Ginter, C., Cassidy, M., Franklin, M. C., Rudolph, M. J., Robine, N., Darnell, R. B., and Hendrickson, W. A. (2015). Structural insights into mis-regulation of protein kinase A in human tumors. *Proceedings of the National Academy of Sciences of the United States of America*, 112(5):1374–1379.

Clarke, P. R. and Hardie, D. G. (1990). Regulation of HMG-CoA reductase: identification of the site phosphorylated by the AMP-activated protein kinase in vitro and in intact rat liver. *The EMBO journal*, 9(8):2439–2446.

Cohen, P. (2002). The origins of protein phosphorylation. *Nature Cell Biology*, 4(5):E127–130.

Colinge, J., César-Razquin, A., Huber, K., Breitwieser, F. P., Májek, P., and Superti-Furga, G. (2014). Building and exploring an integrated human kinase network: Global organization and medical entry points. *Journal of Proteomics*, 107(100):113–127.

Consortium, D. . G. (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167):203–218.

Corwin, T., Woodsmith, J., Apelt, F., Fontaine, J.-F., Meierhofer, D., Helmuth, J., Grossmann, A., Andrade-Navarro, M. A., Ballif, B. A., and Stelzl, U. (2017). Defining Human Tyrosine Kinase Phosphorylation Networks Using Yeast as an In Vivo Model Substrate. *Cell Systems*, 5(2):128–139.e4.

Cozzone, A. J. (1993). ATP-dependent protein kinases in bacteria. *Journal of Cellular Biochemistry*, 51(1):7–13.

Creixell, P., Palmeri, A., Miller, C. J., Lou, H. J., Santini, C. C., Nielsen, M., Turk, B. E., and Linding, R. (2015a). Unmasking Determinants of Specificity in the Human Kinome. *Cell*, 163(1):187–201.

Creixell, P., Pandey, J. P., Palmeri, A., Santa-Olalla, M. C., Ranganathan, R., Pincus, D., and Yaffe, M. B. (2017). Hierarchical organization endows the kinase domain with regulatory plasticity. *bioRxiv*, page 197491.

Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., Perryman, L., Cox, T. R., Zivanovic, N., Palmeri, A., Wesolowska-Andersen, A., Helmer-Citterich, M., Ferkinghoff-Borg, J., Itamochi, H., Bodenmiller, B., Erler, J. T., Turk, B. E., and

Linding, R. (2015b). Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, 163(1):202–217.

Dajani, R., Fraser, E., Roe, S., Yeo, M., Good, V. M., Thompson, V., Dale, T. C., and Pearl, L. H. (2003). Structural basis for recruitment of glycogen synthase kinase 3β to the axin–APC scaffold complex. *The EMBO Journal*, 22(3):494–501.

Dale, S., Wilson, W. A., Edelman, A. M., and Hardie, D. G. (1995). Similar substrate recognition motifs for mammalian AMP-activated protein kinase, higher plant HMG-CoA reductase kinase-A, yeast SNF1, and mammalian calmodulin-dependent protein kinase I. *FEBS letters*, 361(2-3):191–195.

de Beer, T. A. P., Berka, K., Thornton, J. M., and Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Research*, 42(Database issue):D292–296.

de Oliveira, P. S. L., Ferraz, F. A. N., Pena, D. A., Pramio, D. T., Morais, F. A., and Schechtman, D. (2016). Revisiting protein kinase-substrate interactions: Toward therapeutic development. *Science Signaling*, 9(420):re3.

Dente, L., Vetriani, C., Zucconi, A., Pelicci, G., Lanfrancone, L., Pelicci, P. G., and Cesareni, G. (1997). Modified phage peptide libraries as a tool to study specificity of phosphorylation and recognition of tyrosine containing peptides. *Journal of Molecular Biology*, 269(5):694–703.

Dhanasekaran, D. N., Kashef, K., Lee, C. M., Xu, H., and Reddy, E. P. (2007). Scaffold proteins of MAP-kinase modules. *Oncogene*, 26(22):3185–3202.

Dixit, A., Yi, L., Gowthaman, R., Torkamani, A., Schork, N. J., and Verkhivker, G. M. (2009). Sequence and structure signatures of cancer mutation hotspots in protein kinases. *PloS One*, 4(10):e7485.

Douglass, J., Gunaratne, R., Bradford, D., Saeed, F., Hoffert, J. D., Steinbach, P. J., Knepper, M. A., and Pisitkun, T. (2012). Identifying protein kinase target preferences using mass spectrometry. *American Journal of Physiology. Cell Physiology*, 303(7):C715–727.

Duarte, M. L., Pena, D. A., Ferraz, F. A. N., Berti, D. A., Sobreira, T. J. P., Costa-Junior, H. M., Baqui, M. M. A., Disatnik, M.-H., Xavier-Neto, J., Oliveira, P. S. L. d., and Schechtman, D. (2014). Protein folding creates structure-based, noncontiguous consensus phosphorylation motifs recognized by kinases. *Sci. Signal.*, 7(350):ra105–ra105.

Durek, P., Schudoma, C., Weckwerth, W., Selbig, J., and Walther, D. (2009). Detection and characterization of 3d-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC bioinformatics*, 10:117.

Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3):361–365.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763.

Edwards, R. J. and Shields, D. C. (2005). BADASP: predicting functional specificity in protein families using ancestral sequences. *Bioinformatics (Oxford, England)*, 21(22):4190–4191.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Jr, R. K. S., Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C.-C., Gorovsky, M. A., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. A., Krieger, C. J., Toro, C. d., Ryder, H. F., Williamson, S. C., Barbeau, R. A., Hamilton, E. P., and Orias, E. (2006). Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote. *PLOS Biology*, 4(9):e286.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432.

Elledge, S. J. and Spottswood, M. R. (1991). A new human p34 protein kinase, CDK2, identified by complementation of a cdc28 mutation in Saccharomyces cerevisiae, is a homolog of Xenopus Eg1. *The EMBO journal*, 10(9):2653–2659.

Endicott, J. A., Noble, M. E., and Johnson, L. N. (2012). The Structural Basis for Control of Eukaryotic Protein Kinases. *Annual Review of Biochemistry*, 81(1):587–613.

Eswaran, J., Bernad, A., Ligos, J. M., Guinea, B., Debreczeni, J. E., Sobott, F., Parker, S. A., Najmanovich, R., Turk, B. E., and Knapp, S. (2008). Structure of the human protein kinase MPSK1 reveals an atypical activation loop architecture. *Structure (London, England: 1993)*, 16(1):115–124.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–285.

Fischer, E. H. and Krebs, E. G. (1955). Conversion of phosphorylase b to phosphorylase a in muscle extracts. *The Journal of Biological Chemistry*, 216(1):121–132.

Flock, T., Ravarani, C. N. J., Sun, D., Venkatakrishnan, A. J., Kayikci, M., Tate, C. G., Veprintsev, D. B., and Babu, M. M. (2015). Universal allosteric mechanism for Gα activation by GPCRs. *Nature*, 524(7564):173–179.

Frades, I., Resjö, S., and Andreasson, E. (2015). Comparison of phosphorylation patterns across eukaryotes by discriminative N-gram analysis. *BMC Bioinformatics*, 16(1).

Franchin, C., Cesaro, L., Pinna, L. A., Arrigoni, G., and Salvi, M. (2014). Identification of the PLK2-dependent phosphopeptidome by quantitative proteomics [corrected]. *PloS One*, 9(10):e111018.

Franck, W. L., Gokce, E., Randall, S. M., Oh, Y., Eyre, A., Muddiman, D. C., and Dean, R. A. (2015). Phosphoproteome Analysis Links Protein Phosphorylation to Cellular Remodeling and Metabolic Adaptation during Magnaporthe oryzae Appressorium Development. *Journal of Proteome Research*, 14(6):2408–2424.

Freeman, S. and Herron, J. C. (2003). *Evolutionary Analysis*. Pearson, Upper Saddle River, NJ, 3 edition edition.

Freschi, L., Courcelles, M., Thibault, P., Michnick, S. W., and Landry, C. R. (2011). Phosphorylation network rewiring by gene duplication. *Molecular Systems Biology*, 7:504.

Freschi, L., Osseni, M., and Landry, C. R. (2014). Functional Divergence and Evolutionary Turnover in Mammalian Phosphoproteomes. *PLOS Genetics*, 10(1):e1004062.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814):972–976.

Fukunaga, R. and Hunter, T. (1997). MNK1, a new MAP kinase-activated protein kinase, isolated by a novel expression screening method for identifying protein kinase substrates. *The EMBO journal*, 16(8):1921–1933.

Fíla, J. and Honys, D. (2012). Enrichment techniques employed in phosphoproteomics. *Amino Acids*, 43(3):1025–1047.

Gaston, D., Susko, E., and Roger, A. J. (2011). A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics (Oxford, England)*, 27(19):2655–2663.

George, S., Rochford, J. J., Wolfrum, C., Gray, S. L., Schinner, S., Wilson, J. C., Soos, M. A., Murgatroyd, P. R., Williams, R. M., Acerini, C. L., Dunger, D. B., Barford, D., Umpleby, A. M., Wareham, N. J., Davies, H. A., Schafer, A. J., Stoffel, M., O'Rahilly, S., and Barroso, I. (2004). A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science (New York, N.Y.)*, 304(5675):1325–1328.

Giannakouros, T., Nikolakaki, E., Mylonis, I., and Georgatsou, E. (2011). Serine-arginine protein kinases: a small protein kinase family with a large cellular presence. *The FEBS journal*, 278(4):570–586.

Gibbs, C. S. and Zoller, M. J. (1991). Rational scanning mutagenesis of a protein kinase identifies functional regions involved in catalysis and substrate interactions. *The Journal of Biological Chemistry*, 266(14):8923–8931.

Glover, D. M., Hagan, I. M., and Tavares, A. A. M. (1998). Polo-like kinases: a team that plays throughout mitosis. *Genes & Development*, 12(24):3777–3787.

Godfrey, M., Touati, S. A., Kataria, M., Jones, A., Snijders, A. P., and Uhlmann, F. (2017). PP2acdc55 Phosphatase Imposes Ordered Cell-Cycle Phosphorylation by Opposing Threonine Phosphorylation. *Molecular Cell*, 65(3):393–402.e3.

Goldberg, J. M., Griggs, A. D., Smith, J. L., Haas, B. J., Wortman, J. R., and Zeng, Q. (2013). Kinannote, a computer program to identify and classify members of the eukaryotic protein kinase superfamily. *Bioinformatics (Oxford, England)*, 29(19):2387–2394.

Goldberg, J. M., Manning, G., Liu, A., Fey, P., Pilcher, K. E., Xu, Y., and Smith, J. L. (2006). The dictyostelium kinome–analysis of the protein kinases from a simple model organism. *PLoS genetics*, 2(3):e38.

