# The structural and functional identity of the protein kinase superfamily

James David Randall Knight

Thesis submitted to the Faculty of Graduate and
Postdoctoral Studies in partial fulfillment of the requirements
for the Ph.D. degree in Cellular and Molecular Medicine

Department of Cellular and Molecular Medicine
Faculty of Medicine
University of Ottawa

## AUTHORIZATION

**ABSTRACT**

The human protein kinase superfamily consists of over 500 members that individually control specific aspects of cell behavior and collectively control the complete range of cellular processes. That such a large group of proteins is able to uniquely diversify and establish individual identities while retaining common enzymatic function and significant sequence/structural conservation is remarkable. The means by which this is achieved is poorly understood, and we have begun to examine the issue by performing a comparative analysis of the catalytic domain of protein kinases. A novel approach for protein structural alignment has revealed a high degree of similarity found across the kinase superfamily, with variability confined largely to a single region thought to be involved in substrate binding. The similarity detected is not limited to amino acids, but includes a group of conserved water molecules that play important structural roles in stabilizing critical residues and the fold of the kinase domain. The development of a novel technique for identifying kinase substrates on a large scale directly from cell lysate has revealed that substrate specificity is not what discriminates the closely related p38α and β mitogen-activated protein kinases. Instead cellular localization appears to be their distinguishing characteristic, at least during myoblast differentiation. Together these results highlight the extent of conservation, as well as the minimal variability, that is found in the catalytic domain of all protein kinase superfamily members, and that while distantly related kinases may be distinguished by substrate specificity, closely related kinases are likely to be distinguished by other factors. Although these results focus on representative members of the kinase superfamily, they give insight as to how all protein kinases likely diversified and established unique non-redundant identities. In addition, the novel techniques

developed and presented here for structural alignment and substrate discovery offer new tools for studying molecular biology and cell signaling.

**TABLE OF CONTENTS**

## LIST OF TABLES

### CHAPTER 2: CONSERVATION, VARIABILITY AND THE MODELING OF ACTIVE PROTEIN KINASES



### CHAPTER 3: THE ROLE OF CONSERVED WATER MOLECULES IN THE CATALYTIC DOMAIN OF PROTEIN KINASES



### CHAPTER 4: A NOVEL WHOLE-CELL LYSATE KINASE ASSAY IDENTIFIES SUBSTRATES OF THE P38 MAPK IN DIFFERENTIATION MYOBLASTS

## LIST OF FIGURES

### CHAPTER 1: GENERAL INTRODUCTION



### CHAPTER 2: CONSERVATION, VARIABILITY AND THE MODELING OF ACTIVE PROTEIN KINASES



### CHAPTER 3: THE ROLE OF CONSERVED WATER MOLECULES IN THE CATALYTIC DOMAIN OF PROTEIN KINASES



### CHAPTER 4: A NOVEL WHOLE-CELL LYSATE KINASE ASSAY IDENTIFIES SUBSTRATES OF THE P38 MAPK IN DIFFERENTIATION MYOBLASTS

## CHAPTER 5: GENERAL DISCUSSION



## APPENDIX

## LIST OF ABBREVIATIONS

ACK1: activated CDC42 kinase 1
ACN: acetonitrile
ACSW: active-site conserved waters
AGC: protein kinase group named after protein kinases A, G and C
ADP: adenosine diphosphate
Akt: RAC-α serine/threonine-protein kinase
Akt2: RAC-β serine/threonine-protein kinase
ANP: phosphoaminophosphonic acid-adenylate ester
AS: analog-sensitive
ASA: accessible surface area
ATP: adenosine triphosphate
c-Abl: Abelson murine leukemia viral oncogene cellular homolog
c-Raf: murine sarcoma 3611 viral oncogene cellular homolog
c-Src: Rous sarcoma viral oncogene cellular homolog
CAMK: calmodulin/calcium-regulated protein kinases
CARD: caspase-recruiting domain
CDC: cell-division cycle
CDK: cyclin-dependent or cell division protein kinase
ChaK: transient receptor potential channel kinase
CHK: CSK-homologous kinase
CK1: casein kinase or cell kinase I
CK2: casein kinase II subunit α
CLK: CDC-like kinase
CMGC: protein kinase group named after CDK, MAPK, GSK3 and CLK
CNS: central nervous system
COX IV: cytochrome c oxidase IV
CSK: c-Src kinase
DAPK: death-associated protein kinase
DMSO: dimethyl sulfoxide
DYRK1B: dual specificity tyrosine-phosphorylation-regulated kinase 1B
EGF: epidermal growth factor
ERK: extracellular signal-regulated kinases
FAK: focal-adhesion kinase
FBS: fetal bovine serum
FDA: Food and Drug Administration
FGF: fibroblast growth factor
FSBA: 5′-4-Fluorosulfonylbenzoyladenosine
GAPDH: glyceraldehyde 3-phosphate dehydrogenase
GO: gene ontology
GPI: glucose phosphate isomerase
GRP78: glucose-regulated protein of 78kDa
GSK3: glycogen synthase kinase-3
GST: glutathione S-transferase
HSP27: heat-shock protein 27

IEF: isoelectric focusing
ILK: integrin-linked kinase
IPG: immobilize pH gradient
IRK: insulin receptor tyrosine kinase
iTRAQ: isobaric tags for relative and absolute quantitation
JAK: just another or Janus kinase
JNK: c-Jun N-terminal kinase
KESTREL: kinase substrate tracking and elucidation
LC: liquid chromatography
LKB1: liver kinase B1
MAPK: mitogen-activated protein kinase
MARKCS: myristoylated alanine-rich C-kinase substrate
MD: molecular dynamics
MEK: MAP or ERK kinase
MHC: major histocompatibility complex
MKK: mitogen-activated protein kinase kinase
MS: mass spectrometry
MuSK: muscle-specific receptor tyrosine kinase
MyHC: myosin heavy chain
NCAM: neural cell adhesion molecule
NF-κB: nuclear factor κB
NMR: nuclear magnetic resonance
P-p38: phospho p38
PDB: protein data bank
PDGF: platelet-derived growth factor receptor
PDK1: 3-phosphoinositide dependent protein kinase-1
PhK: phosphorylase kinase
Pim-1: proto-oncogene serine/threonine-protein kinase Pim-1
PKA: protein kinase A
PknB: probable serine/threonine-protein kinase PknB
PME: particle mesh Ewald
PP1: protein phosphatase 1
PYK2: proline-rich tyrosine kinase 2
p38: p38 mitogen-activated protein kinase
Rio2: right open 2 kinase
RIP2: receptor-interacting serine/threonine-protein kinase 2
RMSD: root mean square deviation
Rps27: 40s ribosomal protein S27
SAKS1: SAPK substrate protein 1
SAPK: stress-activated protein kinase
SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis
SFK: Src-family kinase
SH2: Src homology 2
Sky1P: SR protein kinase
STE: sterile kinase; protein kinase group named from homology to this yeast kinase
STRADα: STE20-related kinase adapter protein α

TAO2: thousand and one amino-acid protein 2
TK: tyrosine kinase
TKL: tyrosine kinase-like
TrkA: high-affinity nerve growth factor receptor
VEGF: vascular endothelial growth factor
WNK: with no lysine kinase

**ACKNOWLEDGMENTS**

First and foremost I would like to thank my supervisor Dr. Rashmi Kothary. His mentorship during my graduate studies has been critical to my development and I do not believe I could have been as successful to date without his guidance. I consider myself very lucky to have had the opportunity to study in his laboratory and could not imagine a better place to have pursued my PhD. I hope that when I have own lab I can provide my trainees with the same quality of direction that he has provided to me, and be as willing and confident as he is to let students pursue their interests.

I would like to thank the other members of the lab whom I had the benefit of training, collaborating and brainstorming with: Carrie Anderson, Ariane Beauvais, Kunal Bhanot, Melissa Bowerman, Justin Boyer, Yves De Repentigny, Andrew Ferrier, Dr. Karen Lee, Dr. Hong Liu, John-Paul Michlaski, Dr. Lyndsay Murray, Ryan O'Meara, Bruno Pinheiro, Dr. Scott Ryan, Tadasu Sato, Dr. Dina Shafey and Dr. Kevin Young.

I would like to thank my committee members Dr. Valerie Wallace, Dr. Lynn Megeney, Dr. Robert Screaton and Dr. Miguel Andrade for guidance and advice throughout my PhD.

Lastly I would like to that the Canadian Institutes of Health Research and the Multiple Sclerosis Society of Canada for generously funding my graduate studies.

# CHAPTER 1: General Introduction

**The function of a protein kinase**

Cell signaling and cellular behavior are precisely controlled by members of the protein kinase superfamily. Whether it be proliferation, apoptosis, mitosis, senescence, cell adhesion, migration, or the multitude of other cellular processes, the extrinsic and intrinsic signals that initiate these behaviors are mediated and propagated by protein kinases. Fundamentally, the unifying behavior of all protein kinases is enzymatic activity that catalyzes the covalent attachment of phosphate to target proteins. A protein kinase will bind ATP and a substrate protein (in an ordered or randomly ordered manner[1], deprotonate a target serine, threonine or tyrosine residue on the substrate, transfer the terminal phosphate from ATP to the acceptor residue, and then disassociate from the substrate and ADP. In eukaryotes serine, threonine and tyrosine residues are those generally targeted for phosphorylation, although arginine, aspartate, histidine and lysine can also be found in phosphorylated forms[2], particular in prokaryotes[3], and the phosphorylation of these residues by protein kinases may be physiologically relevant in eukaryotes as well[2, 4-7].

Once a phosphate group has been attached to the substrate protein it can alter its function in a few different ways. Phosphate adds a net negative charge of two to serine, threonine and tyrosine residues, creating a drastically different local environment for the affected region of the substrate. Phosphorylation can trigger a conformational change in the substrate from new positive or repulsive interactions between other substrate residues and the phosphorylated amino acid. It may not affect residue interactions on the substrate itself, but instead induce an interaction between the substrate and residues on another protein. Alternatively the phosphorylation may have no direct physiological consequence

for the substrate involved, either being completely irrelevant, or indirectly affecting cell behavior, for example by acting as a phosphate sink to soak up excess kinase activity[8]. When phosphorylation has a direct effect on the substrate the resulting conformational change or induced protein-protein interactions that result can cause an increase or decrease in cellular enzymatic activity (if the substrate is an enzyme or an allosteric regulator of an enzyme) or it can alter the substrate's cellular localization, structural stability or rate of degradation, thereby affecting any processes the substrate is involved in. From the simple covalent attachment of a phosphate group to target proteins it can easily be appreciated how the protein kinase can have such a major impact on cellular behavior.

**The protein kinase superfamily**

There are an estimated 518 human protein kinases and 540 in mouse[9, 10], whose genes occupy ~2% of the genome. They can be divided into seven major groups that display sequence, and in some cases functional, similarity (**Figure 1.1**)[9, 10], and these groups can be further subdivided into families and subfamilies. The AGC group is named for its three principal members, the protein kinases A, G and C. This group also contains the well-known kinase Akt. The CAMK group takes its name from the calmodulin/calcium-regulated protein kinases and includes the first ever-discovered protein kinase, phosphorylase kinase (PhK). Despite the name, this group contains kinases that function independently of calcium. The CMGC group includes many popular and well-studied kinases and its name is taken from the initials of the principal members: cyclin-dependent kinase (CDK), mitogen-activated protein kinase (MAPK), glycogen synthase kinase-3 (GSK3) and CDC-like kinase (CLK). The CK1 group is a relatively small collection of

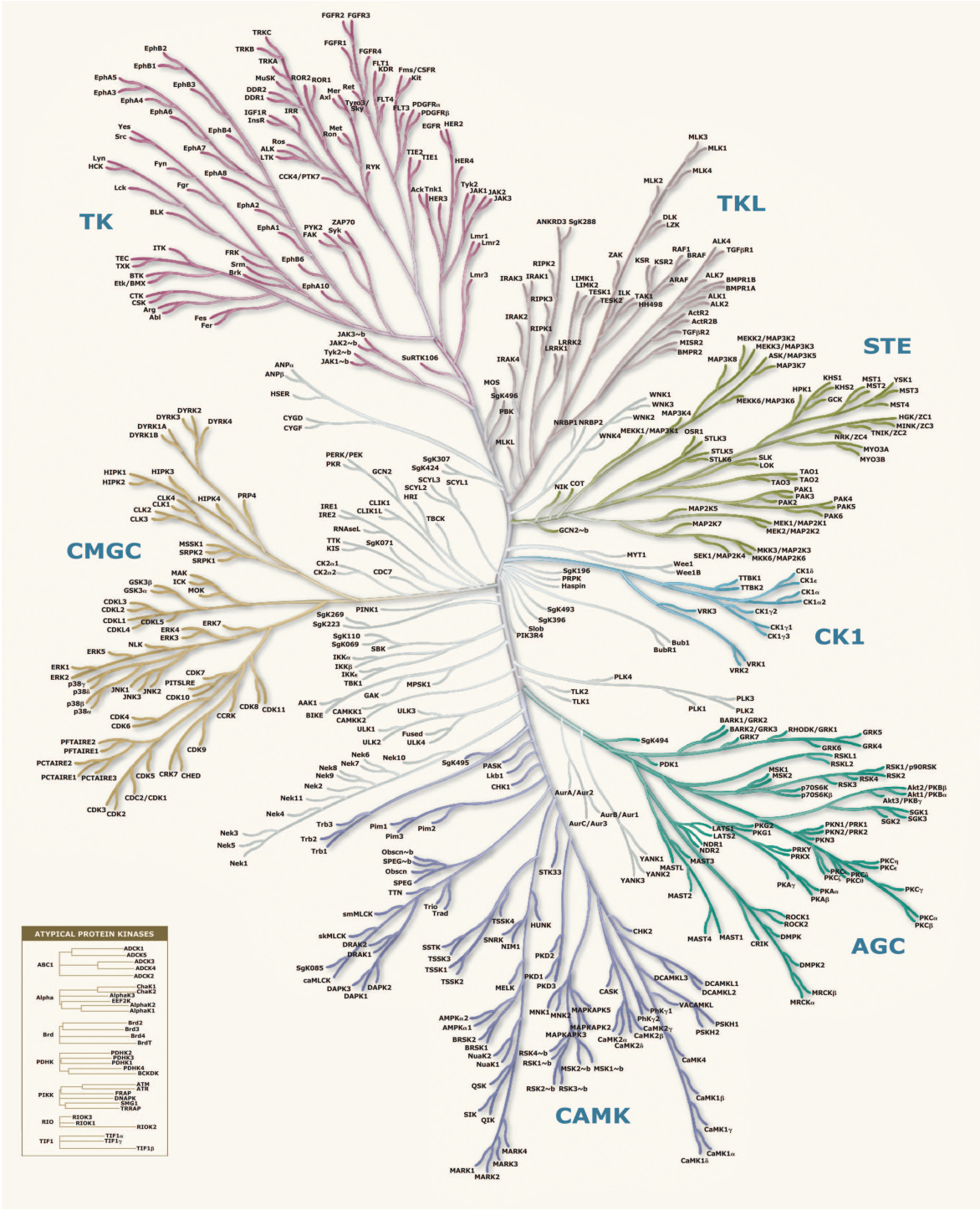



**Figure 1.1**. Phylogeny of the human protein kinase superfamily. Modified and reproduced from Manning, et al. The protein kinase complement of the human genome. *Science*. 2002; 298: 1912-1934. © American Association for the Advancenment of Science.

kinases related to casein kinase (now known as cell kinase). The STE group contains the homologs of the yeast sterile (STE) kinases, many of whose members are involved in initiation of mitogen-activated protein kinase (MAPK) signaling cascades. The TK group contains only tyrosine kinases, both receptor and non-receptor forms, such as the well-known receptors for fibroblast growth factor (FGF), epidermal growth factor (EGF) and vascular endothelial growth factor (VEGF), and cytoplasmic kinases such as Rous sarcoma viral oncogene cellular homolog (c-Src) and Janus kinase (JAK). The final group of related kinases is the TKL or tyrosine kinase-like group whose members are similar to tyrosine kinases but phosphorylate serine/threonine residues. Its most prominent member is the proto-oncogene Raf. These are the seven "related" groups of protein kinases, but there is an eighth group of kinases that are collected together because of their lack of similarity to typical protein kinases. This group is appropriately called the atypical group and its family members bear no sequence similarity to each other or to the previously described seven groups. The members of the atypical group, however, possess phosphotransfer capability.

All protein kinases may have evolved from a single ancestral form, based on the presence of sequence similarity between typical kinases, and from the high structural similarity that all protein kinases, whether typical or atypical, display towards each other[11, 12]. However, not all protein kinases have retained phosphotransfer capability during evolution. An estimated 48 of the 518 human protein kinases, and the same 48 in mice, lack one or more critical catalytic residues at the sequence level and are therefore deemed to be inactive, or as it is more properly termed: pseudokinases[9, 10, 13]. These pseudokinases represent ~10% of the kinase superfamily. However, some of these do

have activity[14-16], and at the same time some kinases with a full complement of catalytic residues may lack activity, making it difficult to precisely determine the number of pseudokinases that exist. Pseudokinases are not just vestigial proteins with no importance, but instead they fulfill other functions, such as regulating the activity of other kinases and serving as structural or signaling scaffolds[13, 17]. Several of these non-catalytic kinases are composed of little more than a kinase domain, highlighting that this domain can be utilized for functions other than phosphotransfer.

**Other functions of protein kinases**

Not everything a kinase does involves catalytic activity and the transfer of phosphate to a target protein. In the case of pseudokinases that lack activity altogether, a kinase has become entirely adapted to fulfill another function. One of the best examples of this is the integrin-linked kinase (ILK). After extensive debate, ILK has been definitively shown to be a *bona fide* pseudokinase that functions as a structural and signaling scaffold at focal adhesion sites[18-20]. ILK consists of four or possibly five ankyrin repeats at its N-terminus and a kinase domain at its C-terminus, but the bulk of its known role as a structural scaffold is performed by the kinase domain[21-23]. This domain has lost the ability to catalyze phosphotransfer and been adapted into a pure interaction domain that binds β1 integrin, and actin-binding proteins such as α-parvin and paxillin. ILK still binds to ATP but rather than using it for phosphorylation, it appears this ligand is used to stabilize the structure of ILK[18]. What the structure and function of ILK demonstrate is how the kinase domain can be used for non-catalytic functions, and in extreme cases be entirely co-opted for functions other than catalysis.

The STE20-related kinase adapter protein α (STRADα) is another well-known pseudokinase co-opted for purposes other than catalysis. STRADα uses its kinase-domain to bind to and activate liver kinase B1 (LKB1)[17]. The manner of binding between STRADα and LKB1 is reminiscent of how any kinase would bind to and activate a protein substrate, except phosphorylation is absent. STRADα adopts a characteristic active kinase conformation during interaction, it requires ATP to activate LKB1, and it uses regions typically involved in substrate binding to interact with this kinase, but rather than activation occurring through phosphorylation, STRADα activates LKB1 through an allosteric mechanism[24, 25].

These types of non-catalytic functions are not limited to pseudokinases, although they are less well studied in *bona fide* kinases. The receptor-interacting serine/threonine-protein kinase 2 (RIP2) is composed of a catalytically active N-terminal kinase domain, followed by an intermediate domain, and a C-terminal caspase-recruiting domain (CARD). RIP2 is a mediator of signaling through both the tumor necrosis factor receptor family and the toll-like receptor family, and is involved in the transcriptional activation of NF-κB, and the activation of the c-Jun N-terminal kinase (JNK)[26-28]. Full-length RIP2, including its kinase domain, is absolutely necessary for the activation of these proteins, but RIP2 kinase activity is not. Instead evidence suggests that RIP2 acts as an adaptor at receptor complexes, connecting and recruiting proteins involved in the activation of downstream components[26-28].

Another kinase that has a function independent of its catalytic activity is the CSK-homologous kinase (CHK). CHK can regulate the activity of the Src-family of kinases (SFK) through both a phosphorylation-dependent and phosphorylation-independent

mechanism[29, 30]. The inactivation of SFKs through the phosphorylation-independent mechanism involves the binding of CHK to SFKs and allosteric inhibition of SFK activity. How this occurs is unknown, but the interaction involves the kinase domain of CHK[30]. Similarly, the cyclin-dependent kinase 5 (CDK5), which contains only a kinase domain, can suppress the neuronal cell cycle through a mechanism that does not require activity[31]. This suppression may involve the interaction with and sequestration of cell-cycle proteins by CDK5, but the details are not currently known[32]. What these studies on RIP2, CHK and CDK5 highlight is that a catalytically active kinase domain can serve functions other than phosphotransfer in a manner similar to what has been shown for pseudokinases.

**An excess of serine/threonine protein kinases?**

The number of protein kinases alone suggests that phosphorylation is a major regulatory mechanism for controlling cell behavior, and this is supported by large-scale phosphoproteome analysis that indicates ~30% of proteins may be phosphorylated at some time[33]. Phosphorylation, being a reversible process, is negatively regulated by protein phosphatases that catalyze the removal of phosphate from target proteins. However, an apparent discrepancy arises when the number of serine/threonine (S/T) phosphatases is considered against the number of S/T kinases. While there are 90 protein tyrosine kinases and a roughly equal number of 107 protein tyrosine phosphatases[34], there are 428 S/T protein kinases but only ~30 S/T protein phosphatases[35]. Why is the number of S/T kinases more than ten times that of their phosphatase counterparts? Initially this discrepancy may seem to suggest that there is extensive redundancy in the S/T group of kinases. There may be some redundancy in this group, as certain members appear to have

overlapping and compensatory functions, in, for example, members of the MAPK family[36], but even in this closely related family isoform specific functions have been identified. The evolutionary conservation of individual S/T kinases between species suggests that redundancy is unlikely, as it would often be accompanied by extensive sequence variation between orthologous members. At the same it is highly doubtful that redundancy would develop to such a high degree exclusively in S/T kinases.

Rather, the reason for the vast difference in the number of S/T protein kinases and phosphatases is due to the different methods S/T phosphatases use for substrate recognition. Some S/T phosphatases exist as one part of a holoenzyme complex that consists of additional and variable regulatory subunits that target the catalytic phosphatase core to distinct substrates[35, 37]. Protein phosphatase 1 (PP1) has 100 or more different regulatory units[35, 38], making the number of potential PP1-containing holoenzymes very large. This general schema of core plus regulatory unit(s) applies to several S/T phosphatases, making the total number of phosphatase-containing holoenzymes likely more on par with the number of S/T kinases. What is interesting is that while some S/T protein phosphatases developed distinct subunits for catalysis and substrate recognition, the S/T protein kinase incorporated both into the same amino acid sequence. The ability for the protein kinase, whether S/T or tyrosine, to combine a conserved catalytic mechanism with a divergent substrate recognition region(s) highlights its perfection as a signaling molecule, one that could be amplified and modified from a single ancestral precursor into a massive superfamily of unique members that specifically control particular aspects of cell behavior.

**The identity of a protein kinase**

Even given that there may be a small degree of redundancy within the protein kinase superfamily, and that certain members can functionally compensate for the loss of another, there are still hundreds of evolutionary-conserved kinases performing unique and non-overlapping functions. With so many kinases, a major question that arises is what gives each kinase its identity, or what distinguishes it from another, particularly when there is high sequence similarity?

Two closely related kinases that have the theoretical potential to functionally compensate for one another may be unable to do so because of differences in tissue expression. In these cases, although the kinases in concern may be virtually identical, there are extra-protein factors that distinguish them, for example DNA regulatory elements. Kinases that are closely related in terms of amino acid or DNA coding sequence are not necessarily closely related in terms of expression pattern. When a gene duplication event occurs, although the coded protein itself may not change much or at all, the regulatory regions of each gene are now subject to different constraints than were present with a single copy, ultimately allowing for the paralogues to take on drastically different expression profiles. This is quite apparent in the protein kinase superfamily. An analysis of 459 human kinases[39] revealed that when kinases are clustered based on tissue expression there is little similarity between the resulting classification scheme and the traditional sequence-based phylogeny presented in **Figure 1.1**. Thus, very closely related and potentially redundant kinases can develop unique tissue expression-based identities that prevent *in vivo* compensation.

Two highly similar kinases expressed within the same tissue or cell type, and even at the same time, can take on unique identities from very small differences in amino acid sequence. Differences in sequence can translate into drastic differences in localization, resulting from signal peptide or protein-protein interaction motifs present on one kinase but not another. These motifs can cause gross differences in localization, such as nuclear or cytoplasmic, or they can cause fine differences in localization, for example localizing to a particular membrane receptor, such that two kinases present within the same cellular compartment can have very different local environments in terms of the proteins and other molecules available to associate with. Differences in sequence may also affect the way a kinase becomes activated/inactivated, allowing for the functional capabilities of two similar kinases to be differentially controlled through their manner of activation. Although the *in vitro* capabilities of close evolutionary related kinases might be identical, their *in vivo* capabilities are quite different because of the way such signal, interaction and activation motifs affect behavior in the complex environment of the cell.

In many cases regions of sequence that affect localization, protein-protein interactions or activation are not small motifs but rather distinct domains. Two-hundred and fifty-eight of the 518 human protein kinases contain additional domains[10], and clustering based on the non-catalytic domain composition of protein kinases produces similar groupings as does classification based on the amino acid sequence of the kinase domain[40]. This demonstrates that multi-domain protein kinases are evolving as whole functional units and not simply discreet domains that can be studied in isolation from each other. A major function of the non-catalytic domains found on protein kinases is to mediate interactions with other proteins or with other biomolecules. A typical example is

the Src homology 2 (SH2) domain that recognizes and docks with particular phosphotyrosine-containing motifs, and is found on many receptor and cytoplasmic tyrosine kinases[41]. Non-catalytic domains, such as the SH2 domain, act to contextualize the function of a kinase. By affecting protein-protein interactions, localization and activation, domains can compartmentalize the function of a kinase and give it a very unique and specific identity as part of an organized signaling pathway.

Differences in expression, localization and protein-protein interactions may be considered contextual differences that exist within a multicellular organism, the environment of the cell or complex protein mixture, but not in isolation from these contexts. These types of differences are extra-catalytic, meaning that they are independent of the catalytic activity of a kinase and serve to restrict the activity of the kinase domain to a particular context. The catalytic activity of a kinase, or rather the substrates that activity is directed towards, is the traditional defining feature that is used to give a kinase its identity: what can it phosphorylate? The protein kinase domain contains a highly variable region known as the substrate-binding groove that is believed to direct substrate specificity, although it is poorly studied and little is known about its function[42, 43]. It has been termed the substrate-binding groove from co-crystal structures, because peptide substrates are seen bound in this region[44, 45]. Several kinases, and maybe even a majority, contain additional regions outside of the substrate-binding groove that also mediate interactions with substrates[46]. Variability in the substrate-binding grove or other regions would theoretically allow different kinases to bind different substrates, giving them their catalytic identity and making the substrate-binding groove the defining feature for any kinase domain.

In some cases a protein kinase can phosphorylate a hundred or more substrates, for example PKA and Akt[47, 48], while others appear highly specific, such as the MAP or ERK kinase (MEK) isoforms and certain yeast kinases[49, 50]. Identifying the substrates a kinase can phosphorylate is critical for understanding the function that each kinase plays, since through this mechanism it has the potential to influence the behavior of so many other proteins. Identifying substrates is no easy task as the number of substrates a kinase may have could range from one on into the thousands. *In vitro* estimates for yeast kinases are available from a large-scale screen undertaken by Ptacek et al.[49] In this study 87 kinases were tested for activity against 4400 yeast proteins on a chip-based assay. The mean number of substrates per kinase was 47, although numbers ranged from one to 256. If we assume that these numbers are proportional to genome size then for the human genome, which has roughly six times as many kinases as yeast[10] and is estimated to have the same proportionally higher number of genes[51], the mean number of substrates per human kinase could be around 300 and range from one to 1500. Most kinases have nowhere near this many identified substrates, meaning that what is known about them is severely lacking and the resulting identities they have been given based on substrate profiles is very partial.