Goldsmith, E. J., Akella, R., Min, X., Zhou, T., and Humphreys, J. M. (2007). Substrate and Docking Interactions in Ser/Thr Protein Kinases. *Chemical reviews*, 107(11):5065–5081.

Gu, X. (2006). A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Molecular Biology and Evolution*, 23(10):1937–1945.

Guo, H., Garcia-Vedrenne, A. E., Isserlin, R., Lugowski, A., Morada, A., Sun, A., Miao, Y., Kuzmanov, U., Wan, C., Ma, H., Foltz, K., and Emili, A. (2015). Phosphoproteomic network analysis in the sea urchin Strongylocentrotus purpuratus reveals new candidates in egg activation. *Proteomics*, 15(23-24):4080–4095.

Gurevich, E. V., Tesmer, J. J. G., Mushegian, A., and Gurevich, V. V. (2012). G protein-coupled receptor kinases: more than just kinases and not only for GPCRs. *Pharmacology & Therapeutics*, 133(1):40–69.

Ha, B. H., Davis, M. J., Chen, C., Lou, H. J., Gao, J., Zhang, R., Krauthammer, M., Halaban, R., Schlessinger, J., Turk, B. E., and Boggon, T. J. (2012). Type II p21-activated kinases (PAKs) are regulated by an autoinhibitory pseudosubstrate. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16107–16112.

Hamza, A., Tammpere, E., Kofoed, M., Keong, C., Chiang, J., Giaever, G., Nislow, C., and Hieter, P. (2015). Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants. *Genetics*, 201(3):1263–1274.

Hanks, S. K. and Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 9(8):576–596.

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J. W. (2010). Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Molecular Biology and Evolution*, 27(9):1988–1999.

Hardie, D. G., Carling, D., and Carlson, M. (1998). The AMP-activated/SNF1 protein kinase subfamily: metabolic sensors of the eukaryotic cell? *Annual Review of Biochemistry*, 67:821–855.

Hemmer, W., McGlone, M., Tsigelny, I., and Taylor, S. S. (1997). Role of the Glycine Triad in the ATP-binding Site of cAMP-dependent Protein Kinase. *Journal of Biological Chemistry*, 272(27):16946–16954.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M. J., Amode, R., Brent, S., Spooner, W., Kulesha, E., Yates, A., and Flicek, P. (2016). Ensembl comparative genomics resources. *Database: The Journal of Biological Databases and Curation*, 2016.

Holt, L. J. (2012). Regulatory modules: Coupling protein stability to phopshoregulation during cell division. *FEBS Letters*, 586(17):2773–2777.

Holt, L. J., Tuch, B. B., Villén, J., Johnson, A. D., Gygi, S. P., and Morgan, D. O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science (New York, N.Y.)*, 325(5948):1682–1686.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research*, 43(Database issue):D512–520.

Hou, Y., Qiu, J., Tong, X., Wei, X., Nallamilli, B. R., Wu, W., Huang, S., and Zhang, J. (2015). A comprehensive quantitative phosphoproteome analysis of rice in response to bacterial blight. *BMC plant biology*, 15:163.

Howard, C. J., Hanson-Smith, V., Kennedy, K. J., Miller, C. J., Lou, H. J., Johnson, A. D., Turk, B. E., and Holt, L. J. (2014). Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. *eLife*, 3:e04126.

Hu, J., Rho, H.-S., Newman, R. H., Hwang, W., Neiswinger, J., Zhu, H., Zhang, J., and Qian, J. (2014). Global analysis of phosphorylation networks in humans. *Biochimica et biophysica acta*, 1844(1 0 0).

Hu, Y., Sopko, R., Chung, V., Studer, R. A., Landry, S. D., Liu, D., Rabinow, L., Gnad, F., Beltrao, P., and Perrimon, N. (2018). iProteinDB: an integrative database of Drosophila post-translational modifications. *bioRxiv*, page 386268.

Huang, C. Y., Yuan, C. J., Blumenthal, D. K., and Graves, D. J. (1995). Identification of the substrate and pseudosubstrate binding sites of phosphorylase kinase gamma-subunit. *The Journal of Biological Chemistry*, 270(13):7183–7188.

Hughes, S., Elustondo, F., Di Fonzo, A., Leroux, F. G., Wong, A. C., Snijders, A. P., Matthews, S. J., and Cherepanov, P. (2012). Crystal structure of human CDC7 kinase in complex with its activator DBF4. *Nature Structural & Molecular Biology*, 19(11):1101–1107.

Hunter, T. (2012). Why nature chose phosphate to modify proteins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1602):2513–2516.

Hunter, T. and Plowman, G. D. (1997). The protein kinases of budding yeast: six score and more. *Trends in Biochemical Sciences*, 22(1):18–22.

Hutti, J. E., Jarrell, E. T., Chang, J. D., Abbott, D. W., Storz, P., Toker, A., Cantley, L. C., and Turk, B. E. (2004). A rapid method for determining protein kinase phosphorylation specificity. *Nature Methods*, 1(1):27–29.

Imamura, H., Sugiyama, N., Wakabayashi, M., and Ishihama, Y. (2014). Large-scale identification of phosphorylation sites for profiling protein kinase selectivity. *Journal of proteome research*, 13(7):3410–3419.

Invergo, B. M., Brochet, M., Yu, L., Choudhary, J., Beltrao, P., and Billker, O. (2017). Subminute Phosphoregulation of Cell Cycle Systems during Plasmodium Gamete Formation. *Cell Reports*, 21(7):2017–2029.

Izarzugaza, J. M. G., Krallinger, M., and Valencia, A. (2012). Interpretation of the Consequences of Mutations in Protein Kinases: Combined Use of Bioinformatics and Text Mining. *Frontiers in Physiology*, 3.

Jackman, M., Firth, M., and Pines, J. (1995). Human cyclins B1 and B2 are localized to strikingly different structures: B1 to microtubules, B2 primarily to the Golgi apparatus. *The EMBO Journal*, 14(8):1646–1654.

Jaillais, Y., Hothorn, M., Belkhadir, Y., Dabi, T., Nimchuk, Z. L., Meyerowitz, E. M., and Chory, J. (2011). Tyrosine phosphorylation controls brassinosteroid receptor activation by triggering membrane release of its kinase inhibitor. *Genes & Development*, 25(3):232–237.

Jo, S., Kim, T., Iyer, V. G., and Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865.

Johnson, L. N. (2011). Substrates of Mitotic Kinases. *Sci. Signal.*, 4(179):pe31–pe31.

Johnson, L. N., Noble, M. E. M., and Owen, D. J. (1996). Active and Inactive Protein Kinases: Structural Basis for Regulation. *Cell*, 85(2):149–158.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339.

Joughin, B. A., Liu, C., Lauffenburger, D. A., Hogue, C. W. V., and Yaffe, M. B. (2012). Protein kinases display minimal interpositional dependence on substrate sequence: potential implications for the evolution of signalling networks. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1602):2574–2583.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., and Poon, A. F. Y. (2016). Ancestral Reconstruction. *PLOS Computational Biology*, 12(7):e1004763.

Kalaivani, R., Reema, R., and Srinivasan, N. (2018). Recognition of sites of functional specialisation in all known eukaryotic protein kinase families. *PLOS Computational Biology*, 14(2):e1005975.

Kamenz, J. and Ferrell, J. E. (2017). The Temporal Ordering of Cell-Cycle Phosphorylation. *Molecular Cell*, 65(3):371–373.

Kannan, N., Haste, N., Taylor, S. S., and Neuwald, A. F. (2007a). The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proceedings of the National Academy of Sciences*, 104(4):1272–1277.

Kannan, N. and Neuwald, A. F. (2004). Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. *Protein Science: A Publication of the Protein Society*, 13(8):2059–2077.

Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C., and Manning, G. (2007b). Structural and functional diversity of the microbial kinome. *PLoS biology*, 5(3):e17.

Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518.

Keck, F., Rimet, F., Bouchez, A., and Franc, A. (2016). phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecology and Evolution*, 6(9):2774–2780.

Kemp, B. E., Benjamini, E., and Krebs, E. G. (1976). Synthetic hexapeptide substrates and inhibitors of 3':5'-cyclic AMP-dependent protein kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 73(4):1038–1042.

Kennelly, P. J. (2002). Protein kinases and protein phosphatases in prokaryotes: a genomic perspective. *FEMS microbiology letters*, 206(1):1–8.

Kennelly, P. J. (2014). Protein Ser/Thr/Tyr Phosphorylation in the Archaea. *Journal of Biological Chemistry*, 289(14):9480–9487.

Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Kumar, N., Liu, Z., Maurel, T., Moore, B., McDowall, M. D., Maheswari, U., Naamati, G., Newman, V., Ong, C. K., Paulini, M., Pedro, H., Perry, E., Russell, M., Sparrow, H., Tapanari, E., Taylor, K., Vullo, A., Williams, G., Zadissia, A., Olson, A., Stein, J., Wei, S., Tello-Ruiz, M., Ware, D., Luciani, A., Potter, S., Finn, R. D., Urban, M., Hammond-Kosack, K. E., Bolser, D. M., De Silva, N., Howe, K. L., Langridge, N., Maslen, G., Staines, D. M., and Yates, A. (2018). Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Research*, 46(D1):D802–D808.

Kettenbach, A. N., Schweppe, D. K., Faherty, B. K., Pechenick, D., Pletnev, A. A., and Gerber, S. A. (2011). Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells. *Science Signaling*, 4(179):rs5.

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K. J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J. B., Morris, A., Nichols, S., Richter, D. J., Salamov, A., Sequencing, J. G. I., Bork, P., Lim, W. A., Manning, G., Miller, W. T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I. V., and Rokhsar, D. (2008). The genome of the choanoflagellate *Monosiga* brevicollis and the origin of metazoans. *Nature*, 451(7180):783–788.

Kishimoto, A., Takai, Y., Mori, T., Kikkawa, U., and Nishizuka, Y. (1980). Activation of calcium and phospholipid-dependent protein kinase by diacylglycerol, its possible relation to phosphatidylinositol turnover. *The Journal of Biological Chemistry*, 255(6):2273–2276.

Knight, Z. A. and Shokat, K. M. (2005). Features of selective kinase inhibitors. *Chemistry & Biology*, 12(6):621–637.

Knighton, D. R., Zheng, J. H., Ten Eyck, L. F., Ashford, V. A., Xuong, N. H., Taylor, S. S., and Sowadski, J. M. (1991). Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science (New York, N.Y.)*, 253(5018):407–414.

Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P., and Brinkworth, R. I. (2005). Substrate specificity of protein kinases and computational prediction of substrates. *Biochimica Et Biophysica Acta*, 1754(1-2):200–209.