**Identifying substrates for a kinase**

One of the most well studied mammalian kinases is PKA. It has approximately 100 *in vivo*[48] and 200 *in vitro* substrates although exact numbers are difficult to determine as web databases are not up-to-date and include substrates on a somewhat arbitrary basis. Over the ten years spanning 1991-2000, 8 *in vivo* substrates were found on average per year for PKA[48]. If such a rate continued it could take decades to find all substrates for this

kinase, highlighting the need for significant advancements in substrate-finding techniques.

There are several strategies available for identifying protein kinase substrates, which can be broadly classified as motif-based, *in vitro* or *in vivo*. Motif-based approaches take many forms, but all rely on identifying substrates that contain a preferred consensus phosphorylation motif. If this is not known, it can be determined by screening a kinase against a library of peptides to identify a candidate sequence. Once a motif is known, *in silico* approaches can be used to search for proteins containing that motif; theoretical substrates that then require validation. Alternatively, a phosphoantibody raised against the consensus motif can be generated and used to screen cell or tissue lysate for proteins containing that epitope, which are again theoretical substrates that require validation. The drawback of motif-based approaches is that they generally produce hypothetical results that may not be legitimate for a variety of reasons, and more importantly they are very biased. Kinases will often have a preferred consensus motif they target for phosphorylation, but these motifs are usually not an absolute requirement. As more data are generated on the mechanisms of substrate recognition, it seems clear that most kinases recognize substrate-docking motifs distinct from the residues around the site of phosphorylation for target recognition. Motif-based strategies over-simply the mechanisms of substrate recognition, and therefore limit substrate identifications to a potentially small subset of actual substrates.

*In vitro* approaches are all based on the traditional *in vitro* kinase assay that involves incubating a purified kinase with purified substrate in the presence of $^{32}P_{\gamma}$- or $^{33}P_{\gamma}$-ATP, and determining if the substrate has been radioactively labeled by, for

example, gel electrophoresis. Although these techniques allow you to identify direct substrates, they do not give information about *in vivo* relevance and thus further validation is required. *In vivo* approaches rely on detecting a difference in phosphorylation between control and inhibitor treated or knock-down cells/tissue (with the inhibitor or knockdown directed at the kinase of interest), with any decrease in phosphorylation attributed to the absence of a specific kinase's activity. However, if a phosphorylation has decreased it is not necessarily direct and could simply be something downstream in the same pathway or an indirect effect from changes in cell behavior due to the absence of a specific kinase. *In vivo* approaches, which can be done on a large cell-wide scale to identify global changes in phosphorylation as a result of kinase inhibition, allow for associations between kinases and phosphorylations to be derived, but not direct associations. An *in vitro* and *in vivo* approach can be coupled to identify direct substrates that are associated/regulated by a kinase *in vivo*, and the drawbacks of each will be compensated for by the other. This has never been done before, but the combination offers an ideal strategy for substrate identification. A major reason why it has not been done is that although *in vivo* approaches for kinase-substrate association are currently easy to use and work well on a large-scale, *in vitro* approaches either work poorly and/or are laborious and thus are rarely used. Developing a simple to use and quick approach for *in vitro* substrate identification that can be used on a large-scale to identify hundreds of direct substrates simultaneously is therefore a necessary step before sophisticated approaches for substrate identification can be employed.

One of the most widely used techniques for *in vitro* substrate finding was developed by Knebel et al.[52] and called the KESTREL (kinase substrate tracking and

elucidation) method. In this approach a kinase of interest is added into cell lysate with a kinase assay buffer containing $^{32}P_{\gamma}$-ATP, and the sample then electrophoresed and imaged. By running parallel samples without an added kinase or with closely related kinases it is possible to identify putative targets of the kinase of interest as bands where the radioactive signal is specifically and significantly up regulated. The proteins in these bands can be identified through scale-up and fractionation strategies combined with mass spectrometry. The problem of background signal is an obvious issue with this method as cell lysate will already contain a number of active kinases and is partly responsible for its inability to identify more than a few substrates at best[52-63]. The authors also state a requirement of 500 mg of cell lysate for fractionation, purification and identification of substrates[64], which limits the number of cell types to which this method can be easily applied.

Troiani et al.[65] modified the KESTREL method to address drawbacks of the original approach. Background activity from endogenous kinases was eliminated by heat-inactivating the lysate prior to performing the assay, which, although successful, will result in partial denaturation of proteins thereby increasing the likelihood of false positive and negative substrate identifications. Identification of putative substrates was achieved through 2D gel-electrophoresis and mass spectrometry, requiring only 200 µg of cell lysate. Despite an autoradiogram showing dozens of potential substrates, the authors were only able to identify two new substrates for their kinase of interest, Aurora-A. They fail to discuss or even mention this discrepancy, which, however, is likely due to phosphorylated proteins being present at quantities too low for gel staining and mass spectrometry to detect.

A large-scale screening method that uses chip technology was described in 2005 that allows for testing a kinase for activity against thousands of proteins simultaneously[49]. Ptacek et al. made protein chips containing 4400 yeast GST-fusion proteins, and these chips were used in a typical kinase radiolabeling assay to test for phosphorylation. This approach allows for the identification of many substrates in one application but has several drawbacks. There is a high rate of false positives for an unknown reason, and presumably a high rate of false negatives from the absence of protein and biochemical interactions, and substrate discovery is limited to the proteins included on the chip. Of course, custom chips could be made but the time and expense of doing so negates the benefits of this approach. This method is similar in principal to one developed a few years earlier that screened proteins from a cDNA library cloned into a phage expression vector[66, 67]. Rather than purified protein, bacterial lysates were immobilized on nitrocellulose and tested for phosphorylation by a kinase of interest. cDNA from positive hits was then sequenced and the target protein identified. The drawbacks of this approach are similar to those for the chip-based technique.

The best method developed to date has come from Kevin Shokat's laboratory. His group began developing what they term analog-sensitive (AS) kinases[68, 69]. These are kinases that have been modified to accept a bulky ATP analog that cannot be utilized by wild-type kinases. Addition of a mutated kinase to cell lysate along with a bulky radioactive ATP analog results in substrate specific labeling that can be visualized following electrophoresis. The first large-scale screens to which this approach was applied used yeast strains overexpressing GST-fusion proteins that served as candidate and easily identifiable substrates[70, 71]. More recently AS-kinases have been used in this

laboratory with a further modification to the ATP analog that causes substrates to be labeled with a unique non-radioactive tag[72]. This new ATP analog contains a sulfur atom in place of an oxygen on the terminal phosphate. An AS-kinase can use this analog to specifically label its substrates in whole-cell lysate and following tryptic digestion, thio-phosphate labeled peptides from substrate proteins can be purified and identified through mass spectrometry. In its first application using whole-cell lysate this technique generated 70+ substrate identifications for CDK1. The drawbacks to this approach are that not all kinases can thiophosphorylate substrates, not all can be modified to accept a bulky ATP analog, the modification process is time-consuming in itself, and the approach used to purify thio-phosphate labeled peptides results in the loss of cysteine-containing peptides (~25% of all peptides) and consequently the loss of the corresponding substrate identifications.

The methods described above are the best currently available and do yield results, but they all have drawbacks that prevent them from becoming widely used. A simple and quick technique that can use cell lysate, work with any kinase and identify numerous substrates is ideally desired and would allow for more comprehensive profiling of protein kinase targets.

**p38 MAPK**

The question of kinase uniqueness/identity and its determinants can be explored by examining closely related members of the superfamily. The p38 MAPK subfamily consists of four isoforms, α, β, γ and δ, which share a high degree of sequence similarity (**Figure 1.2**). All p38 isoforms consist primarily of a kinase domain, with a short ~25 amino acid N-terminal extension and an additional ~60-65 amino acids at the C-terminus

(**Figure 1.3**). p38α and β are the most closely related with an overall sequence identity of 73% (human)[73], while p38β and δ are the most divergent at 56% sequence identity[74]. The different isoforms have quite distinct mRNA expression patterns (**Figure 1.4**)[74]. p38α mRNA is found ubiquitously, p38β is found at higher levels in the CNS, p38δ shows broad expression across non-CNS tissues, while p38γ is restricted largely to skeletal muscle. While these differences might suggest tissue-specific roles for the β, γ and δ isoforms, knockout mice suggest otherwise. Mice that are null for any of these three isoforms are viable and healthy with no overt phenotype[75, 76]. These results suggest either that these isoforms do not play a critical and unique role, or that upon knockout there is functional compensation by another p38 member. However, studies on individual cell types have revealed unique roles for p38β in optimal bone development[77], p38γ in satellite cell maintenance and muscle regeneration[78], and p38δ in keratinocyte biology[79], which, although not critical for mouse viability, are certainly significant. On the other hand, there are no ambiguities surrounding the essential role of the p38α isoform. Mice null for this isoform die around embryonic day 11, or if p38α expression is maintained in the placenta then embryos can reach near term[80-82]. The lack of viability is consistent with cell-specific studies that describe critical roles for p38α in numerous processes and cell types[83, 84].

Despite widespread expression of the p38β and δ isoforms, the lack of any severe defect upon deletion of these genes suggests the possibility that another isoform may compensate for their loss, and the absence of a severe muscle phenotype on deletion of p38γ suggests the same in this tissue. p38α is ubiquitously expressed and may be a good candidate for compensation. This would indicate that p38α may have an overlapping

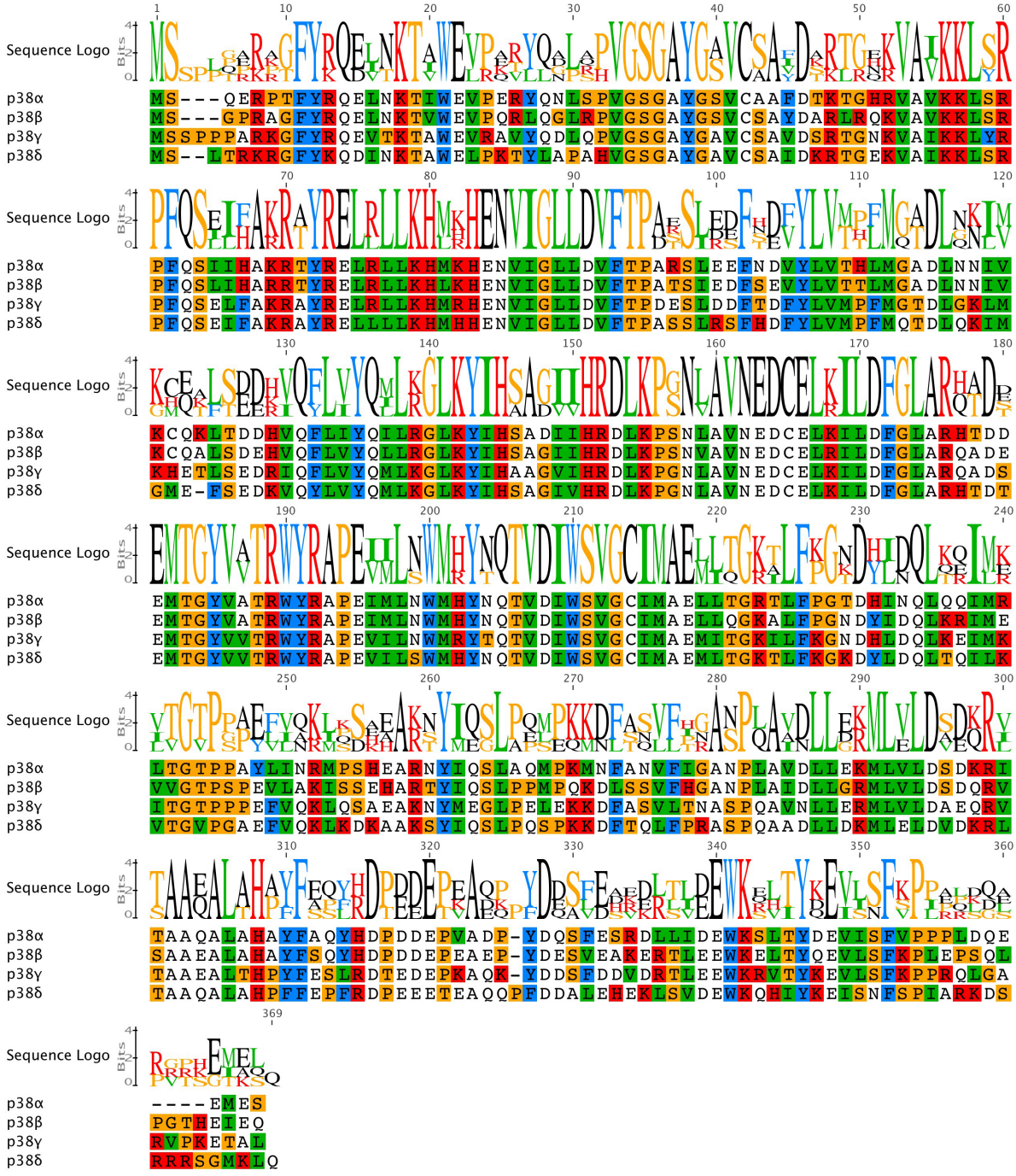



**Figure 1.2**. Sequence alignment of the p38 MAPK subfamily. The consensus sequence is shown as a sequence logo above the alignment.

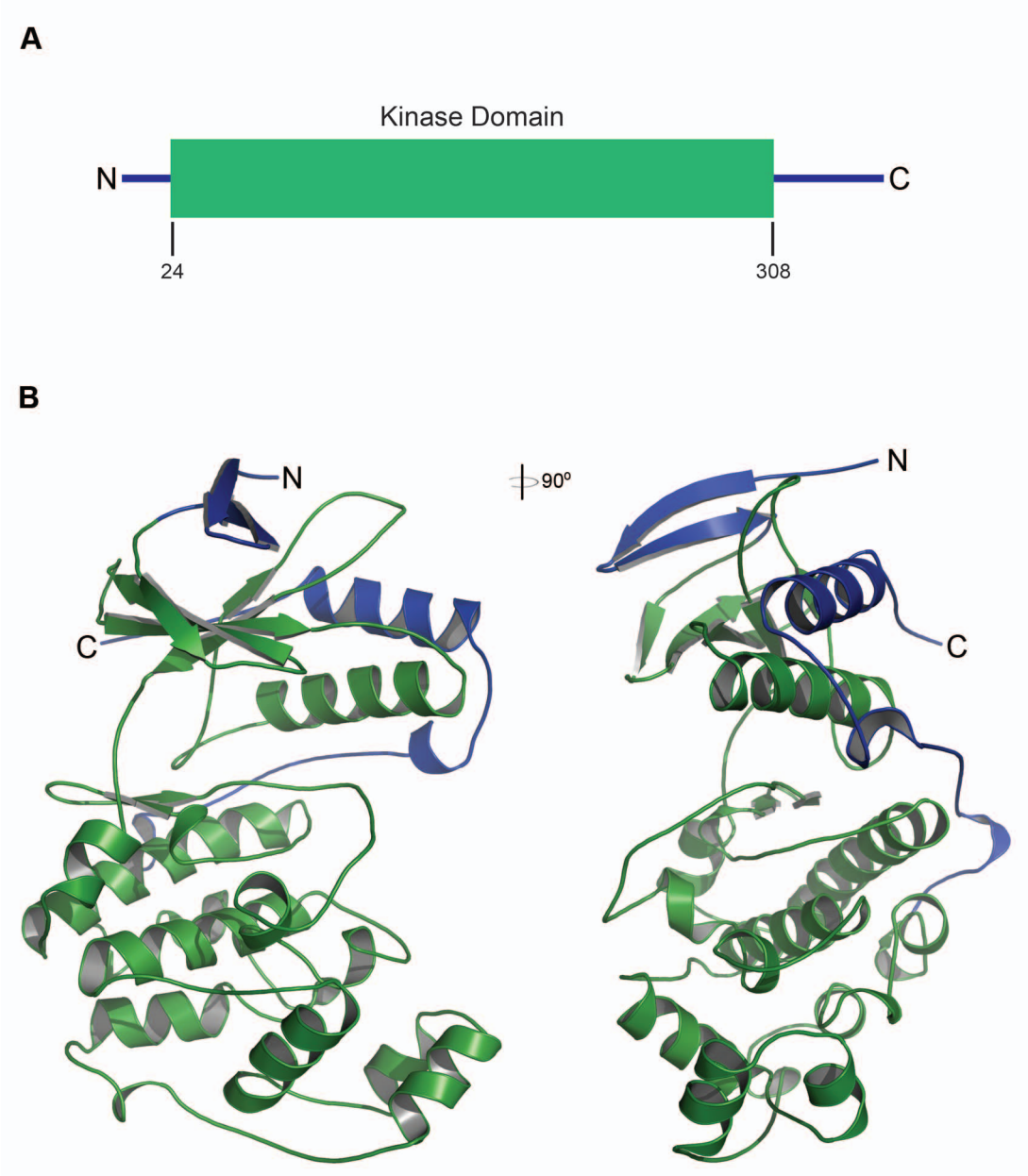



**Figure 1.3**. The structure of p38α. (**A**) A schematic of the domain composition of p38α. (**B**) The structure of p38α in an inactive conformation. The right panel is rotated 90° clockwise around the vertical axis. The kinase domain is shown in green and the N- and C-terminal regions in blue.

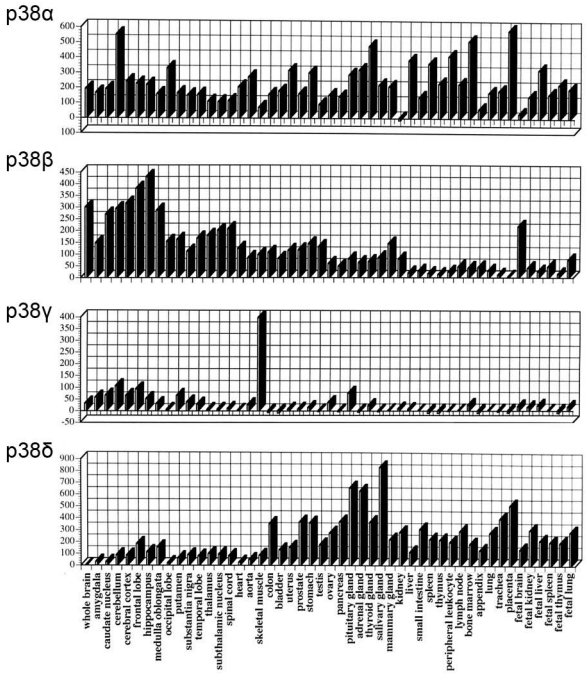



**Figure 1.4**. mRNA expression pattern of the human p38 isoforms. This figure was modified from and originally published in The Journal of Biological Chemistry. Wang, et al. Molecular Cloning and Characterization of a Novel p38 Mitogen-activated Protein Kinase. *The Journal of Biological Chemistry*. 1997; 272: 23668–23674. © American Society for Biochemistry and Molecular Biology.

localization pattern with the other isoforms, and that p38β, γ and δ have no specific substrates; if they did, p38α could not compensate for their loss. Conversely though, why is there no compensation when p38α is knocked out of the whole mouse or individual cell types in culture? There appear to be distinct mechanisms for preferentially activating p38α over the other isoforms[83, 84], meaning that if p38α were knocked-out there may be insufficient activation of p38β, γ or δ to compensate. p38α may also have a unique localization pattern, or of course, specific substrates that the other isoforms cannot phosphorylate. There are no known substrates exclusive to p38α versus β, although there do appear to be specificities between α and γ or δ[74, 85]. To what extent substrate specificity distinguishes the p38 isoforms is totally unknown, and what ultimate features functionally distinguish the p38 subfamily members are poorly understood.

Muscle development or adult regeneration serves as a good example of the differing roles the p38 isoforms are capable of playing. All isoforms are expressed in myoblasts, but while p38β, γ and δ can be knocked out without major consequence, differentiation fails to occur in p38α null myoblasts[80, 86, 87]. Much is known about the role that p38α plays initiating myogenic gene activation at the onset of differentiation[88], but why another isoform cannot fulfill this same role is a mystery. Not only can endogenous p38β and γ not compensate for the loss of α, but overexpression of these isoforms still does not allow for compensation[86, 89] (p38δ has not been examined). p38α must, therefore, have a particular feature and capability that distinguishes it from the other isoforms during myoblast differentiation, but what this may be is unknown. p38γ has been shown to become activated with both myoblast differentiation in culture and during adult regeneration[80, 86], suggesting at least one other p38 isoform is active in myoblasts,

and presumably the others as well. In the absence of any additional domains or signal peptides on any of the p38 isoforms, the most likely explanation for the absolute requirement of p38α is that it is capable of specifically phosphorylating critical myogenic proteins. Upon the loss of p38α these phosphorylations fail to occur and myoblasts do not differentiate. A comprehensive identification of p38 substrates would significantly aid in understanding the unique role that p38α plays in myoblasts, and potentially in the many other cell types that it is critical for as well.

**Objective and Aims**

Given the importance of the protein kinase superfamily to cell biology, understanding how the identity or uniqueness of a kinase is established is of absolute importance. At the same time, there is an urgent need for a simple, comprehensive approach for kinase substrate discovery so the major mechanism by which kinases control cell behavior can be systematically studied. The objective for this project has been to better characterize the features that determine the identity of a protein kinase domain, by examining the similarity and variability that exists at the structural level, and by performing a large-scale substrate screen for two closely related kinases (p38α and β) to determine the extent to which substrate specificity distinguishes close homologs. The following specific aims were created:

1. Determine the structural similarity and variability that is present in the protein kinase domain by examining representative members that span the entire superfamily.
2. Develop a simple approach for kinase substrate discovery that works directly with cell lysate and can be applied to any kinase.

3. Find substrates for the p38α and β isoforms in differentiating myoblasts to identify specific p38α targets and determine to what extent substrate specificity distinguishes these kinases.

The results generated with these aims will create a greater understanding of how protein kinases are able to accomplish unique functions with their catalytic domains, and this in turn will create greater knowledge of the mechanisms by which cell signaling is regulated, added understanding of the process of molecular evolution, and factors that need to be taken into consideration during the design of pharmaceuticals that target protein kinases. Developing the tools needed to meet these aims is in itself a major accomplishment. New approaches are needed for accurate comparison of proteins at the structural level, and having a simple, large-scale technique for kinase substrate discovery will be of great value to the many laboratories that study these proteins.

# CHAPTER 2: Conservation, Variability and the Modeling of Active Protein Kinases

**Conservation, Variability and the Modeling of Active Protein Kinases**

James D. R. Knight[†], Bin Qian[‡], David Baker[‡] and Rashmi Kothary[*]

[†*]Molecular Medicine Program, Ottawa Health Research Institute, Ottawa, Ontario, Canada

[†*]The University of Ottawa Centre for Neuromuscular Disease, Ottawa, Ontario, Canada

[†*]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada

[‡]Department of Biochemistry, University of Washington, Seattle, Washington, USA

[*]Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

**Published in *PLoS ONE*, 2007, 2(10):e982.**

**Author Contributions**

Conceived and designed the experiments: JK RK. Performed the experiments: JK BQ. Analyzed the data: JK BQ. Wrote the paper: JK. Other: Developed and implemented the structure prediction method described: DB BQ JK.

**Abstract**

The human proteome is rich with protein kinases, and this richness has made the kinase of crucial importance in initiating and maintaining cell behavior. Elucidating cell signaling networks and manipulating their components to understand and alter behavior require well designed inhibitors. These inhibitors are needed in culture to cause and study network perturbations, and the same compounds can be used as drugs to treat disease. Understanding the structural biology of protein kinases in detail, including their commonalities, differences and modes of substrate interaction, is necessary for designing high quality inhibitors that will be of true use for cell biology and disease therapy. To this end, we here report on a structural analysis of all available active-conformation protein kinases, discussing residue conservation, the novel features of such conservation, unique properties of atypical kinases and variability in the context of substrate binding. We also demonstrate how this information can be used for structure prediction. Our findings will be of use not only in understanding protein kinase function and evolution, but they highlight the flaws inherent in kinase drug design as commonly practiced and dictate an appropriate strategy for the sophisticated design of specific inhibitors for use in the laboratory and disease therapy.

## Introduction

Protein kinases are the most ubiquitous single family of signaling molecules in the cell, accounting for approximately 2% of the proteins encoded by the human genome[10]. The simple mechanism of attaching an ATP-derived phosphate to a protein involves kinases in every aspect of cell behavior, from apoptosis to survival, proliferation to differentiation, maturation etc. Protein kinases provide a unique opportunity for understanding proteins in general by presenting us with a seeming paradox: wide scale similarity of sequence and structure combined with a diversity of behavioral consequences to their activity. The vast majority of protein kinases have readily detectable sequence similarity, which translates into structure. But even those known protein kinases that show no significant algorithm-detectable similarity at the level of sequence are believed to have very typical structures, as is evidenced by specific examples[90, 91]. As they all have a shared function in transferring the terminal phosphate of ATP to another protein, similarity is understandable. Evidence to date also suggests a common catalytic mechanism (the possible exception may be the integrin-linked kinase[92]), whereby ATP and an active site divalent cation are bound in identical fashions and phosphotransfer is achieved by a shared set of amino acids. Studies in yeast[49, 71] have shown that kinases can be promiscuous, phosphorylating hundreds of proteins, but they also have clear specificities. How is this specificity attained by one family of highly similar proteins? This paradox suggests the perfection of the kinase as an enzyme: a region ideally suited for the common function of catalysis, with another region(s) uniquely modifiable to attain substrate specificity without altering fold, compromising ligand binding or the subsequent reaction mechanism. A thorough understanding of this

family of proteins would generate a tremendous knowledge base for discovering and predicting protein interactions, for designing highly specific and potent inhibitors, and, as a consequence of these facts, for understanding the cell and disease.

As protein kinases are the key players in cell signaling, aberrations in their activity have been directly correlated with numerous disease states (for example, breast cancer[93] and chronic myeloid leukemia[94]) and made them potential targets for drug design in many other diseases (for example, Crohn's[95] and cerebral vasospasm[96]). This has made the kinase the drug target of choice[97]. However, there is an inherent flaw in traditional kinase inhibitor design. Almost all inhibitors target the ATP binding pocket based on a simple principle: if ATP cannot be bound, phosphorylation cannot occur. Building a molecule that can occupy this pocket is relatively simple, but since the ATP binding pocket and the regions in its immediate vicinity are the areas of greatest conservation, building a *specific* inhibitor is impossible. The inherent multi-target nature of inhibitors has been demonstrated by Fabian *et al*.[98], where the twenty compounds tested had multi-target coverage with only 23% of the kinome screened. Other ATP binding proteins could very likely display affinities for these compounds as well, making these inhibitors not just multi-kinase but multi-enzyme. In the laboratory, how can the effect of treating cells with such inhibitors be dissected? And when used for disease, what non-intended effects may arise in the targeted cell type or others over the long term? In the hopes of producing specific inhibitors, what is needed is a new approach to kinase drug design, one which logically targets the region of greatest dissimilarity.

True dissimilarity can be known if similarity or conservation is understood in detail. For this, structure-based comparative approaches are needed to fully extract the

information hidden in the three-dimensional protein-structure space. Traditional structure-driven alignment studies concentrate on maximizing fold overlap, and for the highly-similar protein kinase family which has a largely conserved fold, this can be a useful approach. But it is not necessarily the correct one, particularly where inhibitor design is concerned. Due to the information available in a three-dimensional space, structures can be aligned in other ways, for example by using geometry independent of connectivity. Fold can be ignored and focus directed upon residues free from their covalent associations. The positioning of side-chains and those functional groups involved in enzymatic catalysis and protein interactions can be directly overlain for studying similarity and variability. This type of alignment, and not that of fold, is of greater relevance for understanding protein interactions and therefore in designing small molecules or peptides to act as inhibitors.