Koenig, M. and Grabe, N. (2004). Highly specific prediction of phosphorylation sites in proteins. *Bioinformatics (Oxford, England)*, 20(18):3620–3627.

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, 39(1):309–338.

Kornev, A. P. and Taylor, S. S. (2015). Dynamics driven allostery in protein kinases. *Trends in biochemical sciences*, 40(11):628–647.

Kornev, A. P., Taylor, S. S., and Eyck, L. F. T. (2008). A helix scaffold for the assembly of active protein kinases. *Proceedings of the National Academy of Sciences*, 105(38):14377–14382.

Kotrba, P., Inui, M., and Yukawa, H. (2001). Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism. *Journal of Bioscience and Bioengineering*, 92(6):502–517.

Krebs, E. G. (1983). Historical perspectives on protein phosphorylation and a classification system for protein kinases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 302(1108):3–11.

Krebs, E. G. and Fischer, E. H. (1956). The phosphorylase b to a converting enzyme of rabbit skeletal muscle. *Biochimica Et Biophysica Acta*, 20(1):150–157.

Kuenzel, E. A., Mulligan, J. A., Sommercorn, J., and Krebs, E. G. (1987). Substrate specificity determinants for casein kinase II as deduced from studies with synthetic peptides. *The Journal of Biological Chemistry*, 262(19):9136–9140.

Kumar, N. and Mohanty, D. (2010). Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials. *Bioinformatics*, 26(2):189–197.

Lahiry, P., Torkamani, A., Schork, N. J., and Hegele, R. A. (2010). Kinase mutations in human disease: interpreting genotype–phenotype relationships. *Nature Reviews Genetics*, 11(1):60–74.

Landry, C. R., Freschi, L., Zarin, T., and Moses, A. M. (2014). Turnover of protein phosphorylation evolving under stabilizing selection. *Frontiers in Genetics*, 5:245.

Landry, C. R., Levy, E. D., and Michnick, S. W. (2009). Weak functional constraints on phosphoproteomes. *Trends in genetics: TIG*, 25(5):193–197.

Laporte, D. C., Stueland, C. S., and Ikeda, T. P. (1989). Isocitrate dehydrogenase kinase/phosphatase. *Biochimie*, 71(9-10):1051–1057.

LeBoeuf, B., Gruninger, T. R., and Garcia, L. R. (2007). Food Deprivation Attenuates Seizures through CaMKII and EAG K+ Channels. *PLOS Genetics*, 3(9):e156.

Lee, M. G. and Nurse, P. (1987). Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. *Nature*, 327(6117):31–35.

Lehti-Shiu, M. D. and Shiu, S.-H. (2012). Diversity, classification and function of the plant protein kinase superfamily. *Phil. Trans. R. Soc. B*, 367(1602):2619–2639.

Levene, P. A. and Alsberg, C. L. (1906). The Cleavage Products of Vitellin. *Journal of Biological Chemistry*, 2(1):127–133.

Levy, E. D., Landry, C. R., and Michnick, S. W. (2010). Signaling Through Cooperation. *Science*, 328(5981):983–984.

Levy, E. D., Michnick, S. W., and Landry, C. R. (2012). Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1602):2594–2606.

Li, L., Shakhnovich, E. I., and Mirny, L. A. (2003). Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proceedings of the National Academy of Sciences*, 100(8):4463–4468.

Li, T., Li, F., and Zhang, X. (2008). Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins*, 70(2):404–414.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology*, 257(2):342–358.

Lin, L.-L., Hsu, C.-L., Hu, C.-W., Ko, S.-Y., Hsieh, H.-L., Huang, H.-C., and Juan, H.-F. (2015a). Integrating Phosphoproteomics and Bioinformatics to Study Brassinosteroid-Regulated Phosphorylation Dynamics in Arabidopsis. *BMC Genomics*, 16(1).

Lin, M.-H., Sugiyama, N., and Ishihama, Y. (2015b). Systematic profiling of the bacterial phosphoproteome reveals bacterium-specific features of phosphorylation. *Science Signaling*, 8(394):rs10.

Linch, M., Sanz-Garcia, M., Soriano, E., Zhang, Y., Riou, P., Rosse, C., Cameron, A., Knowles, P., Purkiss, A., Kjaer, S., McDonald, N. Q., and Parker, P. J. (2013). A Cancer-Associated Mutation in Atypical Protein Kinase Cι Occurs in a Substrate-Specific Recruitment Motif. *Sci. Signal.*, 6(293):ra82–ra82.

Lindberg, R. A., Quinn, A. M., and Hunter, T. (1992). Dual-specificity protein kinases: will any hydroxyl do? *Trends in Biochemical Sciences*, 17(3):114–119.

Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B., and Pawson, T. (2008). NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Research*, 36(Database issue):D695–D699.

Lipmann, F. and Levene, P. (1932). Serinephosphoric acid obtained on hydrolysis of vitellinic acid. *J. Biol. Chem.*, 98(1):109–114.

Lippa, B., Pan, G., Corbett, M., Li, C., Kauffman, G. S., Pandit, J., Robinson, S., Wei, L., Kozina, E., Marr, E. S., Borzillo, G., Knauth, E., Barbacci-Tobin, E. G., Vincent, P., Troutman, M., Baker, D., Rajamohan, F., Kakar, S., Clark, T., and Morris, J. (2008). Synthesis and structure based optimization of novel Akt inhibitors. *Bioorganic & medicinal chemistry letters*, 18(11):3359–3363.

Liu, B. A., Shah, E., Jablonowski, K., Stergachis, A., Engelmann, B., and Nash, P. D. (2011). The SH2 Domain–Containing Proteins in 21 Species Establish the Provenance and Scope of Phosphotyrosine Signaling in Eukaryotes. *Sci. Signal.*, 4(202):ra83–ra83.

Lizcano, J. M., Göransson, O., Toth, R., Deak, M., Morrice, N. A., Boudeau, J., Hawley, S. A., Udd, L., Mäkelä, T. P., Hardie, D. G., and Alessi, D. R. (2004). LKB1 is a master kinase that activates 13 kinases of the AMPK subfamily, including MARK/PAR-1. *The EMBO journal*, 23(4):833–843.

Lodowski, D. T., Tesmer, V. M., Benovic, J. L., and Tesmer, J. J. G. (2006). The Structure of G Protein-coupled Receptor Kinase (GRK)-6 Defines a Second Lineage of GRKs. *Journal of Biological Chemistry*, 281(24):16785–16793.

Lubner, J. M., Church, G. M., Chou, M. F., and Schwartz, D. (2016). Reprogramming protein kinase substrate specificity through synthetic mutations. *bioRxiv*, page 091892.

Lubner, J. M., Dodge-Kafka, K. L., Carlson, C. R., Church, G. M., Chou, M. F., and Schwartz, D. (2017). Cushing's Syndrome mutant PKAL205r exhibits altered substrate specificity. *FEBS letters*, 591(3):459–467.

Lv, D.-W., Subburaj, S., Cao, M., Yan, X., Li, X., Appels, R., Sun, D.-F., Ma, W., and Yan, Y.-M. (2014). Proteome and phosphoproteome characterization reveals new response and defense mechanisms of Brachypodium distachyon leaves under salt stress. *Molecular & cellular proteomics: MCP*, 13(2):632–652.

Maichele, A. J., Burwinkel, B., Maire, I., Søvik, O., and Kilimann, M. W. (1996). Mutations in the testis/liver isoform of the phosphorylase kinase γ subunit (PHKG2) cause autosomal liver glycogenosis in the gsd rat and in humans. *Nature Genetics*, 14(3):337–340.

Maiolica, A., De, M. M.-R., Schoof, E. M., Chaikuad, A., Villa, F., Gatti, M., Jeganathan, S., Lou, H. J., Novy, K., Hauri, S., Toprak, U. H., Herzog, F., Meraldi, P., Penengo, L., Turk, B. E., Knapp, S., Linding, R., and Aebersold, R. (2014). Modulation of the chromatin phosphoproteome by the Haspin protein kinase. *Molecular & cellular proteomics : MCP*, 13(7):1724–1740.

Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002a). Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences*, 27(10):514–520.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002b). The protein kinase complement of the human genome. *Science (New York, N.Y.)*, 298(5600):1912–1934.

Manning, G., Young, S. L., Miller, W. T., and Zhai, Y. (2008). The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proceedings of the National Academy of Sciences of the United States of America*, 105(28):9674–9679.

Marchini, F. K., Godoy, L. M. F. d., Rampazzo, R. C. P., Pavoni, D. P., Probst, C. M., Gnad, F., Mann, M., and Krieger, M. A. (2011). Profiling the Trypanosoma cruzi Phosphoproteome. *PLOS ONE*, 6(9):e25381.

Marcon, C., Malik, W. A., Walley, J. W., Shen, Z., Paschold, A., Smith, L. G., Piepho, H.-P., Briggs, S. P., and Hochholdinger, F. (2015). A high-resolution tissue-specific proteome and phosphoproteome atlas of maize primary roots reveals functional gradients along the root axes. *Plant Physiology*, 168(1):233–246.

Mathews, S., Tsai, R. C., and Kellogg, E. A. (2000). Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. *American Journal of Botany*, 87(1):96–107.

Matthews, L. A. and Guarné, A. (2013). Dbf4: the whole is greater than the sum of its parts. *Cell Cycle (Georgetown, Tex.)*, 12(8):1180–1188.

McGowan, C. H. and Russell, P. (1993). Human Wee1 kinase inhibits cell division by phosphorylating p34cdc2 exclusively on Tyr15. *The EMBO journal*, 12(1):75–85.

McKeown, A. N., Bridgham, J. T., Anderson, D. W., Murphy, M. N., Ortlund, E. A., and Thornton, J. W. (2014). Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell*, 159(1):58–68.

McSkimming, D. I., Rasheed, K., and Kannan, N. (2017). Classifying kinase conformations using a machine learning approach. *BMC Bioinformatics*, 18:86.

Miller, C. J. and Turk, B. E. (2018). Homing in: Mechanisms of Substrate Targeting by Protein Kinases. *Trends in Biochemical Sciences*, 43(5):380–394.

Miller, M. L., Jensen, L. J., Diella, F., Jørgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., and Linding, R. (2008). Linear motif atlas for phosphorylation-dependent signaling. *Science Signaling*, 1(35):ra2.

Miller, S. R. (2017). An appraisal of the enzyme stability-activity trade-off. *Evolution; International Journal of Organic Evolution*, 71(7):1876–1887.

Miranda-Saavedra, D. and Barton, G. J. (2007). Classification and functional annotation of eukaryotic protein kinases. *Proteins*, 68(4):893–914.

Miranda-Saavedra, D., Gabaldón, T., Barton, G. J., Langsley, G., and Doerig, C. (2012). The kinomes of apicomplexan parasites. *Microbes and Infection*, 14(10):796–810.

Mochly-Rosen, D., Khaner, H., and Lopez, J. (1991). Identification of intracellular receptor proteins for activated protein kinase C. *Proceedings of the National Academy of Sciences of the United States of America*, 88(9):3997–4000.

Mohanty, S., Oruganty, K., Kwon, A., Byrne, D. P., Ferries, S., Ruan, Z., Hanold, L. E., Katiyar, S., Kennedy, E. J., Eyers, P. A., and Kannan, N. (2016). Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLOS Genetics*, 12(2):e1005885.

Mok, J., Im, H., and Snyder, M. (2009). Global identification of protein kinase substrates by protein microarray analysis. *Nature Protocols*, 4(12):1820–1827.

Mok, J., Kim, P. M., Lam, H. Y. K., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L. N., Sheu, Y.-J., Sassi, H. E., Sopko, R., Chan, C. S. M., De Virgilio, C., Hollingsworth, N. M., Lim, W. A., Stern, D. F., Stillman, B., Andrews, B. J., Gerstein, M. B., Snyder, M., and Turk, B. E. (2010). Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. *Science Signaling*, 3(109):ra12.

Moore, M. J., Adams, J. A., and Taylor, S. S. (2003). Structural Basis for Peptide Binding in Protein Kinase A ROLE OF GLUTAMIC ACID 203 AND TYROSINE 204 IN THE PEPTIDE-POSITIONING LOOP. *Journal of Biological Chemistry*, 278(12):10613–10618.

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 42(D1):D374–D379.

Moses, A. M., Hériché, J.-K., and Durbin, R. (2007). Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biology*, 8(2):R23.

Mushegian, A., Gurevich, V. V., and Gurevich, E. V. (2012). The Origin and Evolution of G Protein-Coupled Receptor Kinases. *PLOS ONE*, 7(3):e33806.

Muñoz-Dorado, J., Inouye, S., and Inouye, M. (1991). A gene encoding a protein serine/threonine kinase is required for normal development of M. xanthus, a gram-negative bacterium. *Cell*, 67(5):995–1006.

Nesić, D., Miller, M. C., Quinkert, Z. T., Stein, M., Chait, B. T., and Stebbins, C. E. (2010). Helicobacter pylori CagA inhibits PAR1-MARK family kinases by mimicking host substrates. *Nature Structural & Molecular Biology*, 17(1):130–132.

Nett, I. R. E., Martin, D. M. A., Miranda-Saavedra, D., Lamont, D., Barber, J. D., Mehlert, A., and Ferguson, M. A. J. (2009). The phosphoproteome of bloodstream form Trypanosoma brucei, causative agent of African sleeping sickness. *Molecular & cellular proteomics: MCP*, 8(7):1527–1538.

Newman, R. H., Hu, J., Rho, H.-S., Xie, Z., Woodard, C., Neiswinger, J., Cooper, C., Shirley, M., Clark, H. M., Hu, S., Hwang, W., Jeong, J. S., Wu, G., Lin, J., Gao, X., Ni, Q., Goel, R., Xia, S., Ji, H., Dalby, K. N., Birnbaum, M. J., Cole, P. A., Knapp, S., Ryazanov, A. G.,

Zack, D. J., Blackshaw, S., Pawson, T., Gingras, A.-C., Desiderio, S., Pandey, A., Turk, B. E., Zhang, J., Zhu, H., and Qian, J. (2013). Construction of human activity-based phosphorylation networks. *Molecular Systems Biology*, 9:655.

Ngo, J. C. K., Chakrabarti, S., Ding, J.-H., Velazquez-Dones, A., Nolen, B., Aubol, B. E., Adams, J. A., Fu, X.-D., and Ghosh, G. (2005). Interplay between SRPK and Clk/Sty kinases in phosphorylation of the splicing factor ASF/SF2 is regulated by a docking motif in ASF/SF2. *Molecular Cell*, 20(1):77–89.

Nguyen, T. H. N., Brechenmacher, L., Aldrich, J. T., Clauss, T. R., Gritsenko, M. A., Hixson, K. K., Libault, M., Tanaka, K., Yang, F., Yao, Q., Paša-Tolić, L., Xu, D., Nguyen, H. T., and Stacey, G. (2012). Quantitative Phosphoproteomic Analysis of Soybean Root Hairs Inoculated with Bradyrhizobium japonicum. *Molecular & Cellular Proteomics*, 11(11):1140–1155.

Niefind, K., Yde, C. W., Ermakova, I., and Issinger, O.-G. (2007). Evolved to Be Active: Sulfate Ions Define Substrate Recognition Sites of CK2α and Emphasise its Exceptional Role within the CMGC Family of Eukaryotic Protein Kinases. *Journal of Molecular Biology*, 370(3):427–438.

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. M., and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4:e04837.

Nolen, B., Taylor, S., and Ghosh, G. (2004). Regulation of Protein Kinases: Controlling Activity through Activation Segment Conformation. *Molecular Cell*, 15(5):661–675.

Nolen, B., Yun, C. Y., Wong, C. F., McCammon, J. A., Fu, X. D., and Ghosh, G. (2001). The structure of Sky1p reveals a novel mechanism for constitutive activity. *Nature structural biology*, 8(2):176–183.

Nuin, P. A., Wang, Z., and Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, 7:471.

Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003). Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, 31(13):3635–3641.

Okuno, S., Kitani, T., and Fujisawa, H. (1997). Studies on the substrate specificity of Ca2+/calmodulin-dependent protein kinase kinase alpha. *Journal of Biochemistry*, 122(2):337–343.

Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006). Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell*, 127(3):635–648.

Onorato, J. J., Palczewski, K., Regan, J. W., Caron, M. G., Lefkowitz, R. J., and Benovic, J. L. (1991). Role of acidic amino acids in peptide substrates of the .beta.-adrenergic receptor kinase and rhodopsin kinase. *Biochemistry*, 30(21):5118–5125.

Oruganty, K. and Kannan, N. (2012). Design principles underpinning the regulatory diversity of protein kinases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1602):2529–2539.

Oruganty, K., Talevich, E. E., Neuwald, A. F., and Kannan, N. (2016). Identification and classification of small molecule kinases: insights into substrate recognition and specificity. *BMC Evolutionary Biology*, 16.

Pan, Z., Wang, B., Zhang, Y., Wang, Y., Ullah, S., Jian, R., Liu, Z., and Xue, Y. (2015). dbPSP: a curated database for protein phosphorylation sites in prokaryotes. *Database: The Journal of Biological Databases and Curation*, 2015:bav031.

Pandya, S., Struck, T. J., Mannakee, B. K., Paniscus, M., and Gutenkunst, R. N. (2015). Testing whether Metazoan Tyrosine Loss Was Driven by Selection against Promiscuous Phosphorylation. *Molecular Biology and Evolution*, 32(1):144–152.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 20(2):289–290.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods11edited by J. Thornton. *Journal of Molecular Biology*, 284(4):1201–1210.

Park, J.-E., Soung, N.-K., Yoshikazu, J., Kang, Y. H., Liao, C., Lee, K. H., Park, C. H., Nicklaus, M. C., and Lee, K. S. (2010). Polo-Box Domain: a versatile mediator of polo-like kinase function. *Cellular and molecular life sciences : CMLS*, 67(12):1957–1970.

Pawson, T. and Scott, J. D. (2005). Protein phosphorylation in signaling – 50 years and counting. *Trends in Biochemical Sciences*, 30(6):286–290.

Pearce, L. R., Komander, D., and Alessi, D. R. (2010). The nuts and bolts of AGC protein kinases. *Nature Reviews Molecular Cell Biology*, 11(1):9–22.

Pearson, G., Robinson, F., Beers Gibson, T., Xu, B. E., Karandikar, M., Berman, K., and Cobb, M. H. (2001). Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. *Endocrine Reviews*, 22(2):153–183.

Pearson, R. B. and Kemp, B. E. (1991). Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. *Methods in Enzymology*, 200:62–81.

Pease, B. N., Huttlin, E. L., Jedrychowski, M. P., Talevich, E., Harmon, J., Dillman, T., Kannan, N., Doerig, C., Chakrabarti, R., Gygi, S. P., and Chakrabarti, D. (2013). Global analysis of protein expression and phosphorylation of three stages of Plasmodium falciparum intraerythrocytic development. *Journal of Proteome Research*, 12(9):4028–4045.

Pereira, S. F. F., Goss, L., and Dworkin, J. (2011). Eukaryote-like serine/threonine kinases and phosphatases in bacteria. *Microbiology and molecular biology reviews: MMBR*, 75(1):192–212.

Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H., Rashmi, B., Ramya, M., Zhao, Z., Chandrika, K., Padma, N., Harsha, H., Yatish, A., Kavitha, M., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003). Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research*, 13(10):2363–2371.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802.

Pike, A. C. W., Rellos, P., Niesen, F. H., Turnbull, A., Oliver, A. W., Parker, S. A., Turk, B. E., Pearl, L. H., and Knapp, S. (2008). Activation segment dimerization: a mechanism for kinase autophosphorylation of non-consensus sites. *The EMBO Journal*, 27(4):704–714.

Pillay, T. S. (2004). A fisherman's tale: Phage display as a discovery tool. *Discovery Medicine*, 4(23):315–318.

Pincus, D., Letunic, I., Bork, P., and Lim, W. A. (2008). Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proceedings of the National Academy of Sciences*, 105(28):9680–9684.

Pinkse, M. W. H., Mohammed, S., Gouw, J. W., van Breukelen, B., Vos, H. R., and Heck, A. J. R. (2008). Highly Robust, Automated, and Sensitive Online TiO2-Based Phosphoproteomics Applied To Study Endogenous Phosphorylation in Drosophila melanogaster. *Journal of Proteome Research*, 7(2):687–697.

Plewczyński, D., Tkacz, A., Godzik, A., and Rychlewski, L. (2005). A support vector machine approach to the identification of phosphorylation sites. *Cellular & Molecular Biology Letters*, 10(1):73–89.

Plowman, G. D., Sudarsanam, S., Bingham, J., Whyte, D., and Hunter, T. (1999). The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):13603–13610.

Potel, C. M., Lin, M.-H., Heck, A. J. R., and Lemeer, S. (2018). Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics. *Nature Methods*, 15(3):187–190.

Prisic, S., Dankwa, S., Schwartz, D., Chou, M. F., Locasale, J. W., Kang, C.-M., Bemis, G., Church, G. M., Steen, H., and Husson, R. N. (2010). Extensive phosphorylation with overlapping specificity by Mycobacterium tuberculosis serine/threonine protein kinases. *Proceedings of the National Academy of Sciences*, 107(16):7521–7526.

Prowse, C. N., Deal, M. S., and Lew, J. (2001). The Complete Pathway for Catalytic Activation of the Mitogen-activated Protein Kinase, ERK2. *Journal of Biological Chemistry*, 276(44):40817–40823.