Understanding the similar/conserved and dissimilar/non-conserved aspects of protein kinases allows for effective drug design. In addition, conservational studies will aid especially in structure prediction. There are at least 518 known human protein kinases[10] and deriving crystal structures for them all would involve a great deal of time and effort. As all known protein kinases have similar structures, homology-driven approaches to structure prediction that incorporate knowledge of conservation should prove fertile. Having a reliable predicted structural kinome would be of great practical use.

In this paper we report on a structural analysis of and a modeling approach to active-conformation protein kinases. We describe the variability found between these kinases in terms of fold and amino-acid side-chain positioning. These results were

produced using a novel structural alignment algorithm that will also be described. This algorithm superimposes structures independent of fold to maximize side-chain similarity. The result is not only an alignment but a consensus structure that depicts residue conservation as a distribution of amino acids and amino-acid categories. This consensus can be used to guide structure prediction, and we report here on its successful use with Rosetta[99] in predicting the structure of 3-phosphoinositide dependent protein kinase-1 (PDK1) and the atypical protein kinase Rio2.

## Results

*Alignment*

To examine conservation and variability between protein kinases we focused on a group of active-conformation structures. Obviously, it is important to examine like conformations so that any observed variability is in fact real. We defined an active kinase structure as one with ATP or a non-hydrolysable ATP analog, at least one divalent cation (always $Mg^{2+}$ or $Mn^{2+}$), and any necessary phosphorylations. Kinases can be constitutively active or be regulated positively or negatively by phosphorylation, which is ultimately kinase specific. Information regarding the kinase structures used in our analysis can be found in **Table 2.1**, and an example of an active-conformation protein kinase is shown in **Figure 2.1**.

Fifteen kinase structures were aligned using the sequence-order independent algorithm outlined in the Methods section to yield a consensus set of forty-four fully and partially conserved residues. The complete set is listed in **Table 2.2**. The structural alignment produced is shown in **Figure 2.2**. From such an image it can be seen that the overall kinase shape is a highly conserved feature. The active site occurs between two lobes: the small lobe above ATP and the large lobe below. Of particular importance for later discussion is the conservation of the substrate-binding groove, located between the catalytic loop, the P+1 loop, helix D, helix F, helix G and helix H. Conserved residues are shown in **Figure 2.3**, in what we term a consensus structure. This is a distribution of amino acids and amino-acid categories conserved between the protein kinases we have examined. The consensus structure has three principle parts: 1) a region of hydrophobic residues clustered around the adenosine of ATP; 2) an area around the γ-phosphate of

**Table 2.1**: Active-conformation kinase structures.

| Kinase | Full name | Species | PDB code | Pfam domain (residues) |
|---|---|---|---|---|
| ACK1 | activated CDC42 kinase 1 | *H. sapiens* | 1U54[100] | Protein tyrosine kinase (126-385) |
| Akt2 | RAC-β serine/threonine-protein kinase | *H. sapiens* | 1O6K[101] | Protein kinase (152-409) |
| CDK2 | cell division protein kinase 2 | *H. sapiens* | 1JST[102] | Protein kinase (4-286) |
| CK1 | casein kinase I | *S. pombe* | 1CSN[103] | Protein kinase (12-237) |
| CK2 | casein kinase II subunit α | *Z. mays* | 1LP4[104] | Protein kinase (34-319) |
| DAPK | death-associated protein kinase | *H. sapiens* | 1IG1[105] | Protein kinase (13-275) |
| IRK | insulin receptor tyrosine kinase | *H. sapiens* | 1IR3[106] | Protein tyrosine kinase (1023-1290) |
| MAPK p38γ | mitogen-activated protein kinase p38γ | *H. sapiens* | 1CM8[107] | Protein kinase (27-311) |
| PhK | phosphorylase kinase | *O. cuniculus* | 1PHK[108] | Protein kinase (19-287) |
| Pim-1 | proto-oncogene serine/threonine-protein kinase Pim-1 | *H. sapiens* | 1XR1[109] | Protein kinase (129-381) |
| PKA | protein kinase A | *M. musculus* | 1ATP[110] | Protein kinase (43-297) |
| PknB | probable serine/threonine-protein kinase PknB | *M. tuberculosis* | 1MRU[111] | Protein kinase (11-273) |
| Rio2 | Rio2 serine kinase | *A. fulgidus* | 1ZAO[112] | Rio1 family (105-275) |
| Sky1P | SR protein kinase | *S. cerevisiae* | 1Q97[113] | Protein kinase (158-706)[†] |
| TAO2 | thousand and one amino-acid protein 2 | *R. norvegicus* | 1U5R[114] | Protein kinase (28-281) |
| ChaK[‡] | transient receptor potential channel kinase | *M. musculus* | 1IA9[91] | Alpha kinase (1596-1814) |

[*]The average pair-wise sequence identity between this set as computed by ClustalW using its default parameters is 17%.
[†]Residues 304-541 constitute a large spacer within the kinase domain[115].
[‡]ChaK lacks a divalent cation in its active site and was not used to generate an initial alignment (see Results section).

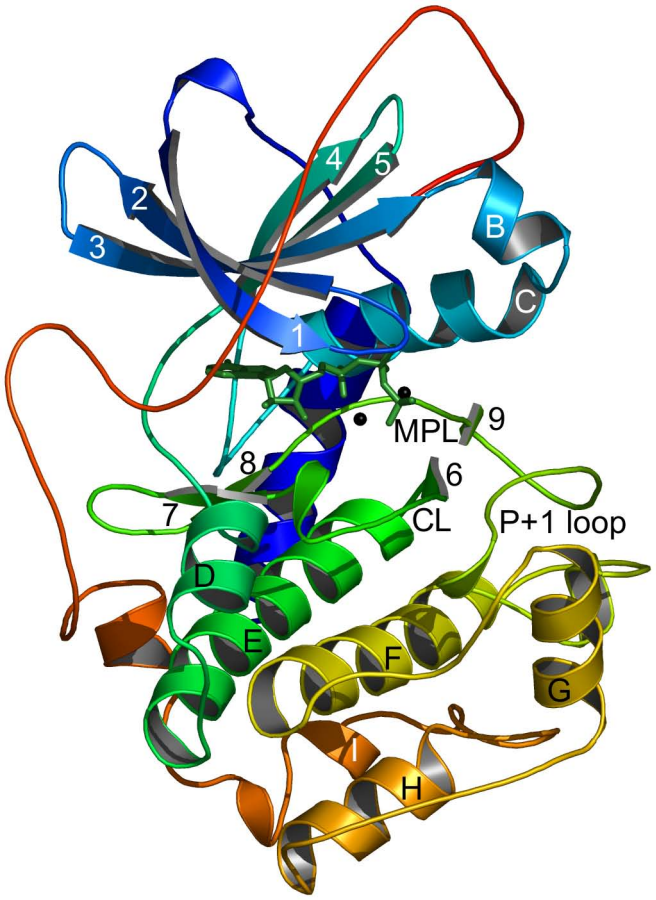



**Figure 2.1**. The structure of protein kinase A (PKA). PKA is shown in its active conformation with ATP in green sticks and $Mn^{2+}$ as black spheres. β-strands, helices and loops are labeled as in Knighton et al.[45] The active site is situated between the small and large lobes, located above and below ATP respectively. CL: catalytic loop; MPL: magnesium-positioning loop.

**Table 2.2**: Conserved residues found in the active-conformation kinases listed in **Table 2.1**.

| | type | ACK1 | Akt2 | CDK2 | CK1 | CK2 | DAPK | IRK | p38γ | PhK | Pim-1 | PKA | PknB | Rio2 | Sky1P | TAO2 | Chak* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | h | L132 | L158 | I10 | I18 | V45 | L19 | L1002 | V33 | L25 | L44 | L49 | L17 | M98 | L164 | I34 | x |
| 2 | G | 133 | 159 | 11 | 19 | 46 | 20 | 1003 | x | 26 | 45 | 50 | 18 | x | 165 | 35 | x |
| 3 | G | x | 161 | 13 | 21 | 48 | 22 | 1005 | x | 28 | x | 52 | 20 | x | 167 | 37 | 1619 |
| 4 | V | 140 | 166 | 18 | I26l | 53 | 27 | 1010 | 41 | 33 | 52 | 57 | 25 | 106 | 172 | 42 | A1624l |
| 5 | h | V155 | Y178 | V30 | V38 | C65 | Y39 | V1027 | V53 | Y45 | V64 | Y69 | V37 | C117 | V184 | V54 | Y1643 |
| 6 | A | 156 | 179 | 31 | 39 | I66l | 40 | 1028 | 54 | 46 | 65 | 70 | 38 | V118l | 185 | 55 | I1644l |
| 7 | h | V157 | M180 | L32 | I40 | I67 | A41 | V1029 | I55 | V47 | I66 | M71 | V39 | V119 | M186 | I56 | I1645 |
| 8 | K | 158 | 181 | 33 | 41 | 68 | 42 | 1030 | 56 | 48 | 67 | 72 | 40 | 120 | 187 | 57 | 1646/R1622b |
| 9 | E | 177 | 200 | 51 | 55 | 81 | 64 | 1047 | 74 | 73 | 89 | 91 | 59 | 154 | 202 | 76 | 1672 |
| 10 | L | M181h | 204 | 55 | Y59h | 85 | 68 | M1051h | 78 | 77 | 93 | 95 | A63l | 158 | 206 | 80 | x |
| 11 | L | 192 | 215 | 66 | V71l | 97 | 79 | 1062 | 89 | 89 | 106 | 106 | V74l | V169l | 228 | Y91h | x |
| 12 | h | M203 | F227 | L78 | L83 | L111 | L91 | V1074 | L107 | L101 | L118 | M118 | I90 | V177 | M244 | L103 | A1716 |
| 13 | V | 204 | 228 | 79 | 84 | I112l | I92l | 1075 | 108 | 102 | I119l | 119 | 91 | L178l | 245 | 104 | 1717 |
| 14 | E | 206 | 230 | 81 | x | 114 | 94 | 1077 | x | x | 121 | 121 | 93 | 180 | x | 106 | x |
| 15 | h | A208 | A232 | L83 | L88 | V116 | V96 | M1079 | M112 | M106 | P123 | V123 | V95 | I182 | L249 | C108 | M1721 |
| 16 | L | x | 237 | 87 | 92 | x | 101 | 1084 | 116 | 111 | 129 | M128h | 100 | 187 | 253 | A112l | x |
| 17 | h | V236 | I259 | L111 | M115 | L140 | I123 | I1116 | M137 | L133 | V151 | I150 | A122 | I202 | L277 | A135 | M1746 |
| 18 | h | M240 | L263 | L115 | V119 | L144 | V127 | M1120 | L141 | I137 | V155 | F154 | L126 | V206 | L281 | L139 | F1749 |
| 19 | H | x | 267 | 119 | 123 | 148 | 131 | x | 145 | 141 | 159 | 158 | 130 | x | 285 | 143 | T1753p |
| 20 | h | F248 | V271 | V123 | L127 | I152 | I135 | F1128 | I149 | I145 | V163 | L162 | I134 | I214 | I290 | M147 | A1678 |
| 21 | h | I249 | V272 | L124 | V128 | M153 | A136 | V1129 | I150 | V146 | L164 | I163 | I135 | V215 | I291 | I148 | x |
| 22 | H | 250 | Y273r | 125 | Y129r | 154 | 137 | 1130 | 151 | 147 | 165 | Y164r | 136 | 216 | 292 | 149 | x |
| 23 | R | 251 | 274 | 126 | 130 | 155 | x | 1131 | 152 | 148 | 166 | 165 | 137 | x | x | 150 | x |
| 24 | D | 252 | 275 | 127 | 131 | 156 | 139 | 1132 | 153 | 149 | 167 | 166 | 138 | 218 | 294 | 151 | 1765 |
| 25 | l | L253 | I276 | L128 | I132 | V157 | L140 | L1133 | L154 | L150 | I168 | L167 | V139 | L219 | I295 | V152 | L1766 |
| 26 | K | R256b | 277 | 129 | 133 | 158 | 141 | R1136b | 155 | 151 | 169 | 168 | 140 | S220p | 296 | 153 | 1727 |
| 27 | N | 257 | 280 | 132 | 136 | 161 | 144 | 1137 | 158 | 154 | 172 | 171 | 143 | 223 | 299 | 156 | Q1767p |
| 28 | h | L258 | L281 | L133 | F137 | V162 | I145 | C1138 | L159 | I155 | I173 | L172 | I144 | V224 | V300 | I157 | x |
| 29 | h | L259 | M282 | L134 | L138 | M163 | M146 | M1139 | A160 | L156 | L174 | L173 | M145 | L225 | L301 | L158 | x |
| 30 | h | L260 | L283 | I135 | I139 | I164 | L147 | V1140 | V161 | L157 | I175 | I174 | I146 | V226 | M302 | L159 | x |
| 31 | l | V266 | I289 | I141 | I150 | L171 | I157 | V1146 | L167 | I163 | L182 | I180 | V152 | I231 | I546 | V165 | x |
| 32 | K | 267 | 290 | 142 | Y151p | R172b | 158 | 1147 | 168 | 164 | 183 | Q181p | 153 | x | 547 | 166 | N1772p |
| 33 | l | I268 | I291 | L143 | V152 | L173 | I159 | I1148 | I169 | L165 | L184 | V182 | V154 | I233 | I548 | L167 | L1773 |
| 34 | D | 270 | 293 | 145 | 154 | 175 | 161 | 1150 | 171 | 167 | 186 | 184 | 156 | 235 | 550 | 169 | x |
| 35 | F | 271 | 294 | 146 | 155 | W176r | 162 | 1151 | 172 | 168 | 187 | 185 | 157 | 236 | L551h | 170 | P1776h |
| 36 | G | 272 | 295 | 147 | 156 | 177 | 163 | 1152 | 173 | 169 | 188 | 186 | 158 | P237s | 552 | 171 | x |
| 37 | P | 299 | 319 | 171 | x | 200 | 186 | 1178 | 194 | 192 | 210 | 207 | 185 | x | 573 | 191 | A1806s |
| 38 | E | 300 | 320 | 172 | N188p | 201 | 187 | 1179 | 195 | 193 | 211 | 208 | 186 | x | 574 | 192 | x |
| 39 | D | 312 | 332 | 185 | 200 | 214 | 199 | 1191 | 208 | 211 | x | 220 | 198 | x | 586 | 207 | x |
| 40 | W | 314 | 334 | 187 | x | 216 | 201 | 1193 | 210 | 213 | 226 | 222 | Y200r | x | 588 | 209 | x |
| 41 | G | 317 | 337 | 190 | 205 | 219 | 204 | 1196 | 213 | 216 | 229 | 225 | 203 | I259l | A591v | 212 | x |
| 42 | h | C367 | L384 | M266 | Y256 | L304 | L255 | C1245 | M291 | F267 | C270 | L272 | A254 | I259 | M686 | C261 | x |
| 43 | L | W368h | 385 | 267 | M257h | 305 | 256 | W1246h | 292 | 268 | 271 | 273 | 255 | x | 687 | 262 | x |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 44 | R | 375 | 392 | 274 | x | 312 | 263 | 1253 | 299 | 275 | 278 | 280 | 262 | x | 694 | 269 | x |

*ChaK was aligned directly onto the consensus generated from the other fifteen samples. See Results section.

The amino acid or amino-acid category of the conserved residue is listed under "type". For each structure the residue identifier corresponding to the conserved point is indicated (listed as x if the point is absent). For category types the amino acid present in each structure is indicated before the identifier. If a sample is missing a conserved amino acid but has a similar residue in the same location then the shared category is listed after the identifier. a, acidic; l, aliphatic; r, aromatic; b, basic; c, charged; h, hydrophobic; p, polar; s, small; v, very small.

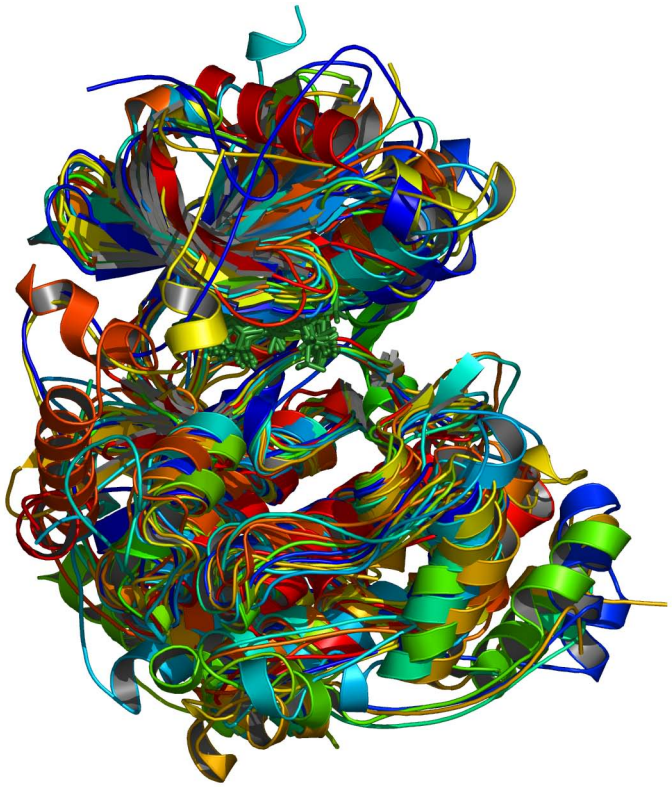


**Figure 2.2**. Multiple kinase alignment. The fifteen active-conformation kinase structures listed in **Table 2.1** were aligned using our modified Procrustes approach. Shown in green sticks is the ATP or ATP analog molecule of each structure. Each kinase is colored uniquely.

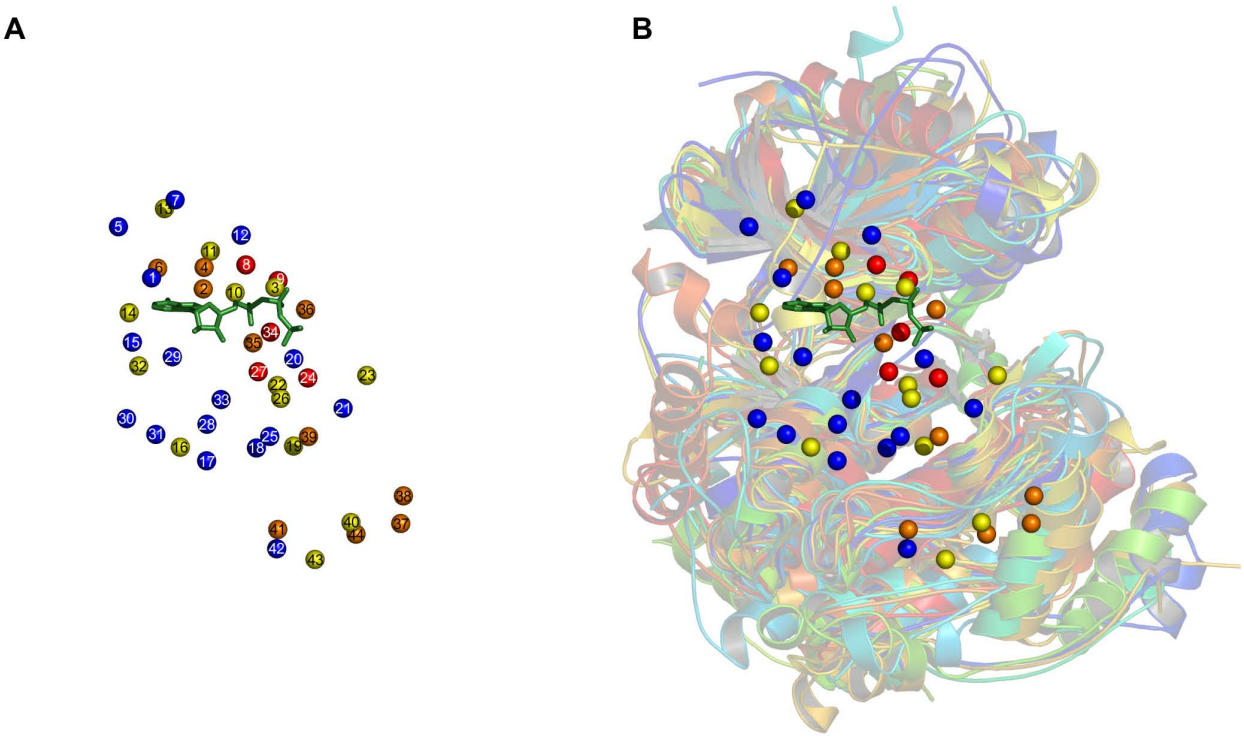



**Figure 2.3**. A kinase consensus structure. Each sphere represents a conserved residue. Red indicates full conservation of a particular amino acid in all fifteen kinase structures; orange, conservation in thirteen or fourteen structures; and yellow, conservation in eleven or twelve structures. Blue spheres indicate full conservation of an amino-acid category. The ATP molecule of protein kinase A is shown in green sticks. (**A**) The consensus structure consisting of the forty-four points listed in **Table 2.2**. (**B**) The consensus structure overlaid on the multiple alignment.

ATP – the active site – enclosed primarily by charged residues; and 3) a region in the large lobe, situated below ATP, of both hydrophobic and polar residues. The hydrophobic region around the adenosine creates a binding pocket for ATP. The charged residues in the active site bind and position the γ-phosphate, as well as the divalent cation, and participate in the catalytic mechanism. The conserved residues located in the large lobe serve to stabilize that region, and may play a role in mediating substrate interactions. Only five specific amino acids are fully conserved in all the kinases. These residues play critical parts in positioning ATP, stabilizing the active-conformation and in the catalytic mechanism. These are lysine 8, which interacts with the α- and β- phosphates of ATP, thereby stabilizing it. Glutamic acid 9, which forms a salt bridge with lysine 8 further stabilizing ATP. Aspartic acid 24 is the catalytic base that initiates phosphotransfer by deprotonating the acceptor serine, threonine or tyrosine. Asparagine 27 interacts with a secondary divalent cation, thereby positioning the γ-phosphate of ATP. And the final fully conserved residue is aspartic acid 34, which chelates the primary divalent cation, indirectly positioning ATP at the same time. Although these are the only residues fully conserved in terms of function, location and amino-acid type, there is one other residue with functional conservation but not locational or type, and in other kinases there is variability in the origin of lysine 8 and the type of amino acid fulfilling its role.

*Rio2, ChaK and conserved residue variability*

Rio2 is the only atypical protein kinase amongst those included in our multiple alignment. No significant sequence similarity can be detected between Rio2 and conventional protein kinases. As such, it is classified with a distinct domain name (see

**Table 2.1**). The structure of the Rio family, as initially determined by Laronde-LeBlanc and Wlodawer[116], shares significant similarity with serine/threonine and tyrosine kinases in the small lobe and in the regions of the large lobe directly adjacent to the active site (**Figure 2.4A)**. Dissimilarity in structure occurs predominantly in the large lobe, in regions involved in substrate specificity, suggesting Rio2 has evolved a novel mechanism of substrate recognition[116]. Our structural alignment algorithm concurs with these findings, showing high similarity in and around the active site (consensus residues 1-36) but not in the large lobe (consensus residues 37-44).

Slightly different results are produced for another atypical protein kinase: channel kinase (ChaK). The structure of this kinase was not included in our initial data set because it lacked a divalent cation in the active site. To examine structural similarity between ChaK and conventional protein kinases, we did align the partially active structure of ChaK (**Figure 2.4B**) with the consensus structure generated from fully active-conformation kinases (see **Table 2.2**). Most similarity is found in the region directly around the γ-phosphate of ATP, although there is some elsewhere. The large lobe, like Rio2, is quite distinct. One of the few differences between this kinase and the others is that the fully conserved asparagine 27 is replaced by the very similar carboxamide containing glutamine. The fully conserved aspartic acid 34 is present but does not align, likely due to the absence of a divalent cation.

We highlight these two kinases for the additional reason that they demonstrate structural variation within the bounds of certain functional constraints. An interesting case of this is lysine/arginine 26. In almost all serine/threonine kinases there is a lysine residue located two positions downstream of the catalytic aspartic acid (consensus residue

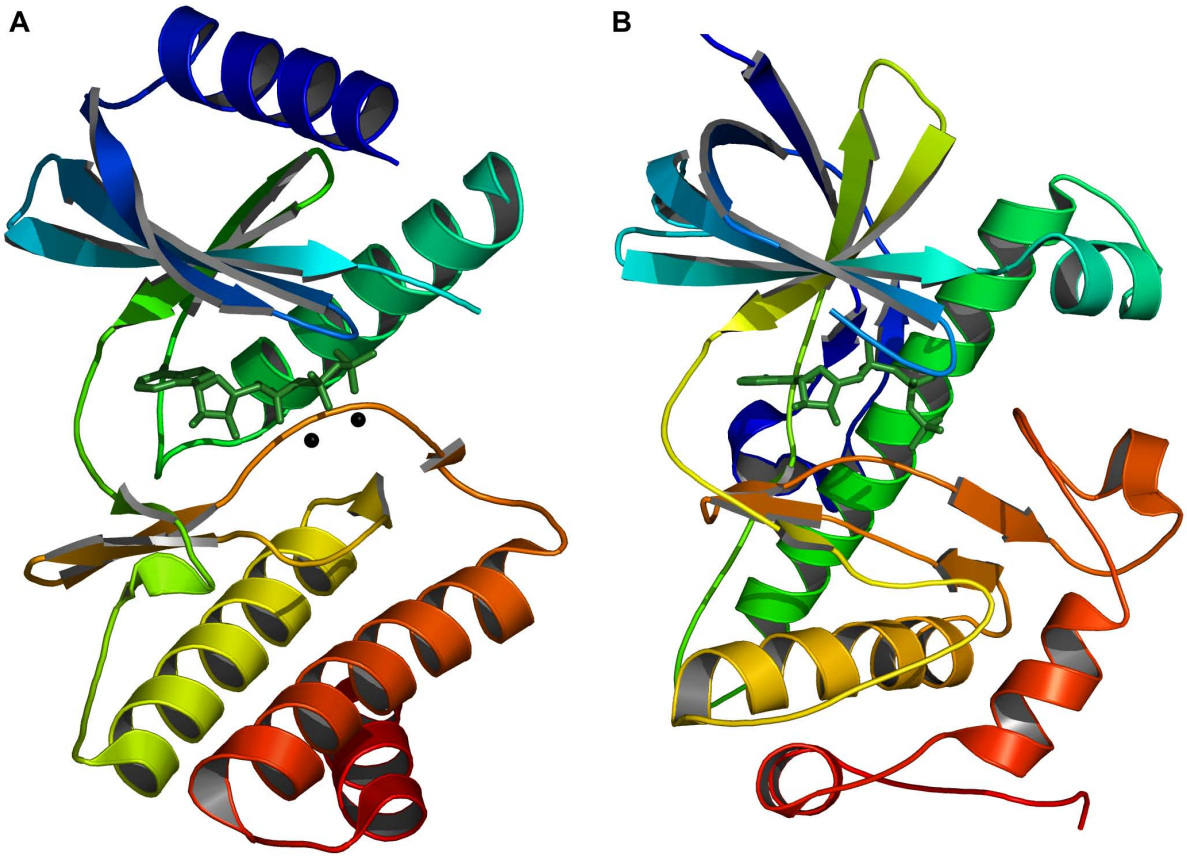



**Figure 2.4**. The atypical kinases (**A**) Rio2 and (**B**) transient receptor potential channel kinase (ChaK).

24). This lysine aids in orientating the γ-phosphate of ATP and it is thought may also act to neutralize the negative charge on this phosphate during catalysis. The position and orientation of these two residues with respect to ATP is shown in protein kinase A (PKA) in **Figure 2.5A**. Tyrosine kinases, like activated CDC42 kinase (ACK1), do not have this lysine but instead have an arginine four positions downstream from the catalytic residue (**Figure 2.5B**). This residue is orientated perpendicular to the lysine but occupies the same location, and since both are positively charged basic residues, both can fulfill the same function. ChaK, a serine/threonine kinase like PKA, utilizes a lysine for this function, which again occupies the same geometric position (**Figure 2.5C**). However, this residue is not located two or four positions downstream, but thirty-eight positions upstream on a β-strand running adjacent to the catalytic loop. This is a good demonstration of the value behind a sequence-order independent alignment algorithm that ignores fold and residue connectivity. This lysine is located on part of a novel fold not found in any of the other kinases examined. Rio2 presents a fourth variation. Our alignment did not reveal Rio2 as having a basic residue at consensus point 26, as it did for all of the other kinases. Further examination led us to the conclusion that the function of this residue is conserved in Rio2 but the location of the residue accomplishing it is not (see also LaRonde-LeBlanc and Wlodawer[116]). This is known as functional residue hopping[117, 118]. The function of consensus residue 26 is performed by a histidine located in the small lobe (**Figure 2.5D**). This histidine is largely unique to the *Archaeoglobus fulgidus* Rio2 ortholog, from which the structure was derived – in most other species it is substituted by an arginine. As *A. fulgidus* is a hyperthermophile, the preference for histidine may be due to the extreme temperature environments in which it is found.