Pérez, J., Castañeda-García, A., Jenke-Kodama, H., Müller, R., and Muñoz-Dorado, J. (2008). Eukaryotic-like protein kinases in the prokaryotes and the myxobacterial kinome. *Proceedings of the National Academy of Sciences*, 105(41):15950–15955.

Randall, R. N., Radford, C. E., Roof, K. A., Natarajan, D. K., and Gaucher, E. A. (2016). An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nature Communications*, 7:12847.

Reimann, J., Esser, D., Orell, A., Amman, F., Pham, T. K., Noirel, J., Lindås, A.-C., Bernander, R., Wright, P. C., Siebers, B., and Albers, S.-V. (2013). Archaeal Signal Transduction: Impact of Protein Phosphatase Deletions on Cell Size, Motility, and Energy Metabolism in Sulfolobus acidocaldarius. *Molecular & Cellular Proteomics : MCP*, 12(12):3908–3923.

Resjö, S., Ali, A., Meijer, H. J. G., Seidl, M. F., Snel, B., Sandin, M., Levander, F., Govers, F., and Andreasson, E. (2014). Quantitative label-free phosphoproteomics of six different life stages of the late blight pathogen Phytophthora infestans reveals abundant phosphorylation of members of the CRN effector family. *Journal of Proteome Research*, 13(4):1848–1859.

Rhoads, T. W., Prasad, A., Kwiecien, N. W., Merrill, A. E., Zawack, K., Westphall, M. S., Schroeder, F. C., Kimble, J., and Coon, J. J. (2015). NeuCode Labeling in Nematodes: Proteomic and Phosphoproteomic Impact of Ascaroside Treatment in Caenorhabditis elegans. *Molecular & cellular proteomics: MCP*, 14(11):2922–2935.

Riley, N. M. and Coon, J. J. (2016). Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. *Analytical Chemistry*, 88(1):74–94.

Ringrose, J. H., Toorn, H. W. P. v. d., Eitel, M., Post, H., Neerincx, P., Schierwater, B., Altelaar, A. F. M., and Heck, A. J. R. (2013). Deep proteome profiling of *Trichoplax adhaerens* reveals remarkable features at the origin of metazoan multicellularity. *Nature Communications*, 4:1408.

Ritz, A., Shakhnarovich, G., Salomon, A. R., and Raphael, B. J. (2009). Discovery of phosphorylation motif mixtures in phosphoproteomics data. *Bioinformatics*, 25(1):14–21.

Rogozin, I. B., Managadze, D., Shabalina, S. A., and Koonin, E. V. (2014). Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biology and Evolution*, 6(4):754–762.

Rose, C. M., Venkateshwaran, M., Volkening, J. D., Grimsrud, P. A., Maeda, J., Bailey, D. J., Park, K., Howes-Podoll, M., den Os, D., Yeun, L. H., Westphall, M. S., Sussman, M. R., Ané, J.-M., and Coon, J. J. (2012). Rapid phosphoproteomic and transcriptomic changes in the rhizobia-legume symbiosis. *Molecular & cellular proteomics: MCP*, 11(9):724–744.

Roskoski, R. (2004). Src protein-tyrosine kinase structure and regulation. *Biochemical and Biophysical Research Communications*, 324(4):1155–1164.

Roskoski, R. (2007). Enzyme Structure and Function. In *xPharm: The Comprehensive Pharmacology Reference*, pages 1–7. Elsevier, New York.

Roskoski, R. (2012). MEK1/2 dual-specificity protein kinases: structure and regulation. *Biochemical and Biophysical Research Communications*, 417(1):5–10.

Roskoski, R. (2015). A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacological Research*, 100:1–23.

Safaei, J., Maňuch, J., Gupta, A., Stacho, L., and Pelech, S. (2011). Prediction of 492 human protein kinase substrate specificities. *Proteome Science*, 9(Suppl 1):S6.

Saito, N. and Shirai, Y. (2002). Protein kinase C gamma (PKC gamma): function of neuron specific isotype. *Journal of Biochemistry*, 132(5):683–687.

Salvi, M., Trashi, E., Cozza, G., Franchin, C., Arrigoni, G., and Pinna, L. A. (2012). Investigation on PLK2 and PLK3 substrate recognition. *Biochimica Et Biophysica Acta*, 1824(12):1366–1373.

Sang, D., Pinglay, S., Vatansever, S., Lou, H. J., Turk, B. E., Gumus, Z. H., and Holt, L. J. (2018). Ancestral resurrection reveals mechanisms of kinase regulatory evolution. *bioRxiv*, page 331637.

Sarno, S., Vaglio, P., Marin, O., Issinger, O.-G., Ruffato, K., and Pinna, L. A. (1997). Mutational Analysis of Residues Implicated in the Interaction between Protein Kinase CK2 and Peptide Substrates. *Biochemistry*, 36(39):11717–11724.

Saunders, N. F. W., Brinkworth, R. I., Huber, T., Kemp, B. E., and Kobe, B. (2008). Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC bioinformatics*, 9:245.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E., and Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(Database issue):D5–15.

Schechtman, D. and Mochly-Rosen, D. (2001). Adaptor proteins in protein kinase C-mediated signal transduction. *Oncogene*, 20(44):6339–6347.

Scheeff, E. D. and Bourne, P. E. (2005). Structural evolution of the protein kinase-like superfamily. *PLoS computational biology*, 1(5):e49.

Schluter, D., Price, T., Mooers, A. O., and Ludwig, D. (1997). LIKELIHOOD OF ANCESTOR STATES IN ADAPTIVE RADIATION. *Evolution; International Journal of Organic Evolution*, 51(6):1699–1711.

Schmitz, R., Baumann, G., and Gram, H. (1996). Catalytic specificity of phosphotyrosine kinases Blk, Lyn, c-Src and Syk as assessed by phage display. *Journal of Molecular Biology*, 260(5):664–677.

Schulman, B. A., Lindstrom, D. L., and Harlow, E. (1998). Substrate recruitment to cyclin-dependent kinase 2 by a multipurpose docking site on cyclin A. *Proceedings of the National Academy of Sciences of the United States of America*, 95(18):10453–10458.

Schwartz, D. and Gygi, S. P. (2005). An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nature Biotechnology*, 23(11):1391–1398.

Schweiger, R. and Linial, M. (2010). Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biology Direct*, 5:6.

Scott, J. W., Norman, D. G., Hawley, S. A., Kontogiannis, L., and Hardie, D. G. (2002). Protein kinase substrate recognition studied using the recombinant catalytic domain of AMP-activated protein kinase and a model substrate. *Journal of Molecular Biology*, 317(2):309–323.

Sente, A., Peer, R., Srivastava, A., Baidya, M., Lesk, A. M., Balaji, S., Shukla, A. K., Babu, M. M., and Flock, T. (2018). Molecular mechanism of modulating arrestin conformation by GPCR phosphorylation. *Nature Structural & Molecular Biology*, 25(6):538–545.

Sharma, K., D'Souza, R. C. J., Tyanova, S., Schaab, C., Wiśniewski, J. R., Cox, J., and Mann, M. (2014). Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Reports*, 8(5):1583–1594.

Shaw, R. J., Kosmatka, M., Bardeesy, N., Hurley, R. L., Witters, L. A., DePinho, R. A., and Cantley, L. C. (2004). The tumor suppressor LKB1 kinase directly activates AMP-activated kinase and regulates apoptosis in response to energy stress. *Proceedings of the National Academy of Sciences of the United States of America*, 101(10):3329–3335.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2016). Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3 (Bethesda, Md.)*, 6(12):3927–3939.

Shoji, S., Parmelee, D. C., Wade, R. D., Kumar, S., Ericsson, L. H., Walsh, K. A., Neurath, H., Long, G. L., Demaille, J. G., Fischer, E. H., and Titani, K. (1981). Complete amino acid sequence of the catalytic subunit of bovine cardiac muscle cyclic AMP-dependent protein kinase. *Proceedings of the National Academy of Sciences of the United States of America*, 78(2):848–851.

Siddiq, M. A., Hochberg, G. K., and Thornton, J. W. (2017). Evolution of protein specificity: insights from ancestral protein reconstruction. *Current Opinion in Structural Biology*, 47:113–122.

Sigrist, C. J. A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, 3(3):265–274.

Simon, B., Huart, A.-S., Temmerman, K., Vahokoski, J., Mertens, H. D. T., Komadina, D., Hoffmann, J.-E., Yumerefendi, H., Svergun, D. I., Kursula, P., Schultz, C., McCarthy, A. A., Hart, D. J., and Wilmanns, M. (2016). Death-Associated Protein Kinase Activity Is Regulated by Coupled Calcium/Calmodulin Binding to Two Distinct Sites. *Structure*, 24(6):851–861.

Skjærven, L., Jariwala, S., Yao, X.-Q., Idé, J., and Grant, B. J. (2016). The Bio3d Project: Interactive Tools for Structural Bioinformatics. *Biophysical Journal*, 110(3):379a.

Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H., and Cantley, L. C. (1994). Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Current biology: CB*, 4(11):973–982.

Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. A., Soderling, T. R., Bartleson, C., Graves, D. J., DeMaggio, A. J., Hoekstra, M. F., Blenis, J., Hunter, T., and Cantley, L. C. (1996). A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Molecular and Cellular Biology*, 16(11):6486–6493.

Sorrell, F. J., Szklarz, M., Abdul Azeez, K. R., Elkins, J. M., and Knapp, S. (2016). Family-wide Structural Analysis of Human Numb-Associated Protein Kinases. *Structure (London, England: 1993)*, 24(3):401–411.

Soundararajan, M., Roos, A. K., Savitsky, P., Filippakopoulos, P., Kettenbach, A. N., Olsen, J. V., Gerber, S. A., Eswaran, J., Knapp, S., and Elkins, J. M. (2013). Structures of Down syndrome kinases, DYRKs, reveal mechanisms of kinase activation and substrate recognition. *Structure (London, England : 1993)*, 21(6):986–996.

Soyer, O. S. and Goldstein, R. A. (2004). Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. *Journal of Molecular Biology*, 339(1):227–242.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. A., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N. H., Stanke, M., Adamska, M., Darling, A., Degnan, S. M., Oakley, T. H., Plachetzki, D. C., Zhai, Y., Adamski, M., Calcino, A., Cummins, S. F., Goodstein, D. M., Harris, C., Jackson, D. J., Leys, S. P., Shu, S., Woodcroft, B. J., Vervoort, M., Kosik, K. S., Manning, G., Degnan, B. M., and Rokhsar, D. S. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, 466(7307):720–726.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stancik, I. A., Šestak, M. S., Ji, B., Axelson-Fisk, M., Franjevic, D., Jers, C., Domazet-Lošo, T., and Mijakovic, I. (2018). Serine/Threonine Protein Kinases from Bacteria, Archaea and Eukarya Share a Common Evolutionary Origin Deeply Rooted in the Tree of Life. *Journal of Molecular Biology*, 430(1):27–32.

Stark, C., Su, T.-C., Breitkreutz, A., Lourenco, P., Dahabieh, M., Breitkreutz, B.-J., Tyers, M., and Sadowski, I. (2010). PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. *Database: The Journal of Biological Databases and Curation*, 2010:bap026.

Starr, T. N., Flynn, J. M., Mishra, P., Bolon, D. N. A., and Thornton, J. W. (2018). Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proceedings of the National Academy of Sciences*, page 201718133.

Steichen, J. M., Iyer, G. H., Li, S., Saldanha, S. A., Deal, M. S., Woods, V. L., and Taylor, S. S. (2010). Global Consequences of Activation Loop Phosphorylation on Protein Kinase A. *The Journal of Biological Chemistry*, 285(6):3825–3832.

Stenberg, K. A. E., Riikonen, P. T., and Vihinen, M. (1999). KinMutBase, a database of human disease-causing protein kinase mutations. *Nucleic Acids Research*, 27(1):362–364.

Stock, A. M., Robinson, V. L., and Goudreau, a. P. N. (2000). Two-Component Signal Transduction. *Annual Review of Biochemistry*, 69(1):183–215.

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *The Biochemical Journal*, 449(3):581–594.

Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ? *Trends in genetics: TIG*, 25(5):210–216.

Studer, R. A., Rodriguez-Mias, R. A., Haas, K. M., Hsu, J. I., Viéitez, C., Solé, C., Swaney, D. L., Stanford, L. B., Liachko, I., Böttcher, R., Dunham, M. J., de Nadal, E., Posas, F., Beltrao, P., and Villén, J. (2016). Evolution of protein phosphorylation across 18 fungal species. *Science (New York, N.Y.)*, 354(6309):229–232.

Su, M.-G. and Lee, T.-Y. (2013). Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures. *BMC bioinformatics*, 14 Suppl 16:S2.

Swaffer, M. P., Jones, A. W., Flynn, H. R., Snijders, A. P., and Nurse, P. (2016). CDK Substrate Phosphorylation and Ordering the Cell Cycle. *Cell*, 167(7):1750–1761.e16.

Talevich, E., Tobin, A. B., Kannan, N., and Doerig, C. (2012). An evolutionary perspective on the kinome of malaria parasites. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1602):2607–2618.

Tamuri, A. U., Reis, M. d., Hay, A. J., and Goldstein, R. A. (2009). Identifying Changes in Selective Constraints: Host Shifts in Influenza. *PLOS Computational Biology*, 5(11):e1000564.

Tan, C. S. H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M. O., Jørgensen, C., Bader, G. D., Aebersold, R., Pawson, T., and Linding, R. (2009a). Comparative Analysis Reveals Conserved Protein Phosphorylation Networks Implicated in Multiple Diseases. *Sci. Signal.*, 2(81):ra39–ra39.

Tan, C. S. H., Pasculescu, A., Lim, W. A., Pawson, T., Bader, G. D., and Linding, R. (2009b). Positive selection of tyrosine loss in metazoan evolution. *Science (New York, N.Y.)*, 325(5948):1686–1688.

Taylor, S. S. and Kornev, A. P. (2011). Protein kinases: evolution of dynamic regulatory proteins., Protein Kinases: Evolution of Dynamic Regulatory Proteins. *Trends in biochemical sciences, Trends in biochemical sciences*, 36, 36(2, 2):65, 65–77.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nature Genetics*, 36(5):492–496.

Telford, M. J., Budd, G. E., and Philippe, H. (2015). Phylogenomic Insights into Animal Evolution. *Current Biology*, 25(19):R876–R887.

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A., and Consortium, o. b. o. t. G. O. (2012). On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLOS Computational Biology*, 8(2):e1002386.

Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews. Genetics*, 5(5):366–375.

Tian, M., Chen, X., Xiong, Q., Xiong, J., Xiao, C., Ge, F., Yang, F., and Miao, W. (2014). Phosphoproteomic Analysis of Protein Phosphorylation Networks in Tetrahymena thermophila, a Model Single-celled Organism. *Molecular & Cellular Proteomics : MCP*, 13(2):503–519.

Tinti, M., Kiemer, L., Costa, S., Miller, M. L., Sacco, F., Olsen, J., Carducci, M., Paoluzi, S., Langone, F., Workman, C., Blom, N., Machida, K., Thompson, C., Schutkowski, M., Brunak, S., Mann, M., Mayer, B., Castagnoli, L., and Cesareni, G. (2013). The SH2 Domain Interaction Landscape. *Cell Reports*, 3(4):1293–1305.

Tokuriki, N., Stricher, F., Serrano, L., and Tawfik, D. S. (2008). How Protein Stability and New Functions Trade Off. *PLOS Computational Biology*, 4(2):e1000002.

Torkamani, A., Kannan, N., Taylor, S. S., and Schork, N. J. (2008). Congenital disease SNPs target lineage specific structural elements in protein kinases. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26):9011–9016.

Traven, A. and Heierhorst, J. (2005). SQ/TQ cluster domains: concentrated ATM/ATR kinase phosphorylation site regions in DNA-damage-response proteins. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 27(4):397–407.

Treeck, M., Sanders, J. L., Elias, J. E., and Boothroyd, J. C. (2011). The phosphoproteomes of Plasmodium falciparum and Toxoplasma gondii reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host & Microbe*, 10(4):410–419.

Trost, B. and Kusalik, A. (2011). Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics (Oxford, England)*, 27(21):2927–2935.

Tsigankov, P., Gherardini, P. F., Helmer-Citterich, M., Späth, G. F., and Zilberstein, D. (2013). Phosphoproteomic analysis of differentiating Leishmania parasites reveals a unique stage-specific phosphorylation motif. *Journal of Proteome Research*, 12(7):3405–3412.

Ubersax, J. A. and Ferrell, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nature Reviews. Molecular Cell Biology*, 8(7):530–541.

Urbaniak, M. D., Martin, D. M. A., and Ferguson, M. A. J. (2013). Global quantitative SILAC phosphoproteomics reveals differential phosphorylation is widespread between the procyclic and bloodstream form lifecycle stages of Trypanosoma brucei. *Journal of Proteome Research*, 12(5):2233–2244.

Vallender, E. J. (2009). Bioinformatic approaches to identifying orthologs and assessing evolutionary relationships. *Methods (San Diego, Calif.)*, 49(1):50–55.

Varjosalo, M., Keskitalo, S., Van Drogen, A., Nurkkala, H., Vichalkovski, A., Aebersold, R., and Gstaiger, M. (2013). The protein interaction landscape of the human CMGC kinase group. *Cell Reports*, 3(4):1306–1320.

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, 41(Database issue):D483–489.

Vlastaridis, P., Kyriakidou, P., Chaliotis, A., Van de Peer, Y., Oliver, S. G., and Amoutzias, G. D. (2017). Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. *GigaScience*, 6(2):1–11.

Wagih, O., Sugiyama, N., Ishihama, Y., and Beltrao, P. (2015). Uncovering phosphorylation-based specificities through functional interaction networks. *Molecular & Cellular Proteomics*, page mcp.M115.052357.

Waller, R. F., Cleves, P. A., Rubio-Brotons, M., Woods, A., Bender, S. J., Edgcomb, V., Gann, E. R., Jones, A. C., Teytelman, L., Dassow, P. v., Wilhelm, S. W., and Collier, J. L. (2018). Strength in numbers: Collaborative science for new experimental model systems. *PLOS Biology*, 16(7):e2006333.

Walsh, D. A., Perkins, J. P., and Krebs, E. G. (1968). An Adenosine 3',5'-Monophosphate-dependant Protein Kinase from Rabbit Skeletal Muscle. *Journal of Biological Chemistry*, 243(13):3763–3765.

Walte, A., Rüben, K., Birner-Gruenberger, R., Preisinger, C., Bamberg-Lemper, S., Hilz, N., Bracher, F., and Becker, W. (2013). Mechanism of dual specificity kinase activity of DYRK1a. *The FEBS journal*, 280(18):4495–4511.

Wang, C., Shang, Y., Yu, J., and Zhang, M. (2012a). Substrate Recognition Mechanism of Atypical Protein Kinase Cs Revealed by the Structure of PKCι in Complex with a Substrate Peptide from Par-3. *Structure*, 20(5):791–801.

Wang, H., Gau, B., Slade, W. O., Juergens, M., Li, P., and Hicks, L. M. (2014). The global phosphoproteome of Chlamydomonas reinhardtii reveals complex organellar phosphorylation in the flagella and thylakoid membrane. *Molecular & cellular proteomics: MCP*, 13(9):2337–2353.

Wang, H.-Y., Lin, W., Dyck, J. A., Yeakley, J. M., Songyang, Z., Cantley, L. C., and Fu, X.-D. (1998). SRPK2: A Differentially Expressed SR Protein-specific Kinase Involved in Mediating the Interaction and Localization of Pre-mRNA Splicing Factors in Mammalian Cells. *The Journal of Cell Biology*, 140(4):737–750.

Wang, T., Kettenbach, A. N., Gerber, S. A., and Bailey-Kellogg, C. (2012b). MMFPh: a maximal motif finder for phosphoproteomics datasets. *Bioinformatics*, 28(12):1562–1570.

Weekes, J., Ball, K. L., Caudwell, F. B., and Hardie, D. G. (1993). Specificity determinants for the AMP-activated protein kinase and its plant homologue analysed using synthetic peptides. *FEBS letters*, 334(3):335–339.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR Analyzing German Business Cycles. In *Data Analysis and Decision Support*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 335–343. Springer, Berlin, Heidelberg.

Wu, W.-L., Lai, S.-J., Yang, J.-T., Chern, J., Liang, S.-Y., Chou, C.-C., Kuo, C.-H., Lai, M.-C., and Wu, S.-H. (2016). Phosphoproteomic analysis of Methanohalophilus portucalensis FDF1t identified the role of protein phosphorylation in methanogenesis and osmoregulation. *Scientific Reports*, 6.

Xing, L., An, Y., Shi, G., Yan, J., Xie, P., Qu, Z., Zhang, Z., Liu, Z., Pan, D., and Xu, Y. (2017). Correlated evolution between CK1δ Protein and the Serine-rich Motif Contributes to Regulating the Mammalian Circadian Clock. *Journal of Biological Chemistry*, 292(1):161–171.

Xu, M., Yu, L., Wan, B., Yu, L., and Huang, Q. (2011). Predicting Inactive Conformations of Protein Kinases Using Active Structures: Conformational Selection of Type-II Inhibitors. *PLOS ONE*, 6(7):e22644.