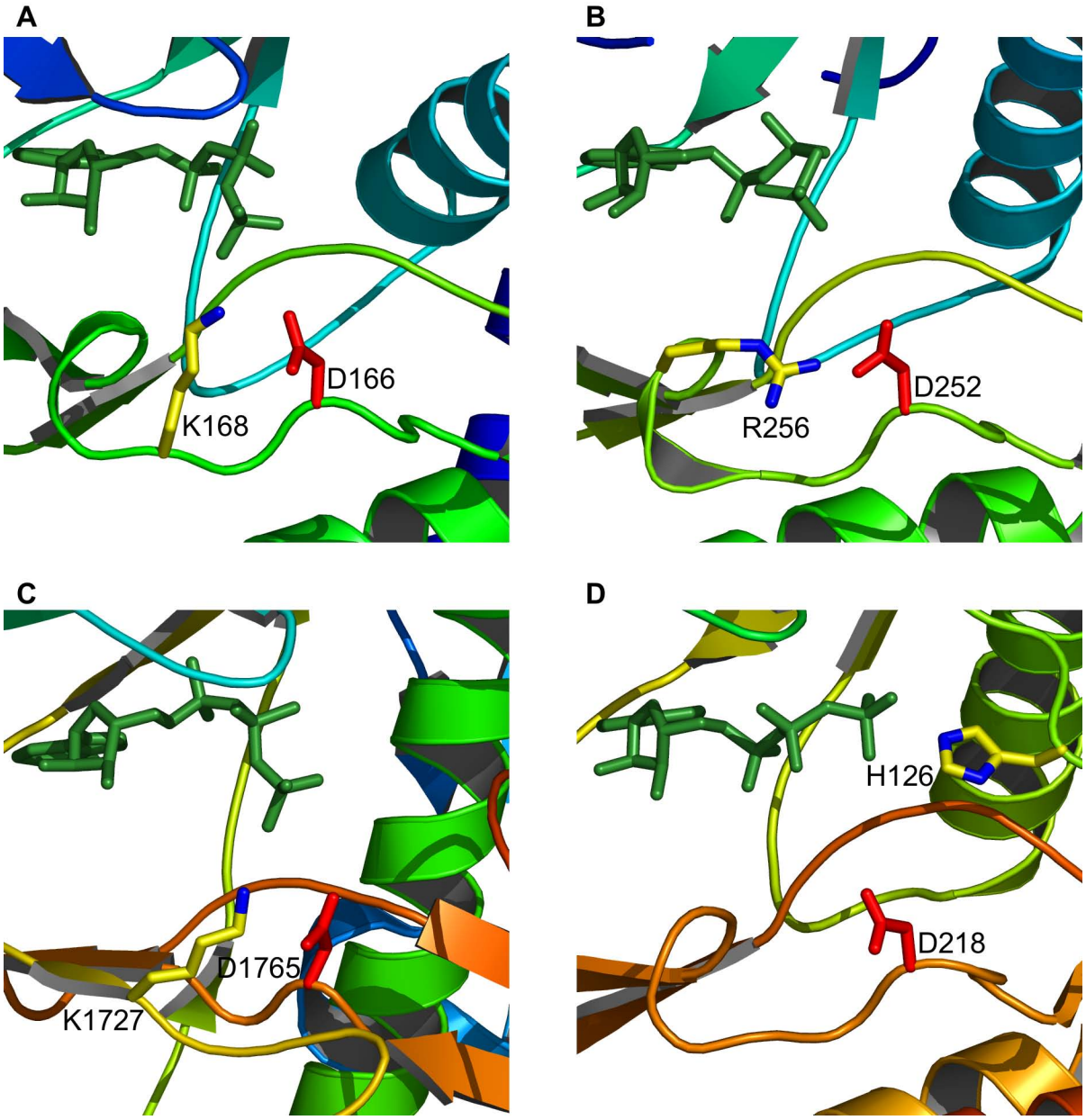



**Figure 2.5**. Residue variability in positioning the γ-phosphate of ATP. In each kinase the catalytic aspartic acid is shown in red and the positively-charged residue that interacts with the γ-phosphate of ATP is shown in yellow with nitrogen atoms colored blue. ATP is shown in green sticks. (**A**) Protein kinase A, (**B**) activated CDC42 kinase 1, (**C**) channel kinase, and (**D**) Rio2 kinase.

The only other variability we have found in conserved residues is for the so-called catalytic lysine (lysine 8). The function of this residue is somewhat debated[119-121], but at the very least it appears to position the phosphates of ATP and is absolutely crucial for catalysis. This residue is normally found on β-strand 3 and interacts with the α- and β-phosphates of ATP (**Figure 2.6A**). In ChaK there is a homologous lysine present but it interacts with the α-phosphate and the adenine ring of ATP (**Figure 2.6B**). Depending upon the alignment parameters used, this lysine can align structurally with that found in other kinases, although for the parameters we have chosen it does not. Instead there is an arginine residue that aligns and this very similar residue interacts with both the α- and β-phosphates of ATP (**Figure 2.6B**), just as the catalytic lysine normally does. Although ChaK lacks a divalent cation, this is unlikely to affect the positions of these residues. The arginine in question, R1622, is fully conserved in the alpha kinase family[122]. It is not known which residue in ChaK is functionally homologous to the catalytic lysine in typical protein kinases. Likely both share the function, and this represents a distinct feature of the alpha-kinase family. One other kinase is known to have a novelty in this area. This is the protein kinase with no lysine (WNK) kinase, named for the apparent absence of the catalytic lysine on β-strand 3. It was predicted by Xu et al.[16] that a lysine on β-strand 2 could be structurally equivalent (**Figure 2.6C**), and the absolute requirement for a lysine at this position was confirmed by this group. A subsequent crystal structure appears to confirm these predictions[123]. It is interesting that this lysine originates from the same position as the arginine in ChaK, showing that variation in conserved residues is possible and highlighting how proteins can incorporate novel features within certain functional constraints.

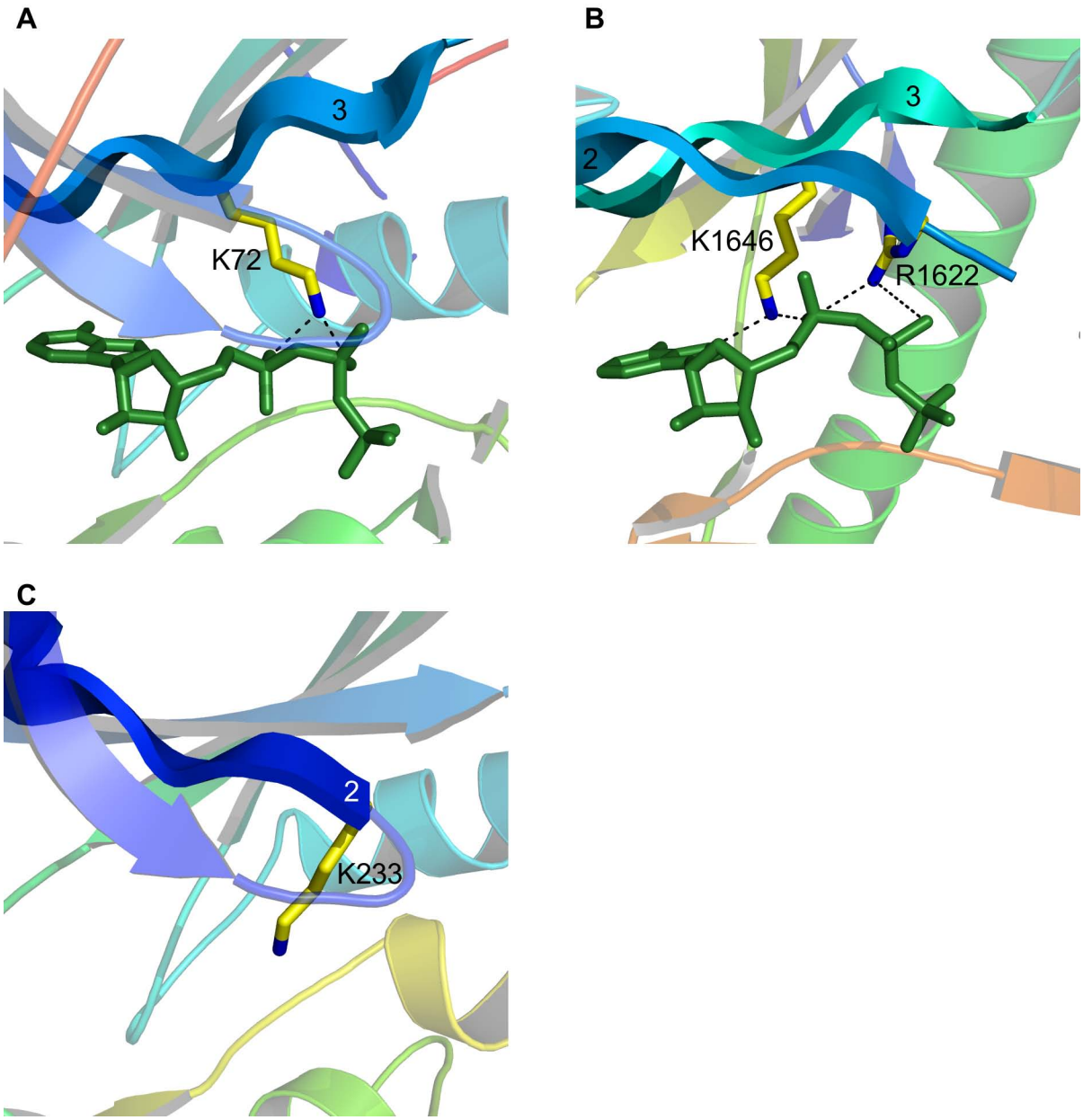



**Figure 2.6**. Variations in the catalytic lysine. (**A**) In almost all protein kinases (such as protein kinase A shown) a lysine residue originating from β-strand 3 interacts with the α- and β-phosphates of ATP. This lysine is required for catalytic activity and has been termed the catalytic lysine. (**B**) In channel kinase (ChaK), the homologous lysine interacts with the α-phosphate and the adenine ring of ATP. An unique arginine residue located on β-strand 2 instead interacts with the α- and β-phosphates. (**C**) In protein kinase with no lysine 1 (WNK1, PDB code: 1T4H[123]), the catalytic lysine is present but

originates from β-strand 2, much like the arginine from ChaK. ATP is shown in green sticks. Potential hydrogen bonds between the positively charged residues and ATP are shown as dashed lines, and nitrogen atoms are colored blue. The WNK structure has no ATP or ATP analog.

*Substrate specific variation*

The conservation we are detecting is found first and foremost in and around the active site, but is present elsewhere in the protein kinase domain. The single exception is in the cleft formed by the catalytic loop, the P+1 loop, helix D and the residues from the end of helix F through to the beginning of helix H. This is known as the substrate-binding cleft, and residues in this region have been shown crucial to substrate binding and regulatory protein-protein interactions[101, 110, 124-128]. The absence of conservation in this cleft is entirely expected. Protein kinases are substrate specific and that specificity is at least in part conferred through residue variability within this groove.

The crystal structures of PKA and Akt we have used in our analysis contain substrate peptides bound in this region. In **Figure 2.7** we show PKA and Akt with bound substrate peptides and display variability in the context of proximity to the substrate and distance from conserved residues. In both PKA and Akt, there are a number of atoms in the substrate-binding cleft adjacent to or near the substrate peptide that are distant from conserved residues. We highlight this for the purpose of discussing inhibitor design. Successful inhibitor design requires a region where small molecules or peptides can be bound with high specificity. The absence of specificity comes from drug targets hitting regions of residue conservation. The substrate-binding cleft obviously has a binding capacity and, as shown, this region is highly variable. Designing inhibitors that target the atoms in this region, mimicking those residues present in substrates or regulatory proteins, would be a fruitful approach.

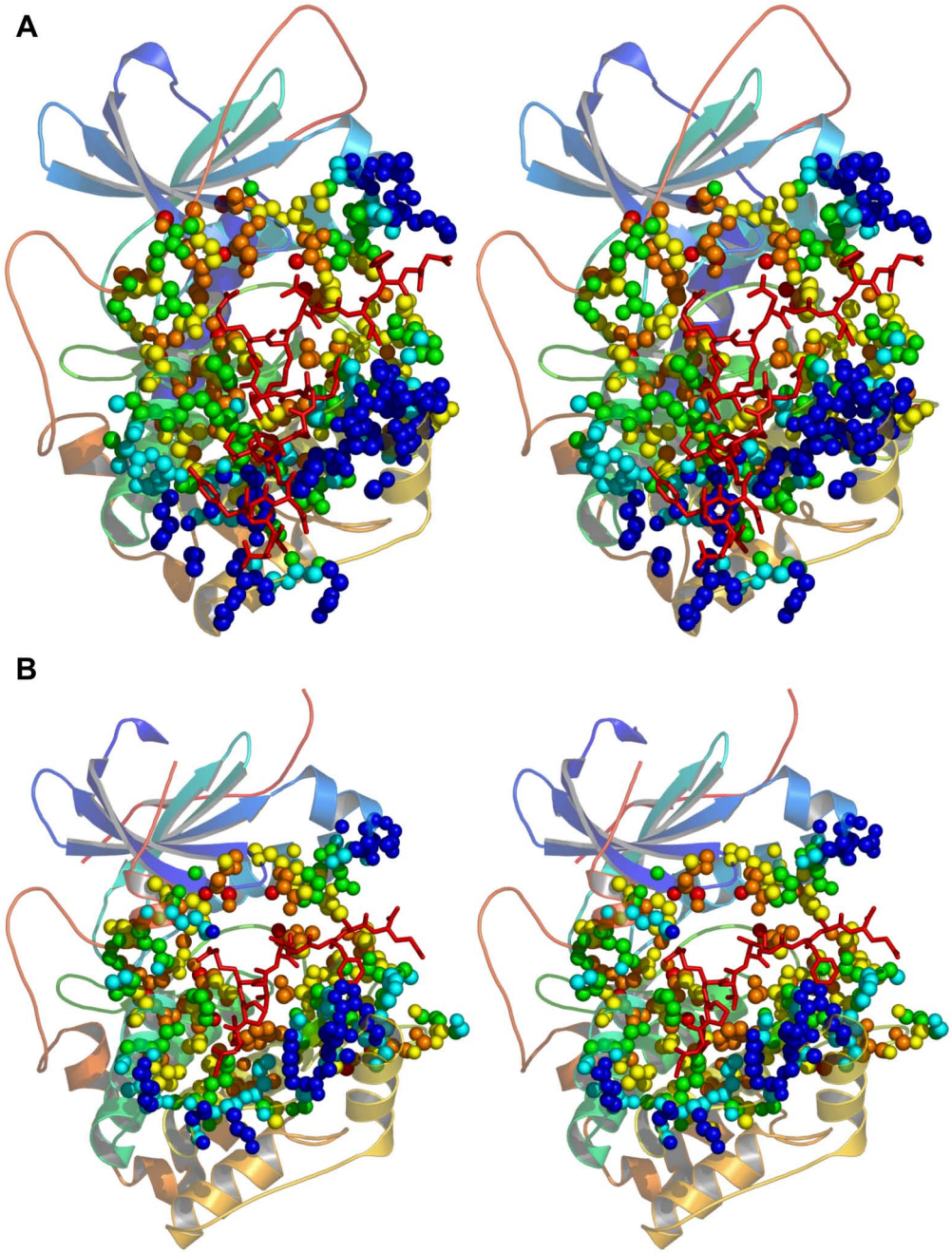



**Figure 2.7**. Substrate-specific variability. Cross-eyed stereo views of (**A**) protein kinase A and (**B**) Akt2. Bound substrate peptides are shown in red sticks. Atoms belonging to non-conserved residues within 10 Å of the substrate peptide are shown as colored spheres. Red: atoms within 2 Å of a conserved residue; orange: atoms between 2 and 4 Å of a conserved residue; yellow: atoms between 4 and 6 Å of a conserved residue; green:

atoms between 6 and 8 Å of a conserved residue; cyan: atoms between 8 and 10 Å of a conserved residue; and blue: atoms more than 10 Å from a conserved residue.

*Structure prediction*

The conservation we are finding can be used for purposes other than understanding protein function, evolution and guiding drug design. It can also be used to predict the structure of protein kinases (and to other proteins if applied). The consensus structure shown in **Figure 2.3** represents the typical position of specific and conserved amino acids or amino-acid categories. If this is a distribution that active protein kinases tend to adopt, then predicted structures of active-conformation protein kinases should be made to meet these criteria. At the most basic level, models can be generated through any method and then screened against a consensus structure to discriminate between good and bad models, or, alternatively, residues known to be equivalent from a sequence alignment can be forced to meet the constraints seen in the consensus structure and the rest of the protein can be modeled within this framework. We have explored these possibilities with the structure prediction method Rosetta[129] and attempted to model two kinases, the typical 3-phosphoinositide dependent protein kinase-1 and the atypical Rio2 kinase.

We used fourteen active-conformation kinases (omitting Rio2) and generated a 52 residue consensus structure (consensus residues are listed in **Supplementary Table 2.1**). A sequence alignment of these fourteen kinases and PDK1 was then generated using ClustalW[130]. Residues from PDK1 apparently equivalent with those of the consensus were selected based on this sequence alignment (see **Supplementary Table 2.1**). These residues were then constrained geometrically in accordance with the consensus while PDK1 was modeled with Rosetta as described in the Methods section. When compared against the partially active-conformation structure of PDK1 (PDB code: 1H1W[131]), the top ranked prediction had 198 of 285 side chains positioned within 2 Å of their actual

location, a $C_\alpha$ RMSD of 1.3 Å and an all-atom RMSD of 1.6 Å (**Figure 2.8A**). The floor of the active site, where most conservation occurs, is highly accurate, with 24 of 25 side chains positioned with 2 Å, $C_\alpha$ RMSD of 0.5 Å and an all-atom RMSD of 0.7 Å (**Figure 2.8B**). In non-conserved regions, such as the substrate-binding groove, good modeling is dependent upon the ability of the prediction method applied. A well-proven method like Rosetta is, therefore, a good complement. 24 of 38 residues in the substrate-binding groove are within 2 Å of their actual position, a $C_\alpha$ RMSD of 0.9 Å and an all-atom RMSD of 1.4 Å (**Figure 2.8C**). Accuracy in this region may be due in part to the constraints applied elsewhere, which would reduce the potential conformational space to search.

Rio2 kinase was modeled initially without constraints as these cannot be derived from a sequence alignment. 80,000 models were generated and the lowest 5% (in full atom energy) were screened against the consensus structure. The top ten were then scrutinized for potential constraints. All ten of the top models unambiguously agreed on the likely equivalent residues for the fully conserved lysine 7 (K120 in Rio2), aspartic acid 24 (D218), asparagine 27 (N223) and aspartic acid 34 (D235). In none of the models could a residue equivalent to the fully conserved glutamic acid 8 be found (it should be E154). We then proceeded with a second modeling phase using constraints from the consensus for K120, D218, N223 and D235, and allowing two candidates for the conserved glutamic acid: E134 or E154. Although models generated with E134 as the conserved glutamic acid scored equally well when compared against the consensus structure, they would likely be deemed implausible by visual inspection as helix C was distorted upwards away from the active site instead of lining the back of the ATP binding

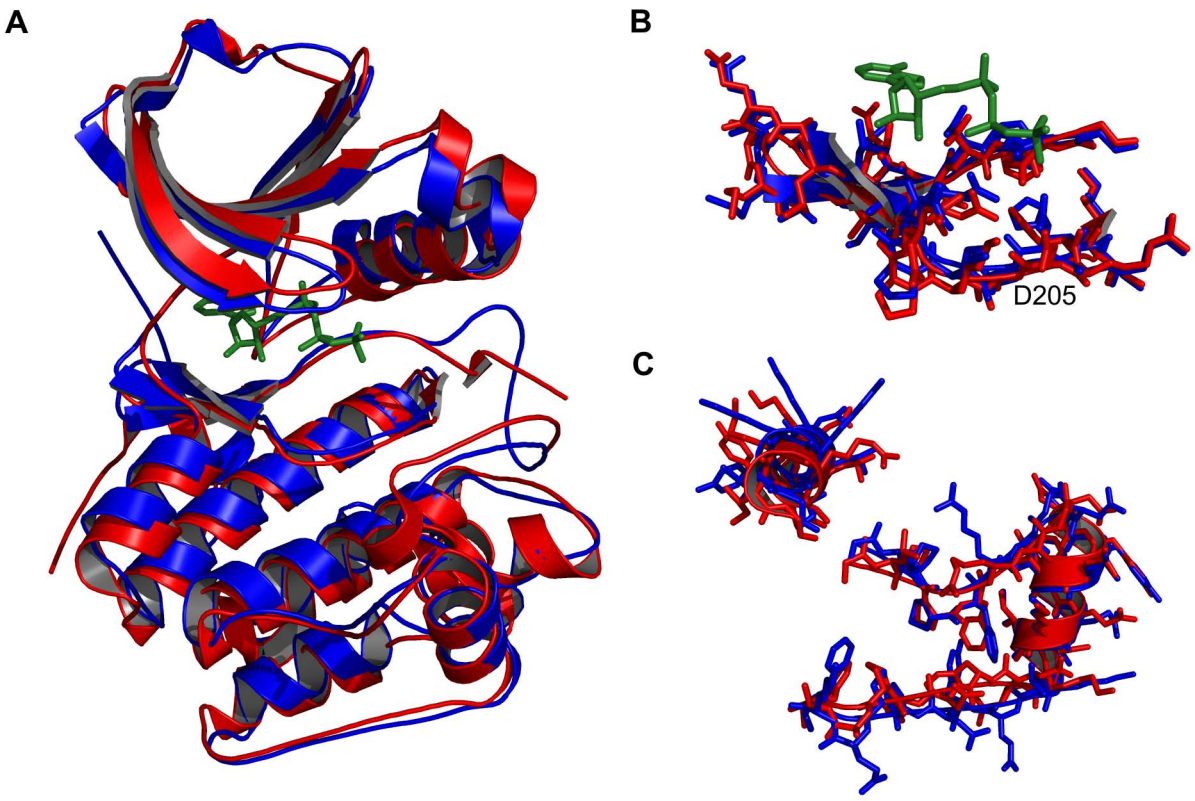



**Figure 2.8**. Modeling a typical kinase. PDK1 was modeled using the consensus guided approach outlined in the Results and Methods sections. (**A**) The best model is in blue and the actual structure in red (PDB code: 1H1W). Structures are shown aligned based on $C_\alpha$ RMSD. (**B**) Residues 201-225, comprising the floor of the active site, are highlighted as aligned on side chains. (**C**) Residues from the non-conserved regions of the substrate binding groove (residues 166-175 and 278-305) are shown as aligned on side chains. ATP from the true structure of PDK1 is shown in green sticks.

pocket as it does in other kinases. The best model for E154 as determined against the consensus structure is shown in **Figure 2.9A** alongside the actual active-conformation structure (PDB code: 1ZAO). Fifty four of 180 side-chains are positioned within 2 Å of their true location. $C_\alpha$ and all-atom RMSD are 6.1 Å and 7.1 Å respectively. However, these values do not accurately reflect the quality of the model. The N-terminus of protein kinases is a non-conserved region and the C-terminal lobe of Rio2 differs greatly from that of other kinases. It should not be expected that a structural-based consensus approach would be able to distinguish a correct model from an incorrect model in these regions. Looking solely at the conserved kinase regions, comprising residues 93-242 of Rio2, the $C_\alpha$ and all-atom RMSD drop to 3.0 Å and 3.7 Å respectively. And all of the correctly positioned residues are found in this region. The floor of the active site has 16 of 24 side chains positioned with 2 Å, $C_\alpha$ RMSD of 1.0 Å and an all-atom RMSD of 1.7 Å (**Figure 2.9B**). A final note: although residues located in the non-conserved large lobe helices are not correctly positioned, this region is topologically correct.

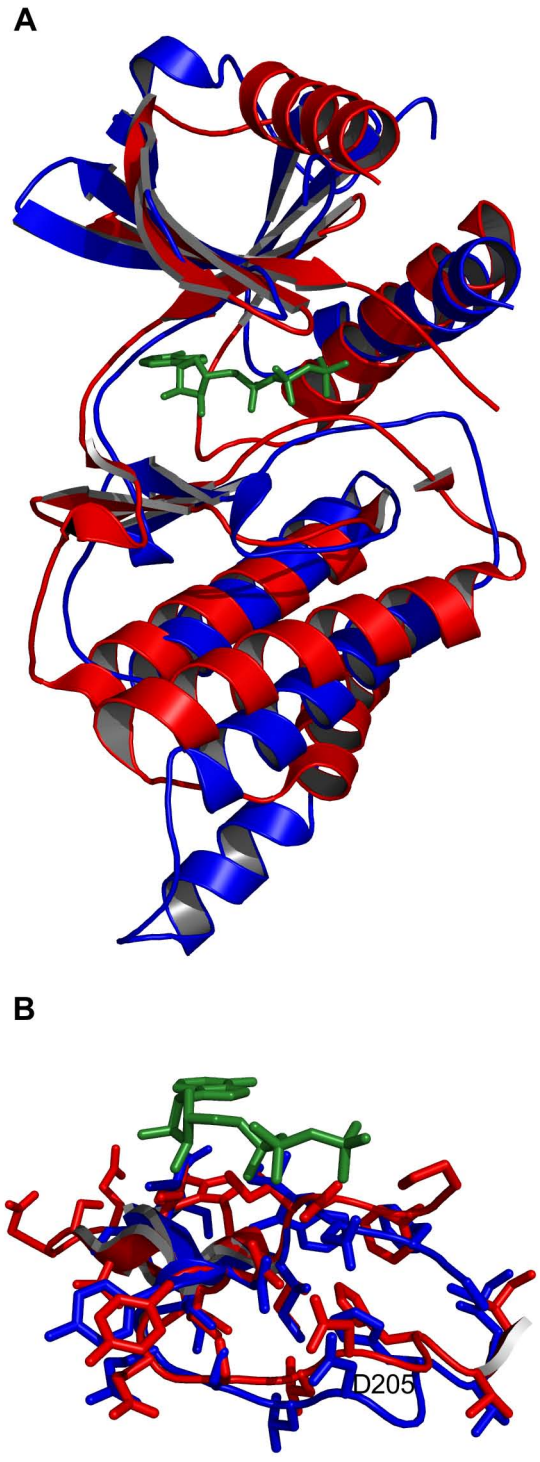



**Figure 2.9**. Modeling an atypical kinase, Rio2. (**A**) The best model is in blue and the actual structure in red (PDB code: 1ZAO). Structures are shown aligned based on $C_{\alpha}$

RMSD. Residues 128-142 from the model are hidden because they are disordered in the actual structure. (**B**) Residues 214-237 are highlighted as aligned on side-chains. ATP is shown in green sticks.