Xu, Q. and Dunbrack, R. L. (2011). The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Research*, 39(Database issue):D761–D770.

Xu, Q., Malecka, K. L., Fink, L., Jordan, E. J., Duffy, E., Kolander, S., Peterson, J. R., and Dunbrack, R. L. (2015). Identifying three-dimensional structures of autophosphorylation complexes in crystals of protein kinases. *Science Signaling*, 8(405):rs13.

Xue, L. and Tao, W. A. (2013). Current technologies to identify protein kinase substrates in high throughput. *Frontiers in biology*, 8(2):216–227.

Yang, F., Sun, S., Tan, G., Costanzo, M., Hill, D. E., Vidal, M., Andrews, B. J., Boone, C., and Roth, F. P. (2017). Identifying pathogenicity of human variants via paralog-based yeast complementation. *PLOS Genetics*, 13(5):e1006779.

Yang, J., Cron, P., Good, V. M., Thompson, V., Hemmings, B. A., and Barford, D. (2002). Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP-PNP. *Nature Structural Biology*, 9(12):940–944.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, 13(5):555–556.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591.

Yao, Q., Ge, H., Wu, S., Zhang, N., Chen, W., Xu, C., Gao, J., Thelen, J. J., and Xu, D. (2014). P³DB 3.0: From plant phosphorylation sites to protein networks. *Nucleic Acids Research*, 42(Database issue):D1206–1213.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. (2015). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics (Oxford, England)*, 31(1):143–145.

Ye, K., Anton Feenstra, K., Heringa, J., IJzerman, A. P., and Marchiori, E. (2008). Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics*, 24(1):18–25.

Yoshizaki, H. and Okuda, S. (2014). Elucidation of the evolutionary expansion of phosphorylation signaling networks using comparative phosphomotif analysis. *BMC Genomics*, 15(1).

Zetterqvist, O., Ragnarsson, U., Humble, E., Berglund, L., and Engström, L. (1976). The minimum substrate of cyclic AMP-stimulated protein kinase, as studied by synthetic peptides representing the phosphorylatable site of pyruvate kinase (type L) of rat liver. *Biochemical and Biophysical Research Communications*, 70(3):696–703.

Zhai, B., Villén, J., Beausoleil, S. A., Mintseris, J., and Gygi, S. P. (2008). Phosphoproteome Analysis of Drosophila melanogaster Embryos. *Journal of Proteome Research*, 7(4):1675–1682.

Zhu, G., Fujii, K., Belkina, N., Liu, Y., James, M., Herrero, J., and Shaw, S. (2005a). Exceptional disfavor for proline at the P + 1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. *The Journal of Biological Chemistry*, 280(11):10743–10748.

Zhu, G., Fujii, K., Liu, Y., Codrea, V., Herrero, J., and Shaw, S. (2005b). A single pair of acidic residues in the kinase major groove mediates strong substrate preference for P-2 or P-5 arginine in the AGC, CAMK, and STE kinase families. *The Journal of Biological Chemistry*, 280(43):36372–36379.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024.

Appendix A

A.1 Chapter 2 commentary on SDRs

Lysine+3

Of the ten basophilic kinases identified, eight are cyclin-dependent kinases of the CMGC *Group*. In CDK2 cocrystal structures, a lysine at the +3 position forms a hydrogen bond with a phospho-threonine at the 155 domain position in the kinase activation loop (Bao et al., 2011; Brown et al., 1999). Brown *et al* hypothesised that the restricted dihedral angles of the proline at position +1 help to orient the +3 side chain towards the primary activation loop phosphate, suggesting some level of inter-positional dependency for CDK specificity. As stated in the *Introduction*, specific contacts are also formed between K+3 and a cyclin subunit, which likely explains why this specificity is observed in CDKs but not in other proline+1 kinases such as the MAPKs.

Position-1

I detected a single significant preference for leucine (10 kinases: 8 CMGC, 1 AGC, 1CAMK) at the -1 position. Six of the the ten kinases with this preference belong to the MAPK *Family*. The identified preference is supported by peptide library analysis of MAPK3, although the identified preference is weak and to a lesser extent extends to other hydrophobic amino acids such as methionine (Songyang et al., 1996). The basis for this specificity is unclear as the -1 residue in the CMGC structures listed in Table 2.1 do not form contacts with the kinase. A single domain position was implicated from the alignment-based methods as a putative SDR at domain position 5 which is also located on the N-terminal loop.

No other significant preferences at the -1 position were detected. This is consistent with a previous screen of yeast kinase specificity that suggested a general lack of sequence constraint at the -1 position (Mok et al., 2010).

Leucine-2

I detected 8 kinases with a moderate preference for leucine at -2: MAPK7 (human), MAPK8 (human and mouse), MAPK9 (human and mouse), MAPK10 (human and mouse), and YCK2. In this grouping there is also minor selectivity for proline and likewise for the 'proline-directed' kinases there is minor selectivity also for leucine. It is likely therefore that both groups represent the same general specificity for hydrophobic residues at position -2. Domain position 131 of the kinase hinge region is implicated as a putative SDR (valine bias in leucine-directed kinases, I/L otherwise), although is unlikely to be a direct determinant given its distance from the active site.

Position -4

Only one specificity (serine / threonine preference) emerges from the clustering procedure employed. However, this is unlikely to represent a physiological kinase preference. For all serine/threonine preferences identified, I failed to identify a general structural mechanism that could explain this feature either from the analysis performed here or from a survey of the relevant literature.

A.2 Kinase mutations in cancer

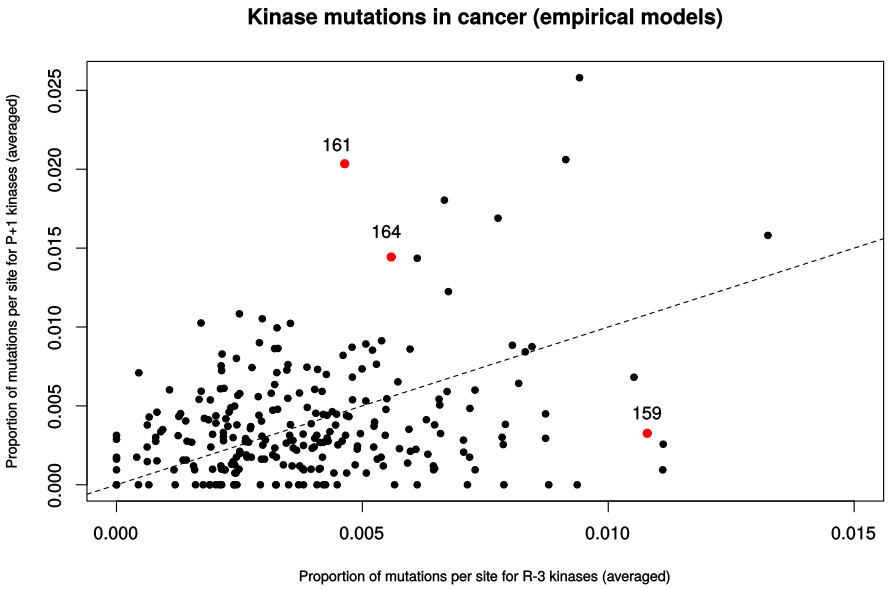


Fig. A.1 The frequency of residue mutations for known R-3 kinases (x-axis) and known P+1 kinases (y-axis). Kinase domain positions 159, 161, and 164 are coloured in red. Here, the kinase classification (P+1 or R-3) was known from kinase specificity models and was not predicted

A.3 Posterior probabilities from the ancestral sequence reconstructions

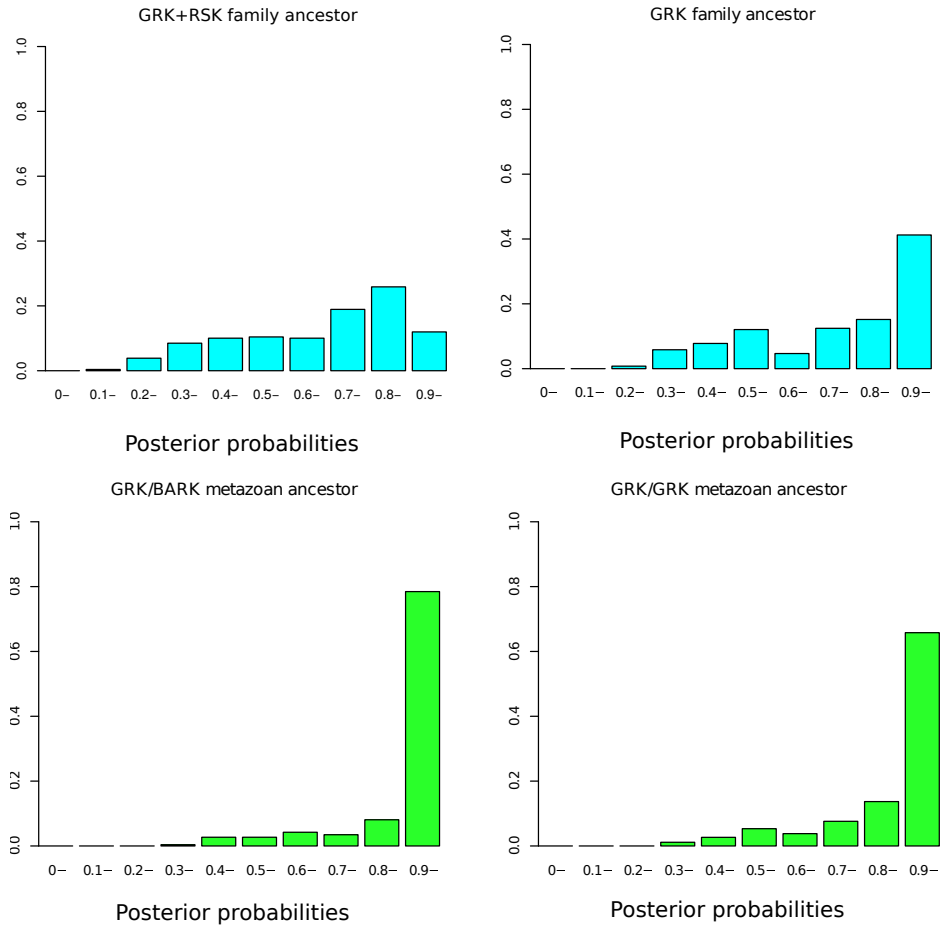


Fig. A.2 Distribution of posterior probabilities across sites for particular nodes of the GRK ancestral sequence reconstruction. The mean probability for the GRK+RSK ancestor is 0.69, for the universal GRK ancestor is 0.77, for the metazoan GRK/BARK ancestral node is 0.92, and for the GRK/GRK ancestral node is 0.88