**Discussion**

The kinase domain can be conceptualized as two functional modules. A highly conserved ATP-binding and catalytic module, located between the small and large lobes, found in all typical and atypical protein kinase structures. Little variation is found here, with only a few kinases, such as Rio2, ChaK and WNK, containing structural (but not functional) novelties. The second module is involved in substrate binding and evidence suggests is localized primarily to the large lobe, in the region named the substrate/peptide binding groove (named from kinase-peptide co-crystal structures). Very little is known about this region of kinases as detailed knowledge would require many kinase structures co-crystallized with full-length substrates, and as yet no single such structure exists. Conservation of the groove fold suggests a common substrate-binding mechanism, while the absence of residue conservation in this region across the kinase family suggests the means by which substrate specificity is determined. The part of this groove located near the active site (between the catalytic and P+1 loops) does contain conserved residues and is entirely consistent with the broad peptide-substrate specificity of kinases. But peptides are not the same as full-length proteins, and studies on the mitogen-activated protein kinase p38 have shown that proteins can be phosphorylated at a hundred to thousand fold greater rate than peptides[132]. More than just the residues immediately attached to and surrounding the target serine, threonine or tyrosine are involved in the kinase-substrate interaction, and the general mechanism of interaction needs to be determined.

Although the substrate/peptide binding groove may be the key feature for understanding substrate specificity in most kinases, it is and need not be in all. The substrate-binding module can vary in three ways: typical kinases that preserve fold and

merely vary the residues in the substrate-binding groove, typical kinases than rely on other regions/regulatory proteins in addition to the substrate-binding groove, and atypical kinases that vary the binding mechanism all together. It should be added that even in very typical kinases substrate interactions are not localized purely to the substrate-binding groove (see Biondi and Nebreda[46]). Much work needs to be done on this area of protein kinases, and current knowledge of substrate-binding mechanisms is rudimentary.

A thorough understanding of substrate interactions will help in discovering new proteins targeted by kinases. Insights into binding can guide computational docking approaches that test the fit of different structures onto kinases as a new means of substrate finding. More experimentally determined structures or advances in structure prediction would be a necessary requirement for this to have broad applicability. This may be a ways off yet, but a closer goal may be in applying structural binding information to the design of specific kinase inhibitors. Currently, inhibitors are primarily designed to target the ATP binding pocket, the region of greatest conservation amongst kinases. The ATP binding pocket does contain unique features between certain subsets of the kinome (the gatekeeper residue is an example[133, 134]), and these novelties can be exploited by drugs. However, this type of design approach will always be plagued by problems of specificity on the simple foundation that this region has evolved to bind a single thing: ATP. On the other hand, the substrate-binding groove has evolved to bind kinase-specific substrates and our analysis demonstrates that this is accompanied by large variability in the substrate-binding groove. Mimicking these substrates with peptides or small molecules that compete with substrates has proven effective in several cases (reviewed in

Bogoyevitch et al.[135]), and a peptide inhibitor that mimics a substrate interaction domain has been shown useful *in vivo* at reducing tumor mass[136].

Of great benefit in this regard would be data on kinase substrates. If wide-ranging data on kinase substrate pools were known, inhibitors could be designed to mimic exclusive targets. Or if there was a desire to inhibit multiple kinases at one time, a small molecule that mimics a common substrate could be made. Ultimately the viability of such a strategy would again require a greater output of crystal and NMR structures, advances in structure prediction and protein docking, but these are popular areas of research and should not prove a hindrance. A dendrogram classification of kinases based upon substrates would be of tremendous use for inhibitor design, cell biology and evolutionary studies. Such an aim would require systematic methods of substrate finding, which are beginning to become available[49, 137].

Protein structures are the key to what we have presented and are discussing. Unless there are significant advances in the rate and ease with which proteins can be cloned, expressed, purified and structures subsequently determined, reliable computational approaches at structure determination will be needed. We have found that structural conservation gives insight not only into protein function but also can be used with success for structure prediction. We have shown how knowledge of residue conservation can help in generating or selecting appropriate protein kinase models. The atypical Rio2 kinase, which possesses no significant sequence similarity to other protein kinases, would be a difficult target for standard approaches at homology modeling. As few as five geometric residue constraints derived from a consensus structure and further screening can select a highly accurate model. For a typical kinase, such as PDK1, where

many constraints can be used (50 in our case), it was a simple matter to generate a highly accurate model. Residue conservation and other types of structural similarity quantified over protein families can provide the basis for model selection and act as guides that reduce the target search space for the computationally demanding task of structure prediction.

Understanding the individual role of every protein kinase in the genome is a daunting task, but necessary because of the important signaling role played by this family of enzymes and the potential for disregulation and subsequent disease. Ultimately, cellular studies that dissect the role of kinases *in vivo* and the treatment of disease will require a library of specific and multi-target inhibitors. Such an aim will require a strong union between computational, molecular and structural approaches, but this goal may be close at hand. Although empirically determined structures of all kinases are a long way off, quality models can be built. Kinome wide data on substrate pools will require a great deal of work, but the techniques are now becoming available. The mechanism(s) of substrate binding will need to be elucidated in detail. This may be the greatest challenge, and towards which maximum focus should be directed, as not a single kinase-substrate structure exists. But, if such can be achieved in a representative sample of cases, kinase-substrate specific docking algorithms may be able to do the rest. The attainment of these aims would allow for the quick and reliable design of specific inhibitors, thereby allowing the precise function(s) of kinases to be determined and creating the means by which signaling networks can be precisely manipulated.

## Materials and Methods

*Structural alignment overview*

The results presented in this paper were generated through the use of a multiple structure alignment algorithm we have developed. The motivation was to have a method for superimposing protein structures independent of residue connectivity, thereby facilitating comparisons between proteins that have distinct folds and maximize side-chain as opposed to main-chain overlap. The field of shape analysis, specifically the Procrustes methods, provided the foundation for accomplishing this task. What follows is an overview of Procrustes and a description of its modification and application to protein structure comparison as we have implemented it.

*Procrustes*

Procrustes is a method for comparing matrices and thereby any group of objects that can be represented in such a way[138, 139]. It is part of the broader field of shape analysis which has its motivation in describing and understanding variation between objects through multivariate analysis. Procrustes is often applied in biological and anthropological studies of morphology, for example in studying the relationship between turtle skull shape and life style across species[140]. It has been used in agriculture, genetics, geography, geology and psychology to name a few fields. Although multivariate analysis is not of interest, the Procrustes method is valuable for its ability to efficiently superimpose objects. Procrustes superimposition initiates from a set of points, each of which has a representative in every object under consideration. These points, known as landmarks, must have some level of correspondence and be located within a geometric space. By manipulating scale, location

and orientation, Procrustes aims to produce an optimal superimposition as defined by some criteria that acts to minimize distance between corresponding points. The term partial Procrustes specifically refers to the subset of methods that maintain scale, i.e. those approaches that only perform rigid body transformations[141].

*Application and algorithm*

An application for Procrustes in protein structure alignment is evident, where landmarks could be viewed as equivalent residues in a space of three dimensions. The term equivalent accounts for functional equivalence and/or homology, although emphasis should be placed on the former. Specifically it is the partial Procrustes method that is applicable since initial aspects of scale should be maintained in the resulting alignment.

Standard implementations of Procrustes require a set of user-specified landmarks with representatives in every input object. These two conditions are inadequate for multiple structure alignments. The primary motivation behind a structural alignment should be to find conserved residues, and a good algorithm must give leeway for less than total conservation. The algorithm we have developed allows for both. It proceeds in two phases. The first is a series of pair-wise comparisons to find residues conserved across all samples. Residue conservation means in terms of both location and amino-acid type (identity/similarity). The second phase begins with an initial multiple alignment using the landmarks found in the first phase. From the multiple alignment the set of landmarks is then extended to include any additional fully conserved residues not found in the first phase, and residues with less than total conservation. The multiple alignment and

extension steps are repeated until a final non-extendable set is obtained or upon completion of a predefined number of iterations.

In this approach, each of $n$ structures is represented by a $l_i \times 3$ coordinate matrix, where $l_i$ equals the number of residues in structure $i$. The mean of the non-hydrogen side chain atoms is used as the coordinate for each residue, with the exception of glycine residues for which the alpha carbon ($C_\alpha$) is used. Most structure alignment algorithms use the $C_\alpha$ for all residues. This is inadequate. Residue function should be viewed largely in terms of side-chain characteristics: what atoms make up the side chain and where they are positioned. It is the side chain that is principally used to interact with ligands, substrates and other residues[142], making knowledge of its position the crucial feature. Superimposed $C_\alpha$'s of similar residues may have side chains positioned in different locations and are therefore unlikely to have the same function. For these reasons focus is placed on side-chain positioning.

One object (structure) can be superimposed onto a second in three dimensions if at least three equivalent residues are known. It is possible these residues may or could be known beforehand, although this is not ideal. An alternative involves a search and score approach: choose three residues from each sample that have corresponding amino-acid labels and share some geometric feature, then score the alignment produced from superimposing these residue triplets. The highest scoring alignment produced from a series of triplet superimpositions can then be used to produce a set of residues conserved across a pair of samples (we term this a pair-wise consensus set). This approach has been employed in a hash table. For all samples every triplet of residues is stored based on the amino-acid labels of its members and the inter-residue distances are indexed. Each entry

of the table can be thought of as a triangle with vertices labeled according to amino acid. Between pairs of samples highly congruent triangles with matching vertices are used as candidate triplets likely to produce high scoring superimpositions. Such criteria reduce the time spent on triplets least likely to be part of the final pair-wise consensus set. We have used congruency criteria of 1-10% as indicated below, i.e. corresponding vertices must have magnitudes within 1-10% of each other.

The method of superimposition is taken from Procrustes. Candidate triplets from each sample *i* are stored in an 3 × 3 coordinate matrix $X_i$. These matrices are first Helmertized (centered), removing translational differences:

$$\dot{X}_i = H^T H X_i, \tag{1}$$

where $\dot{X}_i$ is the centered matrix and $H$ the $(m - 1) \times m$ Helmert sub-matrix whose $j^{\text{th}}$ row is equal to $(h_j,\ldots, h_j, -jh_j,0,\ldots,0)$ for $h_j = -[j(j + 1)]^{-1/2}$. Other approaches for removing translation are available. The optimal rotation of $\dot{X}_i$ onto $\dot{X}_j$ is found by minimizing

$$\left\| \dot{X}_j - \dot{X}_i \Gamma \right\|^2, \tag{2}$$

where $\Gamma$ is a 3 × 3 rotation matrix. It can be shown that if $V \Delta U^T$ is the singular value decomposition of $\dot{X}_j^T \dot{X}_i$, the optimal rotation $\Gamma = UV^T$ (provisions must be made to ensure $\Gamma$ does not also encompass a reflection).

The translation resulting from centering triplets is then applied to each structure and the appropriate sample is rotated by $\Gamma$. This superimposes the equivalent triplet residues and pulls the rest of each structure along. The result is that the initial pair-wise consensus set a triplet comprises may be extended from the structure-wide superimposition. Residues from the superimposed structures are considered to be equivalent (i.e. in the pair-wise consensus set) if they are identical or similar amino acids

within two angstroms of one another. Similar means sharing a category, with those allowed including acidic (D, E, Z), aliphatic (A, G, I, L, V), aromatic (F, H, W, Y), basic (H, K, R), charged (D, E, H, K, R, Z), hydrophobic (A, C, F, G, I, L, M, P, V, W, Y), polar (C, D, E, H, K, N, Q, R, S, T, Y, Z), small (A, C, D, G, N, P, S, T, V) and very small (A, G, S), where Z refers to a phosphorylated serine, threonine or tyrosine.

This procedure is performed on all triplets that match the congruency criteria and all alignments are scored based on the frequency with which the amino acids of any equivalent residues occur. The frequency of occurrence for each amino acid was obtained from the protein knowledge-base release statistics of Swiss-Prot (Release 49.5, 18/04/06). Tryptophan is the lowest occurring amino acid and superimposed tryptophans are therefore given a score of one. All other scores are relative to that of tryptophan: score(amino acid x) = occurrence(W/x). The occurrence of each category is equal to the sum of the occurrences of its members. For *m* equivalent residues, the alignment score is

$$\sum_{i=1}^{m} occurrence(W / x_i), \qquad (3)$$

where $x_i$ is the amino acid or amino-acid category of the $i^{th}$ equivalent residue.

The top 100 scoring triplets are subjected to a second phase of testing to find maximum similarity between two structures. For these triplets the initial pair-wise consensus set is used for a second superimposition. In this case the $X_i$ coordinate matrices take on dimensions $m \times 3$, for *m* equivalent residues in the pair-wise consensus set. The superimposition and extension steps are repeated until a final non-extendable set is obtained.

Computational time and memory are minimized in three ways. First, as has been mentioned, congruency between triangles is used to reduce the number of triplets tested.

Second, the vertices of these triplets are only indexed by amino acid and not by category as well. If vertices could be indexed by amino-acid category, such as hydrophobic, the number of candidate triplets to search scales tremendously, as does demands upon memory if a hash table is employed. Third, we restrict triplets to a core of residues. For a structure of 300 residues there are nearly 27 million candidate triplets. Storing all in a hash table is too demanding, and is not necessary either. The best candidate triplets will be clustered in a core, which may be comprised of residues around an active site, ligand binding domain or phosphorylation site. Only searching this core will provide good candidate triplets and minimize time. By default a geometric core of fifty residues is used.

From the input structures a consensus set $S_n$ is sought for of residues fully conserved across all $n$ samples (a full consensus set). This can be done quickly through a series of pair-wise comparisons. First a consensus set $S_2$ of conserved residues is sought for between samples 1 and 2 using the superimposition procedure outlined above. A second set $S_3$ of conserved residues found in $S_2$ and sample 3 is then sought for using the same procedure of selecting triplets from the consensus set and the sample. This process is repeated until a final full consensus set $S_n$ is produced. From experience the procedure for deriving $S_n$ is generally sample order independent, except in some cases where an extreme outlier is present. This is a structure with few of the conserved residues found across the other $n - 1$ samples. In this case the outlier may align well with a set of residues conserved across the first $x$ samples but not across the final $n - x - 1$. By aligning well with a subset of $S_x$ not in $S_n$, the outlier skews the derivation process away from the true set $S_n$. This can be accounted for by leaving such outliers until the end, by

which time the consensus set has been reduced to a true group of highly conserved residues.

The multiple alignment procedure is a modification of the pair-wise superimposition method outlined above. The full consensus set $S_n$ found through the pair-wise superimpositions is comprised of a set of $m$ landmarks that can be used to simultaneously align the group of $n$ structures. First, the conserved residues from each sample $i$ are stored in an $m \times 3$ matrix $X_i$ and Helmertized. The optimal rotation matrix for each structure $\Gamma_i$ is calculated by an iterative process that seeks to minimize the following value:

$$G = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left\| \dot{X}_i \Gamma_i - \dot{X}_j \Gamma_j \right\|^2 . \tag{4}$$

The rotation $\Gamma_i$ is calculated from the singular value decomposition of $\dot{X}_i$ with the mean of the rest, $\overline{X}_i^T \dot{X}_i$, for

$$\overline{X}_i = \frac{1}{n-1} \sum_{j \neq i} \dot{X}_j . \tag{5}$$

However, with each successive $\dot{X}_i$ rotated, the previous $\overline{X}_j$ for $j < i$ will change and the previously calculated rotation of $\dot{X}_j$ will be incorrect. To account for this, after every sample has been rotated $G$ is calculated and the process of rotation is repeated until the change in $G$ falls below some arbitrary small threshold (i.e. convergence occurs). For $p$ steps until convergence, the optimal rotation of each structure is equal to the product of the successive $\Gamma_i$ rotations.

This procedure superimposes the conserved residues found in the set $S_n$. The next goal is to extend this set. Multiple alignment is more accurate than a series of pair-wise

alignments and this allows for $S_n$ to be extended to include additional fully conserved residues. At the same time residues with less than full conservation can be incorporated. We have defined conserved residues as identical or similar amino acids from different structures within two angstroms of some point, termed a consensus point. Consensus points are considered redundant if they contain more than 50% of the same residues. The point closer to the relevant landmarks is retained. Partially conserved residues must only be present in a threshold of samples and the less than full conservation of an amino acid takes precedence over the conservation of a category. Leeway in the conservation of residues allows for evolutionary freedom, but is also required for practical reasons such as x-ray resolution, model construction, side-chain movement and non-significant sample-dependent positionings. As the superimposition procedure requires all landmarks in all samples, any samples with missing landmarks are temporally given the consensus point as a place filler. If new residues can be added following an alignment then the expanded set $S_n$ is used to re-align the structures. The extension and alignment process is repeated until a final consensus set is converged upon or until the completion of a predetermined number of iterations.

*Aligning active-conformation kinases*

Active-conformation kinases were aligned using the above procedure with a congruency criteria of 1% and a distance constraint for conservation of two angstroms. The partially active-conformation structure of the transient receptor potential channel kinase was aligned directly onto the consensus using congruency criteria of 10% and a distance

constraint of 2.25 angstroms.

*Availability and Performance*

The program is written in C and Fortran. Full source code is freely available under the GNU General Public License by contacting the authors. PDB formatted files are required for input. For multiple alignments output consists of PDB files modified according to the alignment, a consensus structure in PDB format, a csv file indicating the conserved and aligned residues, and a script for viewing the alignment and consensus in the open-source program PyMOL (http://www.pymol.org). For a pair-wise alignment against a consensus structure, output is as above except no consensus structure is produced. The alignment and consensus structure of the fifteen kinases presented in the Results section were generated in ~twenty seconds using a single processing core of an AMD Athlon 64 X2 4200 processor with 1GB of RAM. A single pair-wise alignment takes ~2-8 seconds depending upon the degree of similarity between input samples. Comparisons between a sample and a consensus structure are very rapid (<1 second due to the reduced set of residues in a consensus set).

*Structure prediction*

Constraints from fourteen active-conformation kinases were generated for both the side-chain and the α-carbon. Constraints were derived from the final structural alignment. Side-chain constraints were applied to the atom closest to the median of all side-chain atoms. Structure models were created following the Rosetta homology modeling protocol described in Das *et al*.[129] During full-atom refinement of a model, a penalty score is

applied when the atom-atom distances in the model exceed the upper or lower limit of the corresponding distance constraints. If a distance exceeds the upper or lower constraint limit by $d$ Angstrom, then the penalty score $Ec$ is $d * d$ when $d < 0.5$ and $d - 0.25$ when $d \geq 0.5$. For PDK1, the lowest scoring model (in full atom energy) as output by Rosetta was selected as the best model. For Rio2, the lowest 5% of models were used for screening against the consensus structure as described in the Results section. Models were ranked based on their score from Equation 3.

**Acknowledgements**

We thank Miguel A. Andrade for critical reading and discussions during manuscript preparation.

## Supporting Information

**Supplementary Table 2.1**: Conserved residues for constraint modeling.

| | type | ACK1 | Akt2 | CDK2 | CK1 | CK2 | DAPK | IRK | p38γ | PhK | Pim-1 | PKA | PknB | Sky1P | TAO2 | PDK1* | Rio2* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | h | L132 | L158 | I10 | I18 | V45 | L19 | L1002 | V33 | L25 | L44 | L49 | L17 | L164 | I34 | L88 (CG) | x |
| 2 | G | 133 | 159 | 11 | 19 | 46 | 20 | 1003 | x | 26 | 45 | 50 | 18 | 165 | 35 | 89 (CA) | x |
| 3 | G | x | 161 | x | 21 | 48 | 22 | 1005 | x | 28 | x | 52 | 20 | 167 | 37 | 91 (CA) | x |
| 4 | V | 140 | 166 | 18 | I26l | 53 | 27 | 1010 | 41 | 33 | 52 | 57 | 25 | 172 | 42 | 96 (CB) | x |
| 5 | A | 156 | 179 | 31 | 39 | I66l | 40 | 1028 | 54 | 46 | 65 | 70 | 38 | 185 | 55 | 109 (CB) | x |
| 6 | h | V157 | M180 | L32 | I40 | I67 | A41 | V1029 | I55 | V47 | I66 | M71 | V39 | M186 | I56 | I110 (CG1) | x |
| 7 | K | 158 | 181 | 33 | 41 | 68 | 42 | 1030 | 56 | 48 | 67 | 72 | 40 | 187 | 57 | 111 (CD) | x |
| 8 | E | 177 | 200 | 51 | 55 | 81 | 64 | 1047 | 74 | 73 | 89 | 91 | 59 | 202 | 76 | 130 (CD) | 120 (CD) |
| 9 | L | M181h | 204 | 55 | Y59h | 85 | 68 | M1051h | 78 | 77 | 93 | 95 | A63l | 206 | 80 | M134h (CG) | 134/154 (CD) |
| 10 | L | 192 | 215 | 66 | V71l | 97 | 79 | 1062 | 89 | 89 | 106 | 106 | V74l | 228 | Y91h | 145 (CG) | x |
| 11 | h | M203 | F227 | L78 | L83 | L111 | L91 | V1074 | L107 | L101 | L118 | M118 | I90 | M244 | L103 | F157 (CG) | x |
| 12 | V | 204 | 228 | 79 | 84 | I112l | I92l | 1075 | 108 | 102 | I119l | 119 | 91 | 245 | 104 | x | x |
| 13 | E | 206 | 230 | 81 | x | 114 | 94 | 1077 | x | x | 121 | 121 | 93 | 247 | 106 | x | x |
| 14 | h | A208 | A232 | L83 | L88 | V116 | V96 | M1079 | M112 | M106 | P123 | V123 | V95 | L249 | C108 | A162 (CB) | x |
| 15 | L | x | 237 | 87 | 92 | x | 101 | 1084 | 116 | 111 | 129 | M128h | 100 | 253 | A112l | 167 (CG) | x |
| 16 | h | V236 | I259 | L111 | M115 | L140 | I123 | I1116 | M137 | L133 | V151 | I150 | A122 | L277 | A135 | I189 (CG1) | x |
| 17 | h | A237 | V260 | L112 | L116 | L141 | L124 | A1117 | L138 | L134 | L152 | V151 | C123 | L278 | L136 | V190 (CB) | x |
| 18 | h | M240 | L263 | L115 | V119 | L144 | V127 | M1120 | L141 | I137 | V155 | F154 | L126 | L281 | L139 | L193 (CG) | x |
| 19 | H | x | 267 | 119 | 123 | 148 | 131 | x | 145 | 141 | 159 | 158 | 130 | 285 | 143 | 197 (CG) | x |
| 20 | h | F248 | V271 | V123 | L127 | I152 | I135 | F1128 | I149 | I145 | V163 | L162 | I134 | I290 | M147 | I201 (CG1) | x |
| 21 | h | I249 | V272 | L124 | V128 | M153 | A136 | V1129 | I150 | V146 | L164 | I163 | I135 | I291 | I148 | I202 (CG1) | x |
| 22 | H | 250 | Y273r | 125 | Y129r | 154 | 137 | 1130 | 151 | 147 | 165 | Y164r | 136 | 292 | 149 | 203 (CG) | x |
| 23 | R | 251 | 274 | 126 | 130 | 155 | x | 1131 | 152 | 148 | 166 | 165 | 137 | x | 150 | 204 (NE) | x |
| 24 | D | 252 | 275 | 127 | 131 | 156 | 139 | 1132 | 153 | 149 | 167 | 166 | 138 | 294 | 151 | 205 (CG) | 218 (CG) |
| 25 | I | L253 | I276 | L128 | I132 | V157 | L140 | L1133 | L154 | L150 | I168 | L167 | V139 | I295 | V152 | L206 (CG) | x |
| 26 | K | R256b | 277 | 129 | 133 | 158 | 141 | R1136b | 155 | 151 | 169 | 168 | 140 | 296 | 153 | 207 (CD) | x |
| 27 | N | 257 | 280 | 132 | 136 | 161 | 144 | 1137 | 158 | 154 | 172 | 171 | 143 | 299 | 156 | 210 (CG) | 223 (CG) |
| 28 | h | L258 | L281 | L133 | F137 | V162 | I145 | C1138 | L159 | I155 | I173 | L172 | I144 | V300 | I157 | I211 (CG1) | x |
| 29 | h | L259 | M282 | L134 | L138 | M163 | M146 | M1139 | A160 | L156 | L174 | L173 | M145 | L301 | L158 | L212 (CG) | x |
| 30 | h | L260 | L283 | I135 | I139 | I164 | L147 | V1140 | V161 | L157 | I175 | I174 | I146 | M302 | L159 | L213 (CG) | x |
| 31 | I | V266 | I289 | I141 | I150 | L171 | I157 | V1146 | L167 | I163 | L182 | I180 | V152 | I546 | V165 | I219 (CG1) | x |
| 32 | K | 267 | 290 | 142 | Y151p | R172b | 158 | 1147 | 168 | 164 | 183 | Q181p | 153 | 547 | 166 | Q220p (CD) | x |
| 33 | I | I268 | I291 | L143 | V152 | L173 | I159 | I1148 | I169 | L165 | L184 | V182 | V154 | I548 | L167 | I221 (CG1) | 235 (CG) |
| 34 | D | 270 | 293 | 145 | 154 | 175 | 161 | 1150 | 171 | 167 | 186 | 184 | 156 | 550 | 169 | 223 (CG) | x |
| 35 | F | 271 | 294 | 146 | 155 | W176r | 162 | 1151 | 172 | 168 | 187 | 185 | 157 | L551h | 170 | 224 (CG) | x |
| 36 | G | 272 | 295 | 147 | 156 | 177 | 163 | 1152 | 173 | 169 | 188 | 186 | 158 | 552 | 171 | 225 (CA) | x |
| 37 | T | P293s | 313 | 165 | 181 | S194d | 180 | x | 188 | 186 | 204 | 201 | x | 567 | 185 | 245 (CB) | x |
| 38 | r | W296 | Y316 | Y168 | Y184 | F197 | F183 | W1175 | Y191 | Y189 | Y207 | Y204 | Y182 | Y570 | W188 | Y248 (CD2) | x |
| 39 | s | A298 | A318 | A170 | S186 | G199 | A185 | A1177 | A193 | A191 | P209 | A206 | S184 | S572 | A190 | S250 (CB) | x |
| 40 | P | 299 | 319 | 171 | I187h | 200 | 186 | 1178 | 194 | 192 | 210 | 207 | 185 | 573 | 191 | 251 (CG) | x |
| 41 | E | 300 | 320 | 172 | N188p | 201 | 187 | 1179 | 195 | 193 | 211 | 208 | 186 | 574 | 192 | 252 (CD) | x |
| 42 | D | 312 | 332 | 185 | 200 | 214 | 199 | 1191 | 208 | 211 | x | 220 | 198 | 586 | 207 | 264 (CG) | x |
| 43 | W | 314 | 334 | 187 | x | 216 | 201 | 1193 | 210 | 213 | 226 | 222 | Y200r | 588 | 209 | 266 (CD2) | x |
| 44 | S | x | G335v | 188 | A203v | 217 | 202 | 1194 | 211 | 214 | 227 | A223v | 201 | 589 | 210 | A267s (CB) | x |
| 45 | G | 317 | 337 | 190 | 205 | 219 | 204 | 1196 | 213 | 216 | 229 | 225 | 203 | A591v | 212 | 269 (CA) | x |
| 46 | h | M323 | M343 | M196 | F211 | M225 | L210 | I1202 | M219 | L222 | M235 | M231 | V209 | L597 | L218 | L275 (CG) | x |
| 47 | F | x | 350 | 203 | x | 233 | 217 | x | 226 | 229 | 242 | 238 | 216 | 604 | x | 282 (CG) | x |
| 48 | h | C367 | L384 | M266 | Y256 | L304 | L255 | C1245 | M291 | F267 | C270 | L272 | A254 | M686 | C261 | L316 (CG) | x |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 49 | L | W368h | 385 | 267 | M257h | 305 | 256 | W1246h | 292 | 268 | 271 | 273 | 255 | 687 | 262 | 317 (CG) | x |
| 50 | P | 372 | 389 | 271 | x | x | 260 | 1250 | A296s | 272 | 275 | L277h | 259 | 691 | 266 | A321s (CB) | x |
| 51 | R | 375 | 392 | 274 | x | 312 | 263 | 1253 | 299 | 275 | 278 | 280 | 262 | 694 | 269 | 324 (NE) | x |
| 52 | H | x | 406 | 283 | x | 321 | 272 | x | 308 | 284 | 287 | 294 | D272c | 703 | 278 | 339 (CG) | x |

*Residues (and atoms) used in constraint based modeling.

The amino acid or amino-acid category of the conserved residue is listed under type. For each structure the residue identifier corresponding to the conserved point is indicated (listed as x if the point is absent). For category types the amino acid present in each structure is indicated before the identifier. If a sample is missing a conserved amino acid but has a similar residue in the same location then the shared category is listed after the identifier. a, acidic; l, aliphatic; r, aromatic; b, basic; c, charged; h, hydrophobic; p, polar; s, small; v, very small.

# CHAPTER 3: The role of conserved water molecules in the catalytic domain of protein kinases

**The role of conserved water molecules in the catalytic domain of protein kinases**

James D. R. Knight[1,2,3], Donald Hamelberg[4], J. Andrew McCammon[5] and Rashmi Kothary[1,2,3,6]

[1]Ottawa Hospital Research Institute, Ottawa, Ontario, Canada K1H 8L6

[2]The University of Ottawa Centre for Neuromuscular Disease, Ottawa, Ontario, Canada K1H 8M5

[3]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

[4]Department of Chemistry, Georgia State University, Atlanta, Georgia 30302-4098

[5]Department of Chemistry and Biochemistry and Department of Pharmacology, Center for Theoretical Biological Physics, Howard Hughes Medical Institute, University of California at San Diego, La Jolla, California 92093

[6]Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

**Published in *Proteins: Structure, Function, and Bioinformatics*, 2009, 76(3):527-35.**

**Author Contributions**

JDRK and RK conceived of and designed the project. DH performed molecular dynamics simulations and assisted JDRK with relevant data analysis. All other experiments and analysis were performed by JDRK. JAM supervised DH and advised on data interpretation. JDRK wrote the paper and RK revised and edited it. RK supervised JDRK.

**Abstract**

Protein kinases are essential signaling molecules with a characteristic bilobal shape that has been studied for over fifteen years. Despite the number of crystal structures available, little study has been directed away from the prototypical functional elements of the kinase domain. We have performed a structural alignment of thirteen active-conformation kinases and discovered the presence of six water molecules that occur in conserved locations across this group of diverse kinases. Molecular dynamics simulations demonstrated that these waters confer a great deal of stability to their local environment and to a key catalytic residue. Our results highlight the importance of novel elements within the greater kinase family and suggest that conserved water molecules are necessary for efficient kinase function.

## Introduction

In protein kinases, phosphotransfer capability requires at least two well defined conserved elements in addition to the polypeptide: an active site metal cofactor, usually magnesium or manganese, and ATP. The only element found in crystallography data with potential conservation between kinases that has not been examined in detail is water. It was first noted by Shaltiel, Cox and Taylor[143] that different crystal structures of protein kinase A (PKA) had water molecules in virtually identical locations. They noted several waters within the active site, interacting with both conserved and non-conserved residues, as well as ATP and the active site manganese, and the authors suggested a critical role for these waters in the efficient function of PKA.

That water molecules play critical roles in enzymes/proteins has been known for some time. Condensation and hydrolysis reactions are two clear examples of water functioning as a crucial molecular element – a product or reactant. In chemical reactions water molecules can also function as transition state intermediates. Another role for water at the molecular level is as a structural element, interconnecting the protein through hydrogen bonds and maintaining/stabilizing the positions of residues and/or the fold[144]. These structural waters are often conserved within a protein family, *i.e.* they perform the same functions and occur in nearly identical three-dimensional locations with reference to their associated structures. Structurally conserved waters have been found in several classes of proteins, including ribonuclease T1[145, 146], serine proteases[147], Rossmann fold dinucleotide binding proteins[148], MHC class 1 proteins[149] and aspartic proteinases[150]. In these studies, conserved waters are often found in or near an enzyme's active site, suggesting they play an important function in active site stability, flexibility, ligand

coordination and residue positioning, hence their evolutionary conservation. Rodríguez-Almazán and colleagues[151] have shown recently how even a conservative amino acid change can disrupt conserved water molecules and the connective networks they form, drastically altering enzyme function.

Protein kinases can be divided into two interdependent functional regions: the ATP binding/phosphotransfer region, and the second where substrate binding occurs. These regions are obviously interconnected but they can be discussed separately. The area where ATP is bound, and the tertiary phosphate positioned for transfer, is bordered principally by the 5-stranded β sheet, helix C, the magnesium-positioning loop and the catalytic loop, with the area between them forming the active site (**Figure 3.1**). It is in the immediate area of the transferable phosphate where the greatest conservation is found[152, 153]. Six fully conserved residues play very important roles in positioning this key element, all of which are absolutely critical for efficient function[154]. The ATP-positioning lysine is located on β–strand 3 and interacts with the α- and β- phosphates of ATP, and this lysine is critically stabilized by a glutamic acid originating from helix C. The catalytic aspartic acid that initiates phosphotransfer is located on the catalytic loop directly beneath the γ-phosphate. The γ-phosphate itself is directly and indirectly coordinated by three residues and one or two divalent cations. The magnesium-positioning aspartic acid arises from the magnesium-positioning loop and indirectly positions the γ-phosphate through a magnesium or manganese ion. Similarly, an asparagine residue located on the catalytic loop also indirectly positions this phosphate through a secondary metal ion. And the last fully conserved residue is a lysine located just downstream of the catalytic residue that directly aids in positioning the γ-phosphate, although this residue is exchanged for an

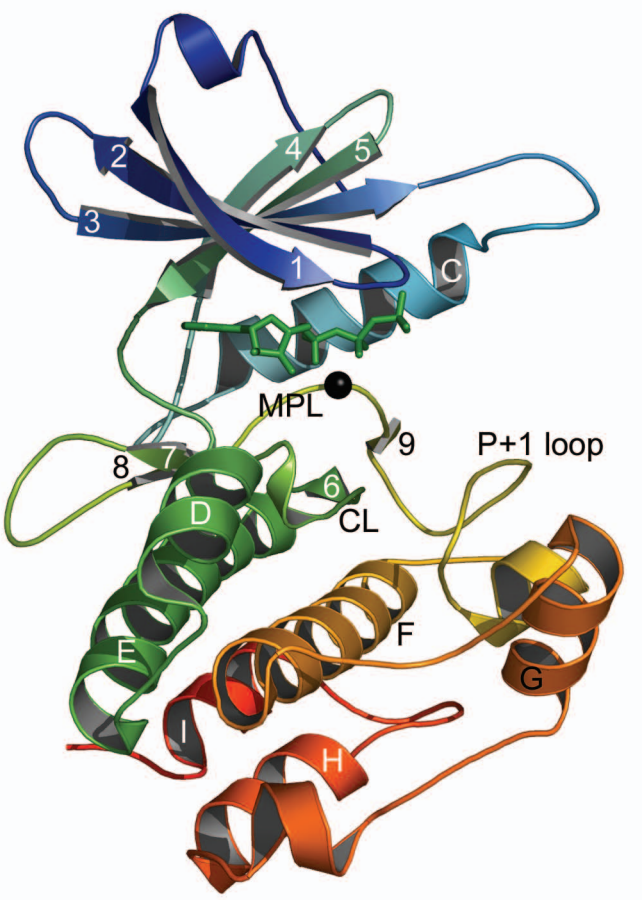



**Figure 3.1**. The crystal structure of the active conformation death-associated protein kinase[105]. β-strands, helices and prominent loops are labelled in accordance with Knighton, *et al*.[45] The non-hydrolyzable ATP analog ANP is shown in green sticks and manganese ion as a black sphere. The small lobe comprises the β-sheet, helix C and the interveening loops. The remainder of the protein, below ANP, constitutes the large lobe. CL: catalytic loop; MPL: magnesium positioning loop.

arginine in tyrosine kinases. While the area of phosphotransfer is relatively well-studied and understood, in contrast the substrate binding region is not. Substrates are bound within a cleft formed from the catalytic loop, the P+1 loop, helix D, and parts of helices F and H. There is little residue conservation within this area[152], which is entirely consistent with its function in substrate binding. However, understanding this area is critical as it mediates substrate binding and contains the information that confers substrate specificity. In neither the substrate binding groove, nor the well-studied active site, has the role of water molecules and the conserved functional roles they might play been examined.

Here we report on an examination of water conservation across a diverse array of kinases from a variety of species to better establish and elucidate the importance of individual water molecules for kinase function. In addition to an analysis of crystallography data, we have performed molecular dynamics simulations to examine the potential roles of conserved waters within the protein kinase domain. Our results reveal the presence of six conserved water molecules in protein kinases. Two waters are located within the heart of the active site, interacting with some of the aforementioned key functional residues. Both of these waters have large free energy benefits and simulations demonstrate how one aids in coordinating the absolutely critical magnesium-positioning aspartic acid. We also discovered a collection of dynamic yet conserved waters trapped within a closed pocket that lend dramatic stability to the substrate binding cleft. And there are a further two water molecules for which we cannot postulate clear reasons for their conservation. In all, our results suggest that conserved waters are crucial to the efficient function of the kinase family and reveal in general the different circumstances and behaviors associated with water molecules in proteins.

## Methods

### *Data Set*

Kinase structures representing active conformations were retrieved from the Protein Data Bank. A kinase was considered "active" if it had any necessary phosphorylations, ATP or a non-hydrolyzable ATP analog, and had at least one divalent cation in the active site. Crystal structures with resolutions poorer than the diameter of water (2.8 Å) were rejected. Thirteen structures met these criteria (**Table 3.1**).

### *Structural Alignment and Conserved Water Identification*

Structures were aligned using a sequence order independent structural alignment algorithm we have developed[152]. After alignment, water molecules were sought for in identical locations between structures. For each structure all water molecules within 3.2 Å of a nitrogen, oxygen or sulfur atom were included in the search. Waters from different structures that fell within a sphere of 2 Å radius were deemed to occupy the same location. Only waters found in at least ten of the thirteen structures were considered conserved. Normalized B-factors were calculated as $(B_i-\langle B\rangle)/\sigma(B)$, where $B_i$ is the B-factor of water i, $\langle B\rangle$ the mean of all waters, and $\sigma(B)$ the standard deviation. ASA was calculated with the program GETAREA 1.1[155] using a probe radius of 1.4. Internal water surfaces were defined using the CASTp server with default parameters[156]. A residue was considered fully conserved if an identical or similar residue could be found in all structures within a sphere of 2 Å radius. Partially conserved residues need only be found in ten of thirteen structures. Conserved categories must be present in all structures (see **Supplementary Table 3.1**).

**Table 3.1**. Protein kinase structures in active conformations. The number of water molecules in each structure within 3.2 Å of a nitrogen, oxygen or sulpfur atom is listed in the final column.

| Kinase | Full Name | PDB Code | Species | Resolution (Å) | No. Waters |
|---|---|---|---|---|---|
| Akt2 | RAC-β serine/threonine-protein kinase | 1O6K | *H. sapiens* | 1.7 | 224 |
| CDK2 | cell division protein kinase 2 | 1JST | *H. sapiens* | 2.6 | 31 |
| CK1 | casein kinase I | 1CSN | *S. pombe* | 2.0 | 83 |
| CK2 | casein kinase II subunit α | 1LP4 | *Z. mays* | 1.86 | 174 |
| DAPK | death-associated protein kinase | 1IG1 | *H. sapiens* | 1.8 | 190 |
| IRK | insulin receptor tyrosine kinase | 1IR3 | *H. sapiens* | 1.9 | 160 |
| MAPK p38-γ | mitogen-activated protein kinase p38-γgamma | 1CM8 | *H. sapiens* | 2.4 | 62 |
| PhK | phosphorylase kinase | 1PHK | *O. cuniculus* | 2.2 | 86 |
| Pim-1 | proto-oncogene serine/threonine-protein kinase Pim-1 | 1XR1 | *H. sapiens* | 2.1 | 21 |
| PKA | protein kinase A | 1ATP | *M. musculus* | 2.2 | 68 |
| Rio2 | Rio2 serine kinase | 1ZAO | *A. fulgidus* | 1.84 | 223 |
| Sky1P | SR protein kinase | 1Q97 | *S. cerevisiae* | 2.3 | 57 |
| TAO2 | thousand and one amino acid protein 2 | 1U5R | *R. norvegicus* | 2.1 | 152 |

***Molecular Dynamics***

The coordinates of the initial structure used in this study was the crystal structure of protein kinase A (1ATP)[110] obtained from the protein data bank (PDB). The hydrogens were added to the heavy atoms using the Leap module in the AMBER 8[157] suite of programs. Furthermore, the atomic charges and force field parameters of the bound ATP were obtained from the work of Meagher *et al*.[158] The complex was solvated with TIP3P[159] water molecules such that the solvent was placed up to about 10 Å away from the protein complex to fill a truncated octahedron box. Three chloride ions were placed around the complex using the Leap module in AMBER 8 to obtain electrostatic neutrality. All molecular dynamics simulations were performed with the sander module in AMBER 8.

The system was equilibrated by using a multistage equilibration protocol. At the start of the equilibration, 100 kcal/mol/Å$^2$ harmonic constraints were placed only on the complex. The water and ions were minimized for 1000 steps. This was followed by another 1000 step minimization on the entire system with a 50 kcal/mol/Å$^2$ constraint placed on the complex. With 50 kcal/mol/Å$^2$ constraint on the complex, the entire system was heated from 100 to 300 K over 50 picoseconds. Finally, the entire system was simulated for another 50 picoseconds without any constraints applied at 300 K. After equilibration, the final equilibrated structure was used to carry out 10 nanoseconds MD simulations in the NPT ensemble with periodic boundary conditions at a constant temperature of 300 K and pressure of 1 bar. The SHAKE algorithm[160] was applied to all bonds involving hydrogen atoms and an integration time step of 2.0 femtoseconds (fs) was used in solving Newton's equation of motion. Long-range electrostatic interactions

were evaluated using the particle mesh Ewald (PME) summation method[161]. The non-bonded interactions were subjected to a 9 Å cutoff that distinguished between the direct space and the reciprocal space in PME. The MD trajectories were collected at every one picosecond (ps) time interval. The resulting trajectories were used to calculate the atomic positional fluctuation of the oxygen atom of the bound water molecules in order to estimate the harmonic force constants used in the free energy simulations, as previously described[162].

***Free Energy Calculations***

The free energy calculations were performed using the double-decoupling method, as previously described[162]. The double-decoupling free energy simulations involve gradually turning off the electrostatic and van der Waals interactions of the bound water molecule from the rest of the system. This involves two sets of simulations: the transfer of the water molecule from bulk water to the gas phase, and transferring the water molecule from the binding pocket of the protein complex to the gas phase during which the water molecule is always constrained to occupy the binding site as defined by the coordinate system of the complex. To perform the change from $H_2O_{sol}$ to $H_2O_{gas}$, the energy term is progressively mapped from U(sol) to U(gas) along a chosen path, where U is the potential energy function. This chosen path is mapped as a function of a coupling parameter λ that varies from 0 to 1, where U can be written in terms of λ as $U(\lambda) = (1- \lambda)^k U(0) + [1-(1-\lambda)^k]U(1)$, and $U(0) = U(sol)$ and $U(1) = U(gas)$. $k = 1$ when decoupling the electrostatic interactions and $k = 6$ when decoupling the van der Waals interactions. The free energy difference between the two states is calculated by the thermodynamic integration

approach using 12-points Gaussian quadrature integration at discrete points of $\lambda_i$ of 0.00922, 0.04794, 0.11505, 0.20634, 0.31608, 0.43738, 0.56262, 0.68392, 0.79366, 0.88495, 0.95206, and 0.99078 along the path. The simulation for each value of $\lambda$ was initially equilibrated for 200.0 picoseconds, and the data sampling was also performed for 200.0 picoseconds after the equilibration. All of the free energy simulations were performed under the same conditions as described above for the molecular dynamics simulation, except for the fact that a time step of 1.0 fs, instead of 2.0 fs, was used when evaluating the free energy component of the van der Waals interactions. Three independent runs were performed for each free energy calculation. The final free energy was then estimated by averaging over the calculated free energies from the three independent simulations. Also the error was estimated by calculating the standard error using the standard deviation of the calculated free energies from the three independent simulations.

## Results and Discussion

### *Conserved Waters*

We present our results with reference to the death associated protein kinase (DAPK) and PKA. The structure of DAPK is shown in **Figure 3.1** for purposes of orientation. Thirteen kinases in active conformations (**Table 3.1**) were aligned (**Figure 3.2(a)**) and six conserved water molecules identified (**Figure 3.2(b)** and **Table 3.2**). These waters have low average B-factors, reflecting a general lack of mobility in these molecules and confidence in their placement. These waters bind both to side-chain and main-chain atoms (**Supplementary Table 3.2**). All but ten of the fifty-seven side-chain interactions are with conserved residues. Water interactions with main-chain atoms occur evenly between conserved and non-conserved residues. Between structures, water interactions with main-chain residues are entirely with equivalent parts of the polypeptide. Waters $A_w$, $C_w$, $D_w$, and $F_w$ have low accessible surface areas.

Of the waters consistently found by Shaltiel, Cox and Taylor[143] across several different crystal structures of PKA, we also found ASCW B (their nomenclature) as conserved across the protein kinase family (our $A_w$), as is ASCW C (our $C_w$). Although they only reported these as potentially interacting with single atoms, both can interact with two (using our distance cut-off, which was also slightly stricter). Interestingly, unlike these authors, we find no conserved waters interacting with either ATP or magnesium/manganese. The functions of these waters in PKA are not likely to be substituted for by amino-acid residues in other kinases, as the region around the γ-phosphate is a very highly conserved area. However, while the active-site divalent cations necessary for catalytic activity are consistently placed throughout the structures we have

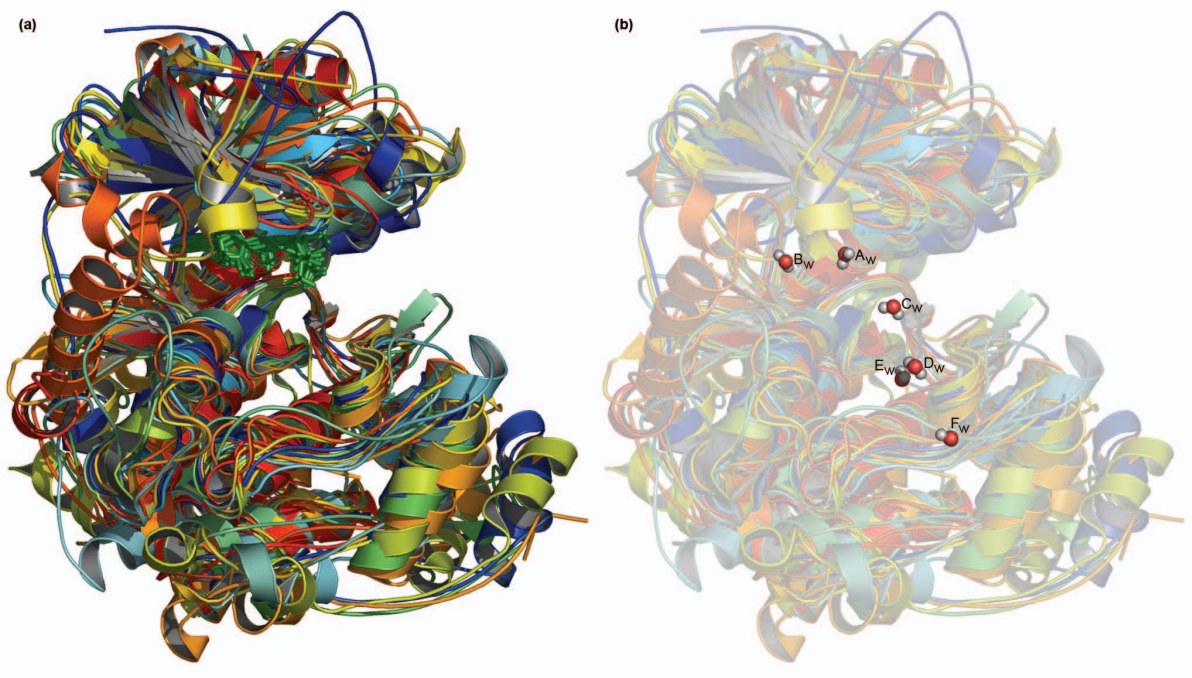



**Figure 3.2**. An active conformation protein kinase structural alignment and consensus. (**a**) The structural alignment of the thirteen kinases listed in **Table 3.1**. The ATP or non-hydrolyzable ATP analog of each structure is shown in green sticks. Each kinase is colored uniquely. (**b**) The conserved water molecules are shown superimposed above the alignment, coloured as red spheres with attached hydrogens in white. Waters are labelled as in **Table 3.2**.

**Table 3.1**. Conserved waters. Identifiers from each structure are listed, along with the normalized B-factors and the acessible surface area (ASA) in $Å^2$.

| | Kinase | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Water | Akt2 | CDK2 | CK1 | CK2 | DAPK | IRK | p38-γ | PhK | Pim-1 | PKA | Rio2 | Sky1P | TAO2 |
| $A_w$ | 31 | 5 | 411 | 20 | 3405 | 44 | 2006 | 401 | | 400 | 11 | | 44 |
| B | -1.207 | 0.191 | -1.483 | -1.230 | -1.402 | -0.834 | -0.786 | -2.053 | - | -0.081 | -1.981 | - | -1.074 |
| ASA | 0.05 | 3.04 | 3.21 | 0.58 | 0.30 | 1.25 | 1.00 | 0.93 | | 0.65 | 1.04 | | 1.30 |
| $B_w$ | 34 | 3 | 409 | | 428 | 45 | 2106 | 426 | | 404 | | 91 | 53 |
| B | -1.394 | -0.279 | -1.031 | - | -0.475 | -0.557 | -0.304 | -1.338 | - | 0.070 | - | -1.058 | -0.538 |
| ASA | 0.55 | 0.88 | 0.82 | | 8.09 | 1.61 | 2.39 | 0.00 | | 9.18 | | 0.00 | 10.52 |
| $C_w$ | 114 | 8 | 441 | 10 | 413 | 29 | 2042 | 405 | 25 | 403 | | | 21 |
| B | -1.433 | -0.025 | -1.394 | -1.548 | -1.351 | -0.844 | -0.603 | -1.360 | 0.302 | -1.324 | - | - | -1.510 |
| ASA | 0.15 | 0.04 | 0.25 | 0.00 | 0.00 | 0.07 | 0.06 | 0.14 | 0.24 | 0.09 | | | 0.00 |
| $D_w$ | 97 | | 427 | 23 | 435 | 7 | 2124 | | 14 | 410 | 110 | | 29 |
| B | -1.325 | - | -0.951 | -1.464 | -1.017 | -1.453 | 1.510 | - | 0.206 | -0.857 | -0.573 | - | -1.567 |
| ASA | 1.53 | | 1.25 | 0.00 | 0.79 | 0.28 | 0.00 | | 0.96 | 1.31 | 3.90 | | 0.96 |
| $E_w$ | 98 | 6 | 444 | 11 | 404 | 1 | 2136 | 415 | 8 | 431 | | 8 | 1 |
| B | -1.141 | -0.788 | -0.197 | -1.672 | -1.604 | -1.940 | 0.440 | -1.687 | 0.738 | -1.406 | - | -1.830 | -2.098 |
| ASA | 4.81 | 0.00 | 2.43 | 0.00 | 1.21 | 1.72 | 0.00 | 1.32 | 5.27 | 3.08 | | 0.00 | 1.18 |
| $F_w$ | 153 | 13 | | 65 | 403 | 46 | 2015 | 413 | 24 | 568 | 132 | 110 | 2 |
| B | -1.819 | -2.526 | - | -0.978 | -1.684 | -0.827 | 0.633 | -1.669 | -0.131 | -0.503 | 0.201 | -1.384 | -2.025 |
| ASA | 0.00 | 0.00 | | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 24.43 | 0.00 | 1.07 |

used, there is a moderate degree of variability in the positioning of the γ-phosphate. We do not know what to attribute this to but variability in the position of this phosphate would be accompanied by variability in the local water environment, and this would manifest as a lack of water molecule conservation in this area. Whether this lack of conservation is real and in contradiction to the findings of Shaltiel, Cox and Taylor[143], or instead is an artifact of the data set we are examining, we cannot say. When more active-conformation kinase structures become available this issue can be better answered.

The conserved water molecules from DAPK are shown in **Figure 3.3** as an example of the typical interactions that are occurring. Water molecule $A_w$ (**Figure 3.3(a)**) interacts with the fully conserved glutamic acid E64 of helix C and connects it to the main chain of the magnesium positioning loop (**Supplementary Table 3.1** lists conserved residues). Molecule $B_w$ (**Figure 3.3(b)**) can potentially interact with three atoms, connecting the terminus of helix C with the end of the loop preceding β-strand 4. Molecule $C_w$ (**Figure 3.3(a)**) can interconnect the magnesium positioning loop through three main-chain atoms and join these to the main chain of the catalytic loop. Molecule $D_w$ (**Figure 3.3(c)**) is only bound to the structure in one place, through the carbonyl of the residue following the catalytic base D139. Molecule $E_w$ (**Figure 3.3(c)**) can interact with the side chains of two partially conserved residues from helix F and connect these to the main chain of the catalytic loop. Lastly, water molecule $F_w$ (**Figure 3.3(d)**) can connect the side chain of the partially conserved tryptophan W201 from helix F with two main-chain atoms of the P+1 loop.