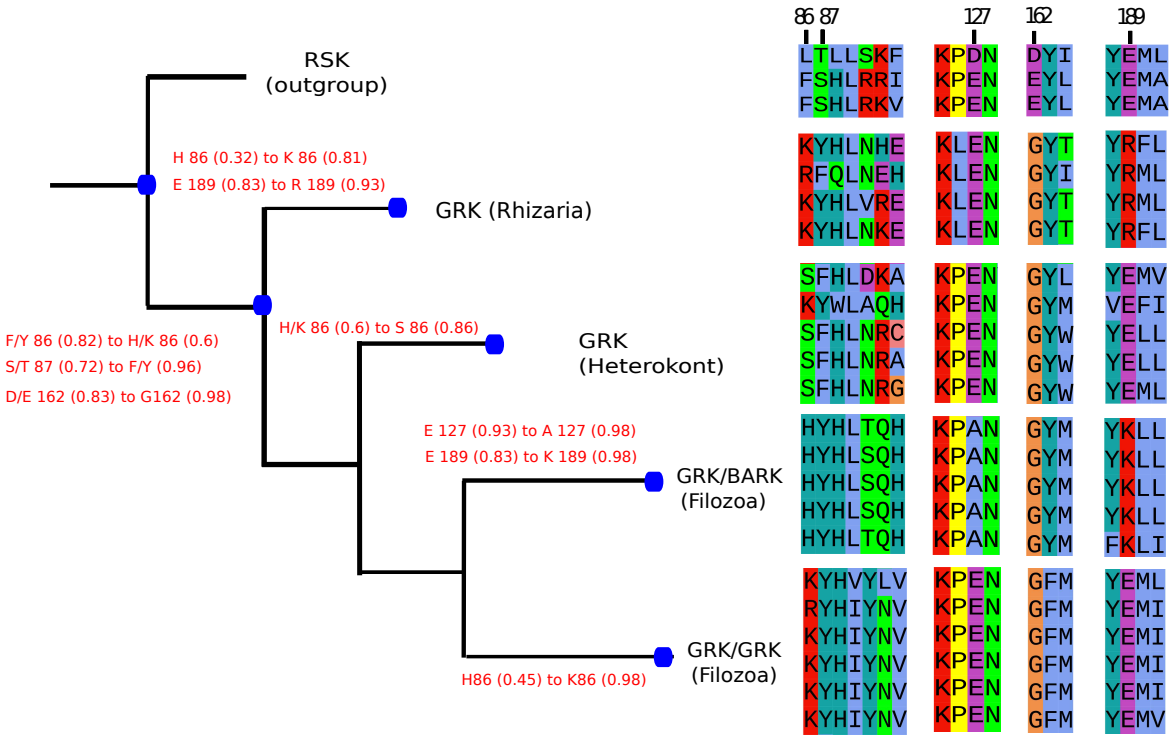


Fig. A.3 Ancestral probabilities for SDR substitutions likely to affect specificity N-terminal to the phosphoacceptor in GRK (as represented in Figure 3.15)

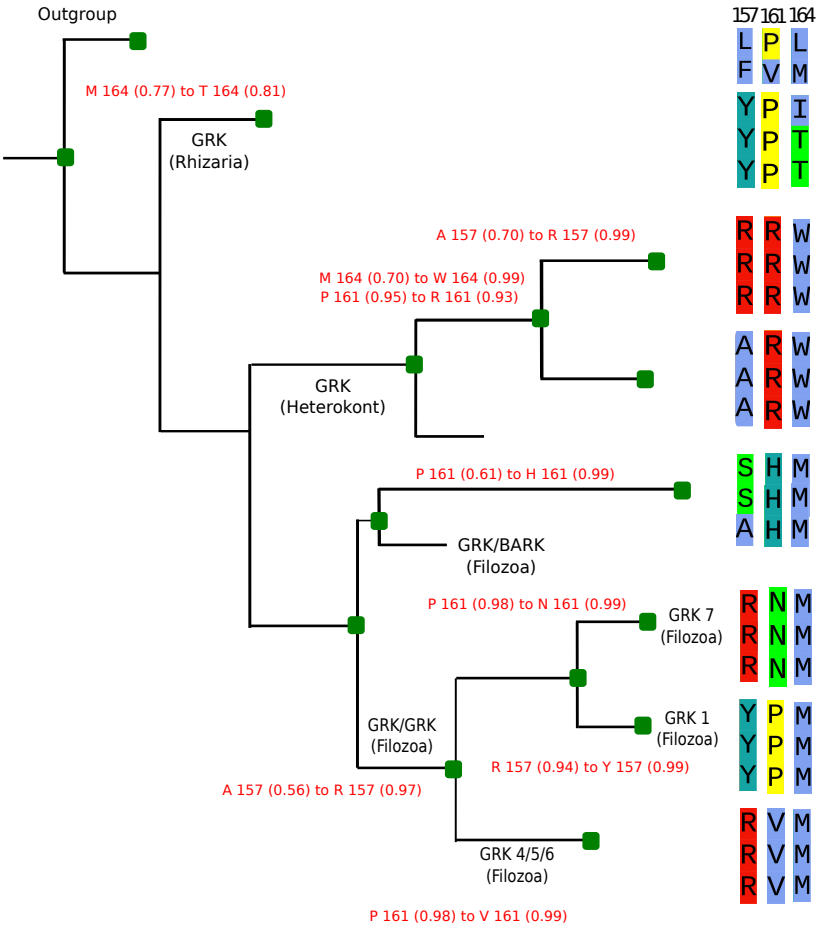


Fig. A.4 Ancestral probabilities for SDR substitutions occurring within the P+1 pocket of GRK (as represented in Figure 3.16)

A.4 Phylogenetic analysis of kinase *Families* and motif enrichments

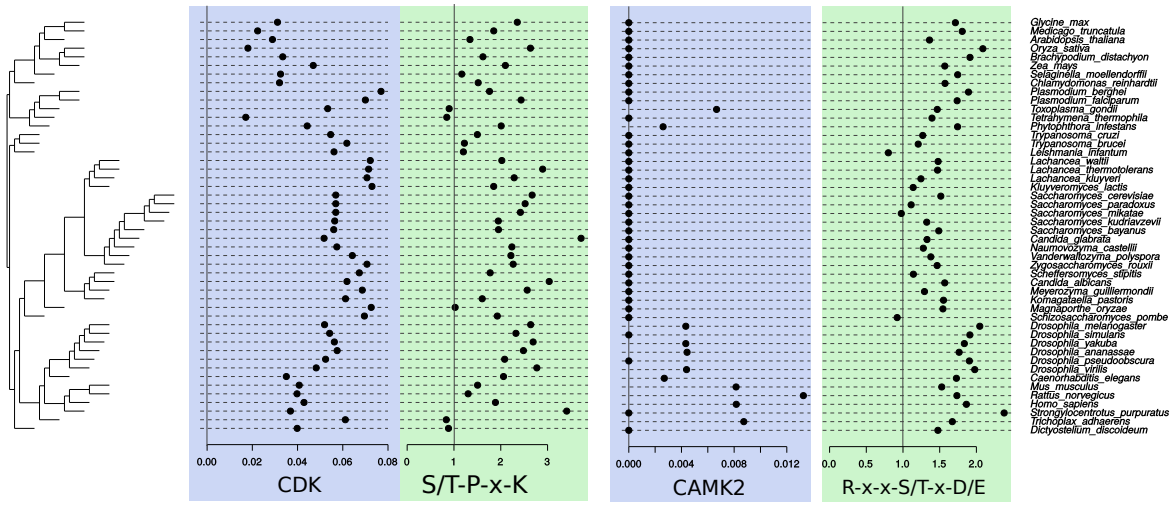


Fig. A.5 For two kinase *Families* (CDK and CAMK2), the relative kinase frequency and the cognate substrate motif enrichment were mapped onto a species phylogeny of the 48 eukaryotes used for this analysis.

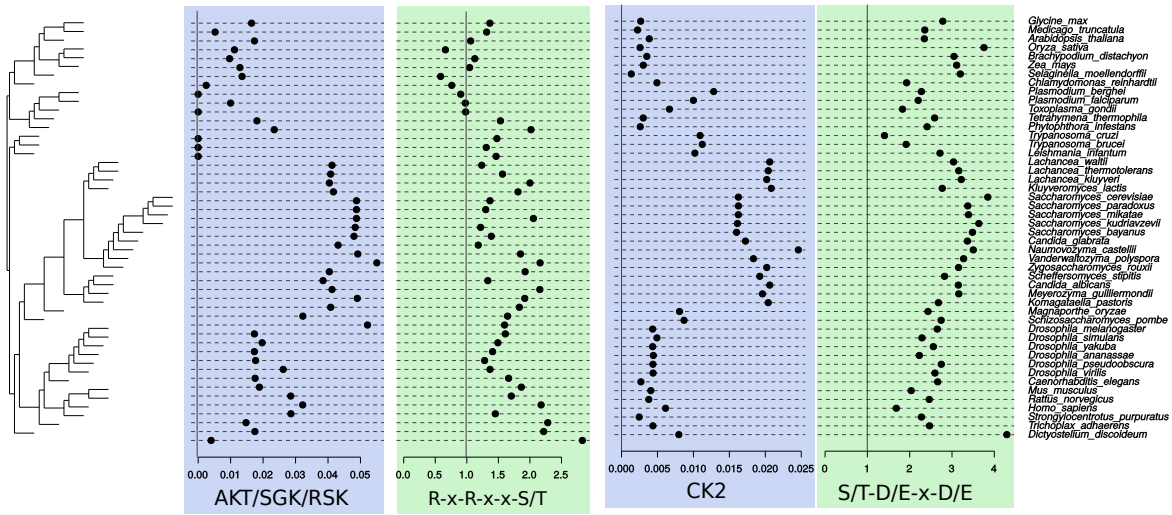


Fig. A.6 For two kinase clades (AKT/SGK/RSK and CK2), the relative kinase frequency and the cognate substrate motif enrichment were mapped onto a species phylogeny of the 48 eukaryotes used for this analysis.

Family	Cmean	I	K	K.star	Lambda
AKT	<0.01	<0.01	<0.01	<0.01	<0.01
CAMK2	<0.01	<0.01	<0.01	<0.01	<0.01
CK2	<0.01	<0.01	<0.01	<0.01	<0.01
CDK	<0.01	<0.01	<0.01	<0.01	<0.01
GSK	<0.01	<0.01	<0.01	<0.01	<0.01
PKA	<0.01	<0.01	<0.01	<0.01	<0.01

Table A.1 Statistical tests for the phylogenetic signal of 6 different kinase *Families* with respect to a species phylogeny of 48 eukartoic species. P-values were less than 0.01 for all *Families* tested and for all statistical tests applied.