From these results we hypothesized that these water molecules may be conserved because of the connectivity they create within the protein kinase domain, and hence the

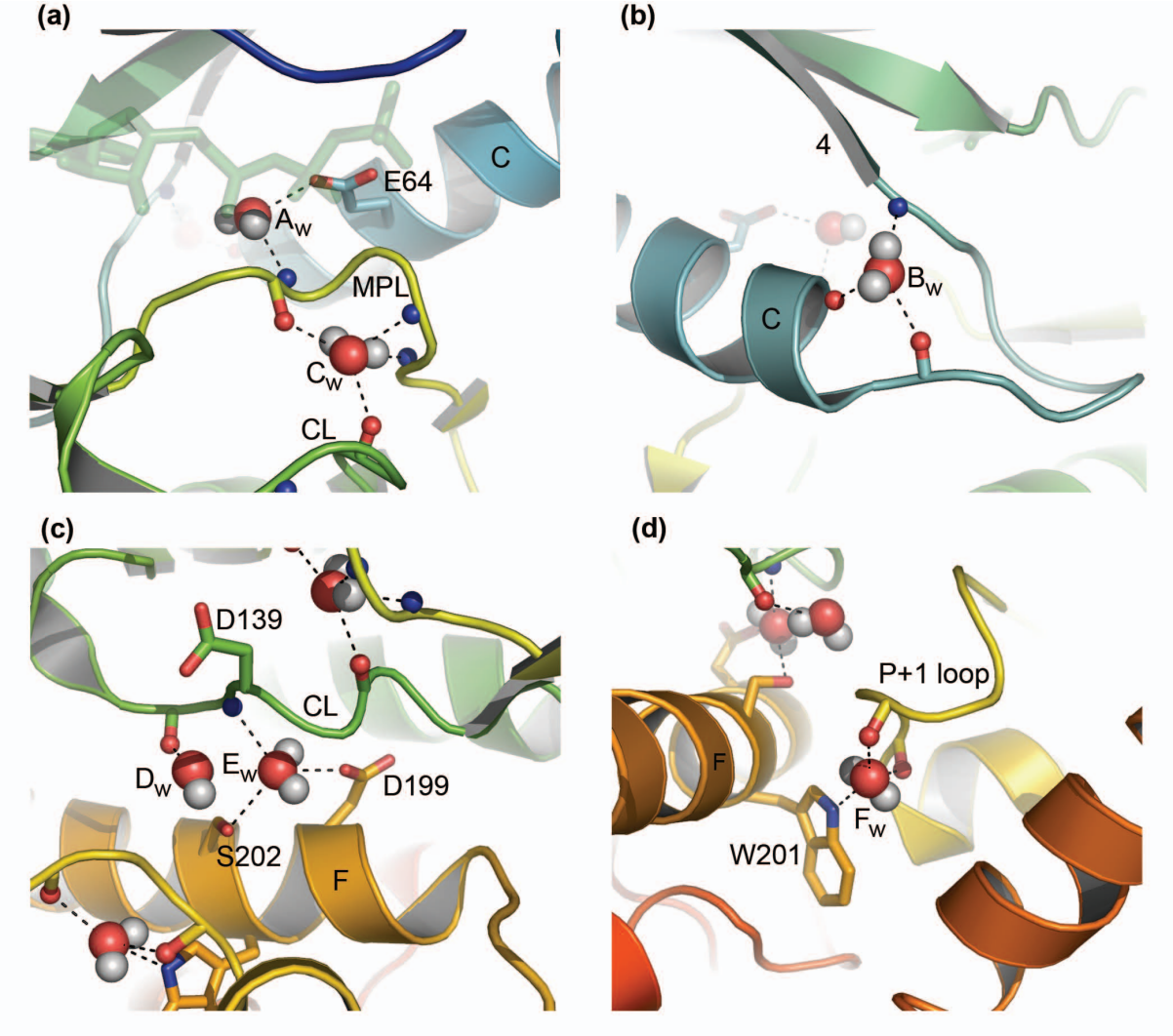



**Figure 3.3**. Conserved waters in DAPK. Each of the conserved waters is labelled as in **Table 3.2**. Dashes indicate potential hydrogen bonds to oxygen and nitrogen atoms within 3.2 Å. Oxygens are colored red, nitrogens blue and hydrogens white. Side chains are shown as sticks, as is the bond between the main-chain carbonyl atoms. CL: catalytic loop; MPL: magnesium positioning loop.

stability they lend to the fold, and/or because of the roles they play in positioning/stabilizing key residues crucial to the efficient function of these enzymes.

### *Molecular Dynamics and Free Energy*

To further elucidate the functions of these water molecules and probe their apparent stabilizing role, molecular dynamics simulations were performed and free energies calculated for the structure of protein kinase A (PKA).

#### *Water $A_w$*

This water is internal and does not exchange with bulk solvent during the time course of simulation. It is fixed within its pocket, with an average fluctuation in position of 1.03 Å. Its free energy contribution is highly favorable at -2.1±0.4 kcal/mol. We also performed simulations to observe the effects of this water molecule on nearby atoms. Simulations were performed with and without this water, and the average positional fluctuation of all atoms from adjacent residues were measured (**Figure 3.4**). The conserved leucine of helix C (PKA: 95, DAPK: 68) and the conserved phenylalanine of the magnesium positioning loop (PKA: 185, DAPK:162) show increased stability in the absence of this water. This is not unexpected as these are hydrophobic residues. In the absence of this water the fully conserved and functionally critical aspartic acid found on the magnesium positioning loop (PKA: 184, DAPK: 161) shows a great increase in fluctuation. This aspartic acid chelates the active site magnesium/manganese, thereby aiding in its positioning and the orientation of the γ-phosphate. A lack of stability in this residue is likely to detrimentally affect catalysis.

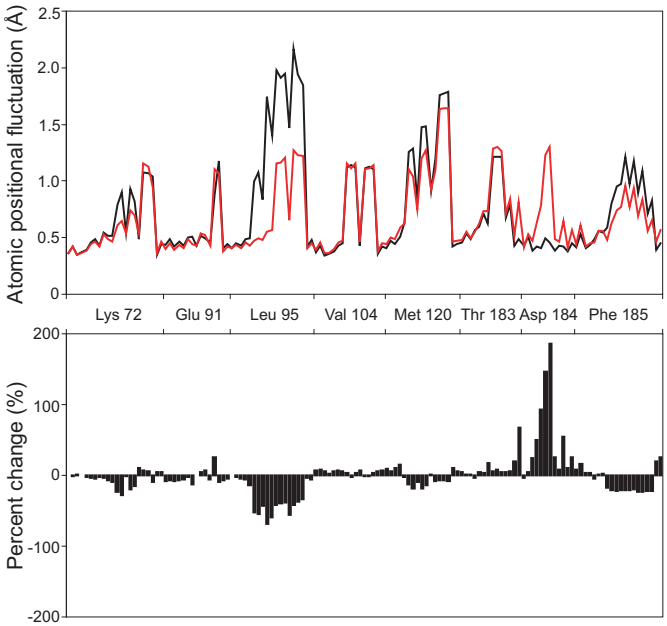



**Figure 3.4**. Water $A_w$ mediated stabilization/destabilization. Molecular dynamics simulations were performed on protein kinase A in the presence (black) and absence (red) of this water. The root-mean-square fluctuation of all atoms in the vicinty of this water is shown in the top panel and the percent difference beneath.

*Water $B_w$*

Water $B_w$ is found on the surface of the protein. It is accessible to and exchanges with bulk solvent. The occupancy for the site this water is found in is ~10%. The residence time for waters in this site is ~2.5 picoseconds. These results suggest this water is not playing an obvious beneficial role in so far as regional stability is concerned, however we cannot rule out that it may be positively involved in the "fluidity"/movement of its local environment, hence its loosely bound nature and simultaneous conservation.

*Water $C_w$*

This water is internal and does not exchange with bulk solvent. Its atomic positional fluctuation is only ~0.78 Å. In other proteins in our dataset (apart from PKA) this water is not adjacent to any other waters but only contacts atoms of the polypeptide. This is not the case in PKA, where it is found in contact with a second water. This creates unfavorable conditions, resulting in a free energy contribution of 0.4±1.6 kcal/mol. However, we believe the second water (424) to be erroneously assigned, as it displays a large overlap with the carbonyl oxygen of tyrosine 164 and has a high average B-factor. When the second water is removed, the free energy contribution of water $C_w$ is very favorable: -5.1±0.7 kcal/mol. This significant lowering of free energy is of obvious benefit in a region so crucial to function as the catalytic loop.

*Waters $D_w$ and $E_w$*

These two waters are found in a pocket that contains four additional waters and will be discussed together. This pocket is found in all structures and contains between two and

six water molecules. Only those found in the positions of $D_w$ and $E_w$ are consistently placed. In the first 1-2 nanoseconds of the simulations two water molecules of PKA exit through a mouth that opens between three hydrophobic residues (PKA: leucine 198, leucine 205 and isoleucine 209). The mouth then closes trapping the remaining four waters. The waters in this pocket do not exchange with bulk solvent but do exchange with one another. Their average positional fluctuation is ~2.30 Å. The site occupied by water $D_w$ is occupied ~19% of time, while site $E_w$ is occupied ~54% of the time. The respective residence times of waters in these sites are 27.8 picoseconds and 46.6 picoseconds. The free energy contribution of adding one water into this pocket is -4.1±1.2 kcal/mol, followed by a second -3.8±1.4 kcal/mol, third -2.6±1.8 kcal/mol and fourth -3.8±1.9 kcal/mol. Their presence obviously lends a great deal of stability.

In several of the structures, water $D_w$ interacts with a carboxyl oxygen from the catalytic aspartic acid and a main-chain carbonyl oxygen from the adjacent downstream residue. From the orientations of these atoms in crystals it appears that this water may aid in positioning the catalytic aspartic acid. However, the crystal structure orientation of the carbonyl is strained and high energy, and during the equilibration stage of the simulations this functional group flips inwards towards the active site and does not change position.

An important point emerges from consideration of waters $D_w$ and $E_w$: none of the water molecules we have studied is permanently trapped, nor are hydrogen bonds made permanent. Water molecules, even those found in internal pockets, will exchange with bulk solvent. This occurs over nanosecond to microsecond timescales[163], outside of the range of our simulations, and results from large-scale fluctuations in structure[164]. Exchange between water molecules in deep pockets can be rapid and have a small

enthalpic cost when the water is weakly hydrogen bonded[163]. Hydrogen bonds must obviously be broken prior to exchange. But even when a water molecule is present in a closed pocket and in contact with the polypeptide, they often have great rotational freedom[163]. Interactions between such waters and the polypeptide should be viewed as transient, not permanent. These points do not contradict the notion of water-molecule mediated stability. Although on/off, hydrogen bonds between water and other functional groups may only need to exist for a continuous series of brief instants to stabilize and orient atoms of the polypeptide. What we see in crystal structures, and even to a degree in molecular dynamics simulations, are the average or end effects of a water molecule's presence. Differences between waters, such as rate of exchange and occupancy, should be considered against one another rather than on their own when assessing any particular water's functional role.

*Water $F_w$*

This water is internal and does not exchange with bulk solvent. It has the most restricted motion at ~0.53 Å. Its free energy contribution is -0.4±0.7 kcal/mol. As with water $A_w$ the dynamics of adjacent residues was also examined in the presence and absence of this water. However, all atoms with the exception of one showed an increase or decrease in atomic positional fluctuation of less than 10%. The exception showed an ~25% decrease in fluctuation. The location of this water in the substrate binding groove may suggest that its interaction with adjacent residues/regions participates beneficially in this process, facilitating residue movement or stability on substrate binding, or assisting in large-scale domain motions. We examined the apo(free) enzyme state and found a negative free

energy for this water (-2.9±0.5), alternatively suggesting it may play a beneficial role in stabilizing this region prior to ATP or substrate binding. As we are only examining short timescales of a single context in our simulations, an understandable solution may exist in another unexplored context.

## Conclusion

Our examination of active-conformation protein kinases has revealed the presence of six conserved water molecules that create a high degree of inter-protein connectivity. Two of these waters are located adjacent to the magnesium-positioning and catalytic loops, and three more are located within the relatively unstudied substrate binding groove. Molecular dynamics simulations on these waters and their local environments suggest that one water molecule ($A_w$) plays a key role in stabilizing an important catalytic residue and that (very) favorable free energy benefits (and hence stability) are conferred by five of the six waters we uncovered. In our study, if we consider conserved waters against conserved amino acids in protein kinases, then we find water is more abundant than all of the twenty amino acids except leucine. The lack of reported proteins with conserved water molecules is striking in this regard. Although the waters we have discovered are not directly involved in the phosphotransfer mechanism, and hence cannot be considered cofactors or ligands, our results suggest they may be essential structural and functional elements necessary for the efficient function of protein kinases.

**Acknowledgements**

We wish to thank Dr. Lynn Megeney for critical reading of the manuscript. J.D.R.K. was supported by a Canadian Institutes of Health Research (CIHR) Canadian Graduate Scholarship and a Multiple Sclerosis Society of Canada (MSSC) Studentship. D.H. was supported in part by the Georgia Cancer Coalition (GCC). J.A.M. was supported in part by the NSF, NIH, HHMI, NBCR, CTBP, SDSC and Accelrys. This work was supported by grants from CIHR and the MSSC to R.K.

**Supporting Information**

**Supplementary Table 3.1**. Kinase consensus residues. The thirteen kinases listed in **Table 3.1** were strucutrally aligned. The amino acid or amino acid category of each consensus residue is listed under "type". Under each kinase is listed the residue identifier corresponding to the consensus. For conserved amino acid categories the specific amino acid found in each kinase is listed before the identifier. Samples missing a consensus residue have an "x". Those samples that have a similar residue in place of a conserved amino acid have the shared category listed after the identifier. b, basic; h, hydrophobic; l, aliphatic; p, polar; r, aromatic; s, small; v, very small.

| | | Kinase | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Type | Akt2 | CDK2 | CK1 | CK2 | DAPK | IRK | p38-γ | PhK | Pim-1 | PKA | Rio2 | Sky1P | TAO2 |
| 1 | h | L158 | I10 | I18 | V45 | L19 | L1002 | V33 | L25 | L44 | L49 | M98 | L164 | I34 |
| 2 | G | 159 | 11 | 19 | 46 | 20 | 1003 | x | 26 | 45 | 50 | x | 165 | 35 |
| 3 | G | 161 | 13 | 21 | 48 | 22 | 1005 | x | 28 | x | 52 | x | 167 | 37 |
| 4 | V | 166 | 18 | I26l | 53 | 27 | 1010 | 41 | 33 | 52 | 57 | 106 | 172 | 42 |
| 5 | A | 179 | 31 | 39 | I66l | 40 | 1028 | 54 | 46 | 65 | 70 | V118l | 185 | 55 |
| 6 | K | 181 | 33 | 41 | 68 | 42 | 1030 | 56 | 48 | 67 | 72 | 120 | 187 | 57 |
| 7 | E | 200 | 51 | 55 | 81 | 64 | 1047 | 74 | 73 | 89 | 91 | 154 | 202 | 76 |
| 8 | L | 204 | 55 | Y59h | 85 | 68 | M1051h | 78 | 77 | 93 | 95 | 158 | 206 | 80 |
| 9 | L | 215 | 66 | V71l | 97 | 79 | 1062 | 89 | 89 | 106 | 106 | V169l | 228 | Y91h |
| 10 | h | F227 | L78 | L83 | L111 | L91 | V1074 | L107 | L101 | L118 | M118 | V177 | M244 | L103 |
| 11 | l | V228 | V79 | V84 | I112 | I92 | V1075 | V108 | V102 | I119 | V119 | L178 | V245 | V104 |
| 12 | h | M229 | F80 | I85 | F113 | L93 | M1076 | M109 | F103 | L120 | M120 | M179 | F246 | M105 |
| 13 | E | 230 | 81 | x | 114 | 94 | 1077 | x | x | 121 | 121 | 180 | 247 | 106 |
| 14 | h | A232 | L83 | L88 | V116 | V96 | M1079 | M112 | M106 | P123 | V123 | I182 | L249 | C108 |
| 15 | L | 237 | 87 | 92 | x | 101 | 1084 | 116 | 111 | 129 | M128h | 187 | 253 | A112l |
| 16 | h | F247 | I99 | F103 | L128 | L111 | P1104 | L125 | L121 | L139 | F138 | V192 | I265 | L123 |
| 17 | h | I259 | L111 | M115 | L140 | I123 | I1116 | M137 | L133 | V151 | I150 | I202 | L277 | A135 |
| 18 | h | L263 | L115 | V119 | L144 | V127 | M1120 | L141 | I137 | V155 | F154 | V206 | L281 | L139 |
| 19 | h | L266 | C118 | I122 | C147 | L130 | L1123 | I144 | L140 | C158 | L157 | F209 | M284 | L142 |
| 20 | H | 267 | 119 | 123 | 148 | 131 | x | 145 | 141 | 159 | 158 | x | 285 | 143 |
| 21 | h | V271 | V123 | L127 | I152 | I135 | F1128 | I149 | I145 | V163 | L162 | I214 | I290 | M147 |
| 22 | h | V272 | L124 | V128 | M153 | A136 | V1129 | I150 | V146 | L164 | I163 | V215 | I291 | I148 |
| 23 | H | Y273r | 125 | Y129r | 154 | 137 | 1130 | 151 | 147 | 165 | Y164r | 216 | 292 | 149 |
| 24 | R | 274 | 126 | 130 | 155 | x | 1131 | 152 | 148 | 166 | 165 | x | x | 150 |
| 25 | D | 275 | 127 | 131 | 156 | 139 | 1132 | 153 | 149 | 167 | 166 | 218 | 294 | 151 |
| 26 | l | I276 | L128 | I132 | V157 | L140 | L1133 | L154 | L150 | I168 | L167 | L219 | I295 | V152 |
| 27 | K | 277 | 129 | 133 | 158 | 141 | R1136b | 155 | 151 | 169 | 168 | S220p | 296 | 153 |
| 28 | N | 280 | 132 | 136 | 161 | 144 | 1137 | 158 | 154 | 172 | 171 | 223 | 299 | 156 |
| 29 | h | L281 | L133 | F137 | V162 | I145 | C1138 | L159 | I155 | I173 | L172 | V224 | V300 | I157 |
| 30 | h | M282 | L134 | L138 | M163 | M146 | M1139 | A160 | L156 | L174 | L173 | L225 | L301 | L158 |
| 31 | h | L283 | I135 | I139 | I164 | L147 | V1140 | V161 | L157 | I175 | I174 | V226 | M302 | L159 |
| 32 | l | I289 | I141 | I150 | L171 | I157 | V1146 | L167 | I163 | L182 | I180 | I231 | I546 | V165 |
| 33 | l | I291 | L143 | V152 | L173 | I159 | I1148 | I169 | L165 | L184 | V182 | I233 | I548 | L167 |
| 34 | D | 293 | 145 | 154 | 175 | 161 | 1150 | 171 | 167 | 186 | 184 | 235 | 550 | 169 |
| 35 | F | 294 | 146 | 155 | W176r | 162 | 1151 | 172 | 168 | 187 | 185 | 236 | L551h | 170 |
| 36 | G | 295 | 147 | 156 | 177 | 163 | 1152 | 173 | 169 | 188 | 186 | P237s | 552 | 171 |
| 37 | s | C297 | A149 | V158 | A179 | A165 | T1154 | A175 | S171 | G190 | A188 | S239 | A554 | A173 |
| 38 | T | 313 | 165 | 181 | S194s | 180 | x | 188 | 186 | 204 | 201 | x | 567 | 185 |
| 39 | P | 319 | 171 | x | 200 | 186 | 1178 | 194 | 192 | 210 | 207 | x | 573 | 191 |
| 40 | E | 320 | 172 | N188p | 201 | 187 | 1179 | 195 | 193 | 211 | 208 | x | 574 | 192 |
| 41 | D | 332 | 185 | 200 | 214 | 199 | 1191 | 208 | 211 | x | 220 | x | 586 | 207 |
| 42 | W | 334 | 187 | x | 216 | 201 | 1193 | 210 | 213 | 226 | 222 | x | 588 | 209 |
| 43 | G | 337 | 190 | 205 | 219 | 204 | 1196 | 213 | 216 | 229 | 225 | I259l | A591v | 212 |
| 44 | h | M343 | M196 | F211 | M225 | L210 | I1202 | M219 | L222 | M235 | M231 | F262 | L597 | L218 |
| 45 | h | L384 | M266 | Y256 | L304 | L255 | C1245 | M291 | F267 | C270 | L272 | A275 | M686 | C261 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 46 | L | 385 | 267 | M257h | 305 | 256 | W1246h | 292 | 268 | 271 | 273 | I272l | 687 | 262 |
| 47 | R | 392 | 274 | x | 312 | 263 | 1253 | 299 | 275 | 278 | 280 | x | 694 | 269 |
| 48 | H | 406 | 283 | x | 321 | 272 | x | 308 | 284 | 287 | 294 | x | 703 | 278 |

**Supplementary Table 3.2**. Conserved water interactions. Oxygen and nitrogen atoms within 3.2 Å of each conserved water are listed by residue number. Conserved residues are in bold and organized by row. Dashes indicate samples missing the conserved water.

| | Kinase | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Water | Akt2 | CDK2 | CK1 | CK2 | DAPK | IRK | p38-γ | PhK | Pim-1 | PKA | Rio2 | Sky1P | TAO2 |
| $A_w$ | **OE2-E200** | **OE2-E51** | **OE2-E55** | **OE2-E81** | **OE2-E64** | **OE2-E1047** | **OE2-E74** | **OE2-E73** | | **OE2-E91** | **OE2-E154** | | **OE1-E76** |
| | **N-F294** | **N-F146** | | | **N-F162** | **N-F1151** | | **N-F168** | | **N-F185** | | | **N-F170** |
| | | | OH-Y59 | N-W176 | | | | | - | | N-D235 | - | |
| | | | N-D154 | | | | | | | | | | |
| | $O\text{-}H_2O(265)$ | $O\text{-}H_2O(74)$ | $O\text{-}H_2O(403)$ | $O\text{-}H_2O(13)$ | $O\text{-}H_2O(3400)$ | $O\text{-}H_2O(31)$ | | $O\text{-}H_2O(402)$ | | | $O\text{-}H_2O(13)$ | | $O\text{-}H_2O(265)$ |
| $B_w$ | | **O-L55** | **O-Y59** | | **O-L68** | **O-M1051** | | **O-L77** | | | | **O-L206** | |
| | | **N-L66** | | | **N-L79** | **N-L1062** | | **N-L89** | | | | **N-L228** | **N-Y91** |
| | O-T207 | | O-P69 | | O-I71 | O-F1054 | O-M81 | O-I87 | | O-L95 | | OD1-N210 | O-L83 |
| | OG1-T207 | | | - | | | O-I87 | | - | | - | ND2-N210 | |
| | O-T213 | | | | | | | | | | | | |
| | $O\text{-}H_2O(33)$ | | $O\text{-}H_2O(442)$ | | $O\text{-}H_2O(3021)$ | $O\text{-}H_2O(41)$ | | | | $O\text{-}H_2O(406)$ | | | $O\text{-}H_2O(315)$ |
| | $O\text{-}H_2O(36)$ | | | | | | | | | | | | |
| $C_w$ | **O-Y273** | **O-H125** | **O-Y129** | **O-H154** | **O-H137** | **O-H1130** | **O-H151** | **O-H147** | **O-H165** | | | | **O-H149** |
| | **O-D293** | **O-D145** | | **O-D175** | **O-D161** | **O-D1150** | **O-D171** | **O-D167** | **O-D186** | **O-D184** | | | **O-D169** |
| | N-L296 | N-L148 | N-M157 | N-L178 | N-L164 | N-M1153 | N-L174 | N-F170 | N-S189 | N-F187 | - | - | N-S172 |
| | | N-A149 | | N-A179 | N-A165 | N-T1154 | | | N-G190 | | | | N-A173 |
| | | | | | | OG1-T1154 | | | | | | | |
| | | | | | | | | | | $O\text{-}H_2O(424)$ | | | |
| $D_w$ | **OD1-D275** | | | **OD1-D156** | | **OD1-D1132** | **OD1-D153** | | **OD1-D167** | | | | **OD1-D151** |
| | **O-I276** | - | **O-I132** | **O-V157** | **O-L140** | **O-L1133** | **O-L154** | - | **O-I168** | **O-L167** | **O-L219** | - | **O-V152** |
| | | | | N-S194 | | | | | | | ND2-N257 | | |
| | $O\text{-}H_2O(96)$ | | | | $O\text{-}H_2O(458)$ | $O\text{-}H_2O(63)$ | | | | $O\text{-}H_2O(576)$ | | | $O\text{-}H_2O(54)$ |
| | | | | | $O\text{-}H_2O(439)$ | | | | | | | | |
| $E_w$ | | | **N-D131** | | | | **N-D153** | | | | | **N-D294** | **N-D151** |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **N-I276** | **N-L128** | **N-I132** | **N-V157** | **N-L140** | **N-L1133** | **N-L154** | | **N-I168** | **N-L167** | | **N-I295** | **N-V152** |
| | **OD1-D332** | **OD1-D185** | | **OD1-D214** | **OD1-D199** | **OD1-D1191** | **OD1-D208** | **OD1-D211** | | **OD1-D220** | - | **OD1-D586** | **OD1-D207** |
| | | OG-S188 | | OG-S217 | OG-S202 | OG-S1194 | OG-S211 | OG-S214 | OG-S227 | | | OG-S589 | OG-S210 |
| | | | O-$H_2O$(457) | | O-$H_2O$(439) | O-$H_2O$(34) | | O-$H_2O$(422) | O-$H_2O$(22) | O-$H_2O$(411) | | | |
| $F_w$ | **NE1-W334** | **NE1-W187** | | **NE1-W216** | **NE1-W201** | **NE1-W1193** | **NE1-W210** | **NE1-W213** | **NE1-W226** | **NE1-W222** | | **NE1-W588** | **NE1-W209** |
| | O-E315 | O-W167 | - | O-Y196 | O-E182 | O-R1174 | O-W190 | O-S188 | O-V206 | O-E203 | OD1-N257 | O-E569 | O-Y187 |
| | O-L317 | O-R169 | | O-K198 | O-V184 | OH-Y1210 | O-R192 | O-L190 | O-S208 | O-L205 | O-H2O235 | O-R571 | O-M189 |
| | | NE2-Q211 | | | | | | | | | | NE2-H618 | |
| | | | | | | O-$H_2O$(20) | | | | | | | O-$H_2O$(278) |

# CHAPTER 4: A novel whole-cell lysate kinase assay identifies substrates of the p38 MAPK in differentiating myoblasts

# A novel whole-cell lysate kinase assay identifies substrates of the p38 MAPK in differentiating myoblasts

James D. R. Knight[1,2], Ruijun Tian[3,4], Robin E. C. Lee[1,2], Fangjun Wang[3,5], Hanfa Zou[5], Lynn A. Megeney[1,2,6], Daniel Figeys[3,4] and Rashmi Kothary[1,2,6]

[1]Regenerative Medicine Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada K1H 8L6

[2]Department of Cellular and Molecular Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

[3]Ottawa Institute of Systems Biology, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

[4]Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

[5]CAS Key Lab of Separation Sciences for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China 116023

[6]Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada K1H 8M5

**Manuscript submitted.**

**Author Contributions**

JDRK and RK conceived of and designed the project. JDRK conceived of the substrate finding technique. RT performed dimethyl labeling, phosphopeptide enrichments and MS analysis, and assisted JDRK with the implementation of the substrate finding technique. RECL performed 2D gel work and assisted JDRK with the development of the radioactive approach. FW performed quantitative MS analysis and validated phosphopeptide data. All other experiments and analysis were performed by JDRK. JDRK wrote the paper and RK revised and edited it. RECL and LAM provided criticism during manuscript preparation. HZ supervised FW. LAM supervised RECL and gave advice on project design. DF supervised RT and FW, and provided MS facilities and expertise. RK supervised JDRK.

**Abstract**

Although techniques for identifying substrates for protein kinases exist, they are difficult to use and/or generally work poorly, and as a result are not widely used. We describe here a simple technique that identifies kinase substrates and their phosphorylation sites, and can be used to compare multiple kinases in the same experiment. Applying the technique to the p38α MAPK has resulted in the identification of 156 phosphorylation sites on 93 proteins in lysate derived from differentiating myoblasts. A comparison with p38β has revealed that substrate specificity is not what discriminates these two isoforms, but instead cellular localization is their distinguishing characteristic. The substrate screen has also lead to the identification of a potentially critical role for p38α in the cytoplasm during myoblast differentiation. The method presented here provides a necessary tool for studying the hundreds of protein kinases that exist, and for uncovering the deeper mechanisms of phosphorylation-dependent cell signaling.

## Introduction

Protein kinases are well-known regulators of cell signaling and cellular behavior that execute their function through the covalent attachment of an ATP-derived phosphate to protein substrates. To understand the function of any protein kinase on a large and cell-wide scale first requires the development of a substrate screening technique that allows for the proteins phosphorylated by a kinase of interest to be comprehensively identified, ideally in a single experiment. Although substrate-finding techniques exist, they are hindered by problems that prevent them from becoming easily or readily employed[165-167]. Such problems include incomplete scope and fabrication difficulties associated with protein (or peptide) arrays[49], or for those techniques that work with cell lysate the results are either poor (two to three substrates identified at best) or they require mutation of the kinase of interest, if permissible, and the use of non-physiological ATP analogs that cannot be used by all kinases[64, 72]. Having a simple to use technique free from these complications is an obvious need for the field.

The mitogen-activated protein kinase p38α is involved in several cellular processes, but its critical role during differentiation, and particularly the differentiation of myoblasts, has been a major focus. At the initiation of myoblast differentiation p38α is known to phosphorylate several transcription factors and chromatin remodeling proteins, thereby inducing the expression of a myogenic gene program[88]. Although much is known about p38α's role in this process, it is likely very partial, and whether or not p38α is playing an important role in other processes during myoblast differentiation, such as cell fusion or sarcomere formation, is unknown. At the same time there are questions regarding the other p38 isoforms and their role in myogenesis, or lack thereof. p38β is

also expressed in myoblasts and is activated in the same manner as p38α, but despite having a kinase domain 75% identical to p38α (72% sequence identity overall), p38β is unable to compensate for the loss of p38α, even when overexpressed[80, 89, 168]. The obvious and suspected explanation is that there are critical myogenic phosphorylations specific to the α isoform, but these have yet to be discovered, and whether this assumption is correct is unknown.

Here we describe a simple approach for substrate finding that begins with treatment of cell lysate to inactivate endogenous kinases, followed by an *in vitro* assay using a kinase of interest, and concludes with quantitative mass spectrometry to identify phosphorylation sites specific to the added kinase. Applying this technique to p38α with lysate from differentiating myoblasts has resulted in the identification of many new substrates that suggest novel functions for p38α during myogenesis. We do not identify a single phosphorylation specific to the p38α isoform when compared with p38β, and propose that their distinguishing characteristic during myoblast differentiation is cellular localization rather than substrate specificity.

## Results

### FSBA inhibits endogenous protein kinases

A substrate finding approach that works with cell lysate has to overcome the obstacle of endogenous protein kinase activity. Lysate contains tens if not hundreds of active kinases, making it difficult to attribute individual phosphorylations that occur during a lysate based assay to a particular kinase. 5′-4-Fluorosulfonylbenzoyladenosine (FSBA) offers a simple solution. FSBA is an ATP analog that inhibits protein kinases by occupying the ATP binding site and covalently attaching to an invariant lysine[169-172], the fully conserved and so-called catalytic lysine[152]. As FSBA irreversibly occupies the ATP binding site, a bound kinase will permanently lose activity. Treatment of whole-cell C2C12 myoblast lysate with this compound can completely eliminate the endogenous kinase signal present (**Fig. 4.1a**).

### Kinase-specific substrate labeling

Cell lysate treated with FSBA can be desalted to remove any unbound inhibitor and a pool of protein with no inherent kinase activity is generated. A kinase of interest can then be added with a kinase assay buffer and any labeling that subsequently occurs is due to the added kinase as opposed to an endogenous one. To specifically label substrates of p38α, a kinase assay buffer containing $^{32}$P-γ-ATP was added to FSBA-treated C2C12 lysate along with recombinant p38α and the sample assayed at 30°C. Substrates labeled by p38α appear as bands following one dimensional (1D) gel electrophoresis or as spots in 2D with no contaminating signal from endogenous kinases (**Fig. 4.1b,c**). Although this type of approach is excellent for visualizing phosphorylation, it is very difficult to

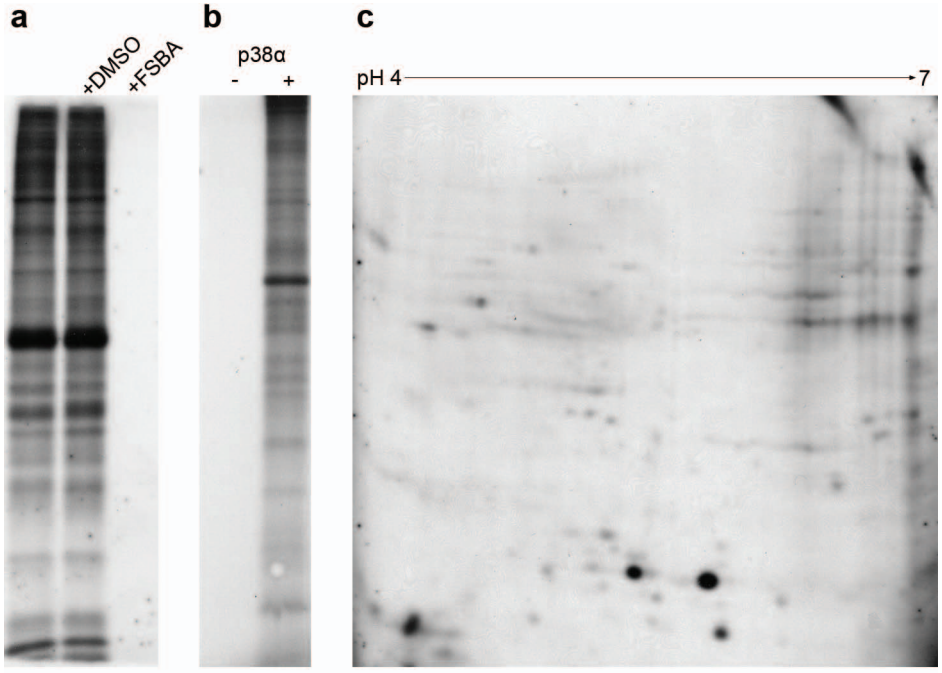



**Figure 4.1**. FSBA is a pan-kinase inhibitor that allows for kinase-specific substrate labeling of cell lysate. (**a**) C2C12 cell lysate was treated with either nothing (lane 1), DMSO or FSBA solubilized in DMSO. Subsequently, samples were desalted, kinase assay buffer added containing $^{32}$P-γ-ATP and the samples assayed for 1.5 hours. Treatment with FSBA abolishes the labeling of endogenous protein kinase substrates. (**b**, **c**) After pre-treatment of C2C12 cell lysate with FSBA, purified p38α was added with a kinase assay buffer to specifically label its substrates, visualized via 1D SDS-PAGE (**b**) or 2D gel electrophoresis (**c**).

identify phosphorylated proteins through spot picking and mass spectrometry. We therefore sought an alternative gel- and radioactive-free approach for identifying phosphorylated proteins.

**Quantitative MS coupled with a phosphopeptide enrichment to identify substrates**

The approach we devised to identify substrates is outlined in **Figure 4.2**. Cell lysate is treated with FSBA and desalted as described in the previous section. This is followed by the addition of a non-radioactive kinase assay buffer and the sample is split in two. To one sample is added active kinase (kinase-added) and to the other the same amount of heat-inactivated kinase (control). After assaying at 30°C the two samples are digested and peptides from each sample differentially tagged using isotopomeric dimethyl labels. As a final step before mass spectrometry (MS), the samples are combined and an enrichment for phosphopeptides is performed using $TiO_2$. Phosphopeptides are then identified using 2D LC-MS/MS and their relative abundance between samples quantified via the differential dimethyl labeling of peptides. Phosphopeptides that are more abundant in the kinase-added sample are from proteins labeled by the added kinase during the *in vitro* assay, and by this means substrates can be identified. This approach results not only in substrate identification but also identification of the site of phosphorylation, and as up to three samples can be compared using dimethyl labeling, the phosphorylation profile of two kinases (plus a control) can be compared in a single experiment.

To identify p38α substrates during myoblast differentiation, lysate from C2C12 cells that had been differentiating for 48 hours was used as a substrate pool. This time point was chosen because of our interest in identifying novel functions for p38α.

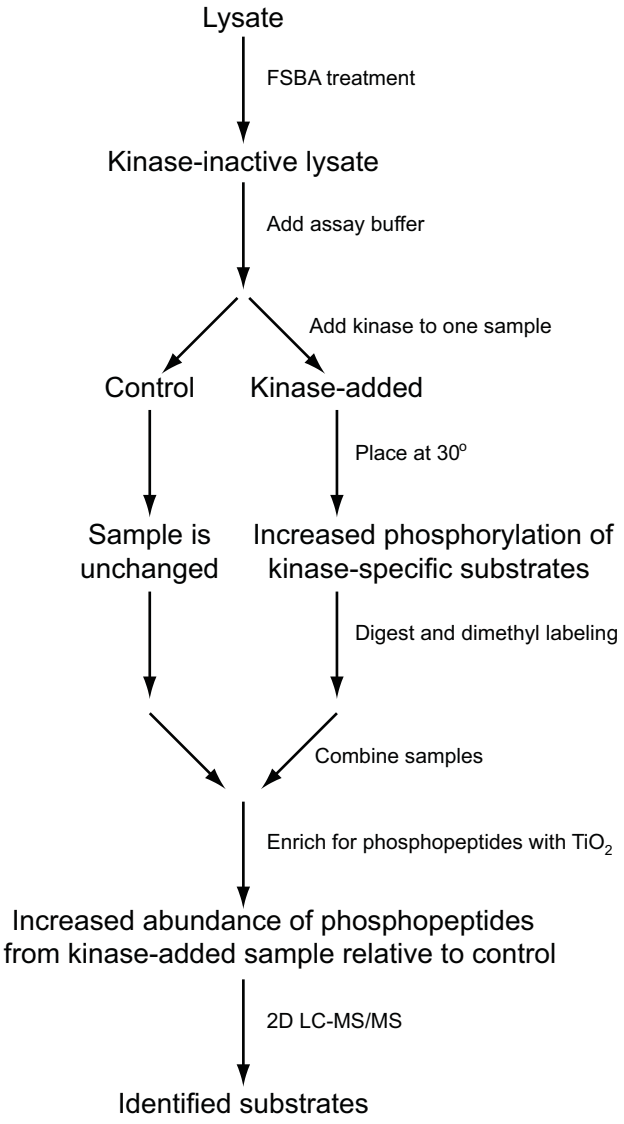



**Figure 4.2**. Methodology to identify kinase substrates using FSBA and quantitative MS.

Myogenic gene activation occurs within the first 48 hours of differentiation[173] and is followed by cell fusion, sarcomere formation and other processes. If differentiating C2C12 myoblasts are treated with a p38 inhibitor at this time, there is a reduction in cell fusion and overall differentiation (**Supplementary Fig. 4.1**), indicating there is a requirement for long-term p38 activity beyond the initial stage of myogenic gene activation. Substrate finding began by treating 1.5 mg of C2C12 lysate with FSBA, and for the assay the sample was split into three equal parts. Heat-inactivated p38α was added to the control, active p38α to the second sample and active p38β to the third. The rationale behind comparing p38α with p38β was to identify specific p38α phosphorylations that may explain why p38β cannot compensate for the loss of p38α in differentiating myoblasts.

In total 395 unique serine/threonine phosphopeptides were identified (**Supplementary Table 4.1**). A histogram of their relative abundance ratios (p38α/control) is shown in **Supplementary Figure 4.2**. A 2-fold increase in the abundance ratio was selected as a cutoff for high-confidence substrates as this is beyond the range of inherent variability, and 156 phosphopeptides from 93 different proteins show at least a 2-fold increase in the p38α-assayed sample relative to the control. The list of p38α phosphorylation sites is presented in **Supplementary Table 4.2**. We identified two previously known substrates (caldesmon and SAKS1) and 91 previously unknown. Three of these previously unknown targets were validated to determine if the technique was discovering true *in vitro* p38α substrates. **Figure 4.3a** shows the validation of three full-length proteins as substrates, and in **Figure 4.3b** is shown a validation of peptides to confirm the site of phosphorylation. After aligning substrates on their phosphorylation

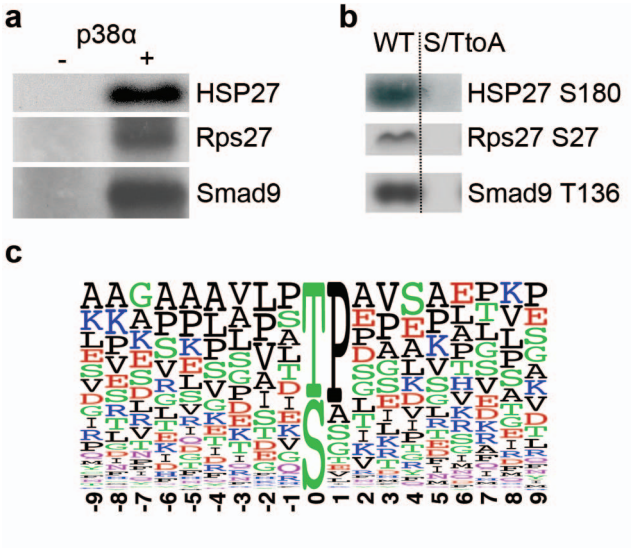



**Figure 4.3**. Validation of newly discovered p38α substrates. (**a**) Validation of three newly identified targets as *in vitro* p38α substrates was performed using full-length substrate incubated with a kinase assay buffer and p38α. (**b**) Wild-type peptides containing the phosphorylation site, or peptides harboring a S/TtoA mutation of the phosphorylation site were incubated with p38α and a kinase assay buffer to confirm the location of the phosphorylation site. (**c**) The 156 phosphorylations sites identified were aligned and a frequency logo generated using WebLogo[174]. p38α has a consensus phosphorylation motif of hXS/TP, where h is a small hydrophobic residue (not tryptophan or phenylalanine).

sites we find a consensus phosphorylation motif is present in many substrates, although not an absolute requirement (**Fig. 4.3c**). The motif contains a proline immediately downstream of the target serine or threonine, and a small hydrophobic residue two residues upstream. This motif is in agreement with the literature for p38[175], in further support of our substrate finding approach.

**Substrate specificity does not distinguish p38α and p38β**

We first assayed p38α and p38β on cell lysate using the radioactive approach and were surprised to see no obvious differences in the substrate banding patterns produced by the two isoforms (**Fig. 4.4a**). Consistent with this observation, of the 156 p38α phosphorylations identified using the quantitative approach, none of these appear specific to this isoform. The 156 quantitative values for the p38α phosphorylations are plotted from highest to lowest in **Figure 4.4b** (blue line for p38α). The red line is for the corresponding p38β values. Although p38α and p38β show differences in affinity, they have very similar profiles. There are only four phosphorylations that fall above the 2-fold cutoff for p38α while for p38β they are well below. However, the values are still positive for p38β, suggesting these could likely be real phosphorylations but there is simply less confidence in them. To determine if these may represent specific phosphorylations, we performed *in vitro* assays with either p38α or p38β using purified substrate for one candidate. The purified substrate was the 40s ribosomal protein S27 (Rps27) and both p38 isoforms were able to phosphorylate this protein, and at the same site (**Fig. 4.4c,d**). Therefore, none of the 156 phosphorylations found for p38α appear to be specific.

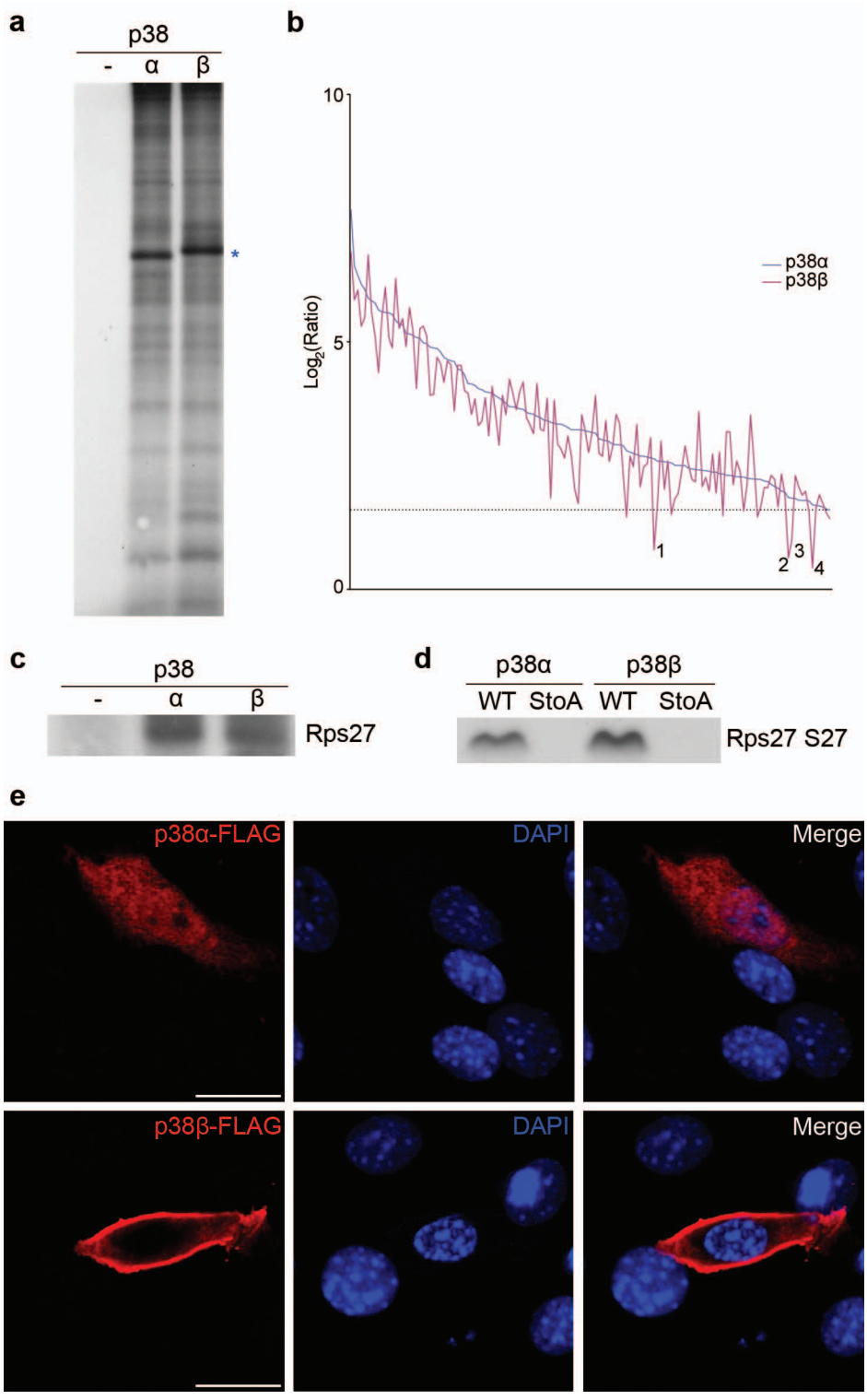



**Figure 4.4**. Localization, not substrate specificity, distinguishes p38α from p38β. (**a**) FSBA treated C2C12 cell lysate was incubated with a kinase assay buffer and p38α, p38β or no kinase as a control. Both isoforms produce similar banding patterns. The prominent

band marked by the asterisk in each of the p38α and p38β lanes represents autophosphorylation of the added recombinant kinase. (**b**) The quantitative values for all 156 identified p38α phosphorylations are plotted in blue on a $\log_2$ scale, with the corresponding values for p38β in red. The dashed line represents the fold-cutoff for accepted substrates. Four p38β phosphorylation fall appreciably below the fold cut-off (point 2 corresponds to serine 27 of the 40S ribosomal protein S27 – Rps27). (**c**) Incubation of p38α or p38β with purified Rps27 and a kinase assay buffer. (**d**) Incubation of p38α or p38β with a peptide from Rps27 containing serine 27 (or mutation of serine 27 to alanine) and a kinase assay buffer. (**c**) and (**d**) demonstrate that serine 27 of Rps27 is not a specific p38α phosphorylation site. (**e**) FLAG-tagged p38α or p38β were transfected into C2C12 cells, and at 48 hours of differentiation immunofluorescence staining for FLAG was performed. p38α has a ubiquitous distribution while p38β is found solely at the cell periphery. Scale bar = 20 μm.

If substrate specificity does not distinguish the two p38 isoforms, and yet p38β is unable to compensate for the loss of p38α, there must be an alternative characteristic that discriminates them. An obvious possibility is cellular localization. If p38α and p38β localize differently within the cell, then p38β would simply be unable to fulfill p38α's role even though it may have the potential to do so. To study this, we overexpressed FLAG-tagged p38α and p38β in C2C12 cells and assessed their localization during differentiation (**Fig. 4.4e**). While p38α has a ubiquitous localization pattern, p38β is found only at the periphery of the cell. p38α would therefore have access to a substrate pool that p38β does not, highlighting a major reason as to why p38β cannot compensate for the loss of p38α.

**Characterization of the cytoplasmic role of p38α**

An advantage of a substrate screening technique is that the results propose novel avenues of research for the kinase being studied. In the case of p38α we have identified dozens of cytoplasmic substrates with no obvious nuclear role, suggesting that p38α could have an important function in the cytoplasm during myoblast differentiation. From fractionations and western blotting we find that p38α is indeed present in the cytoplasm during differentiation (**Fig. 4.5a**), in agreement with the immunofluorescence staining of FLAG-tagged p38α. The levels of phosphorylated p38, the active form, increase in the cytoplasm with differentiation (**Fig. 4.5a, Supplementary Fig. 4.3**), suggesting p38α may be playing an important role there. Validation of the phospho-p38 antibody is shown in **Supplementary Figure 4.4**. Further fractionation of the cytoplasm reveals that p38α and phospho-p38 are present only in the cytosolic fraction (**Fig. 4.5b**), so, for example,

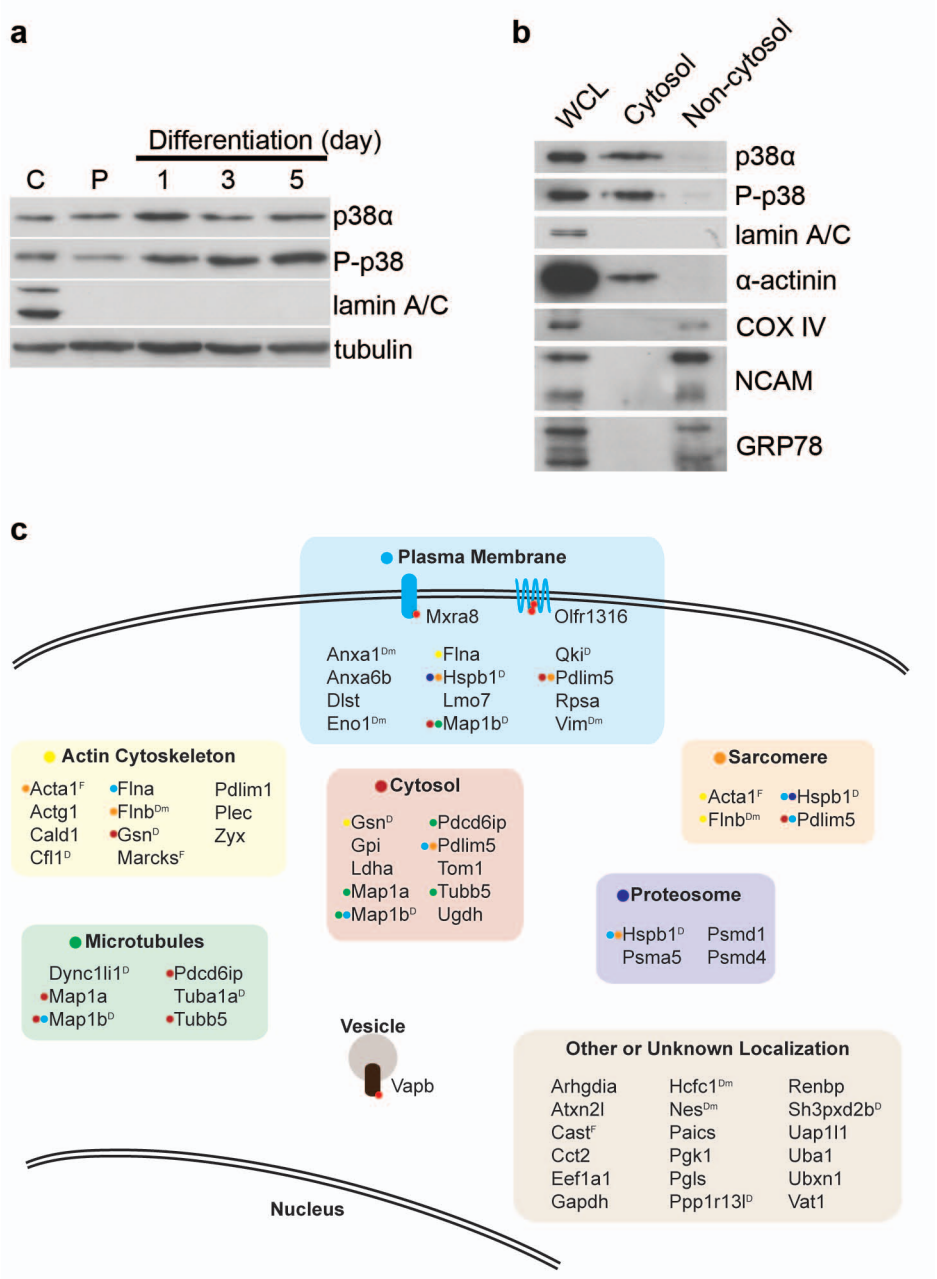



**Figure 4.5**. Cytoplasmic characterization of p38α during C2C12 cell differentiation. (**a**) C2C12 cells were fractionated over a differentiation time-course (C: control whole-cell lysate; P: proliferating lysate). Western blotting for phospho-p38 shows a clear increase in the cytoplasm as differentiation proceeds. Lamin A/C was used to demonstrate that there were no contaminating nuclei in cytoplasmic extracts. Quantification of phospho-p38 expression is shown in **Supplementary Figure 4.3**. (**b**) After 48 hours of

differentiation C2C12 cells were lysed, the cytoplasmic fraction collected and further fractionated into cytosolic (including the cytoskeleton) and non-cytosolic fractions. Nuclear marker: lamin A/C; cytoskeletal marker: α-actinin; mitochondrial marker: COX IV; membrane marker: NCAM; ER marker: GRP78. (**c**) Cytosolically accessible p38α substrates. Of the 93 substrates identified, 55 are known to be either in the cytosol or accessible from the cytosol based on GO annotations (displayed by compartment and gene name). Three are known to be involved in myocyte fusion (marked with a superscript F), six are known to be involved in myoblast differentiation (superscript Dm), and nine more are known to be involved in the differentiation of some other cell type (superscript D). See **Supplementary Table 4.3** for references. Colored circles to the left of gene names indicate additional compartments in which the protein product is found. The red circles on Olfr1316, Mxra8 and Vapb indicate that the phosphorylation sites identified on the protein products are cytosolically accessible.

although our screen identified eighteen mitochondrial proteins that can be phosphorylated by p38α, they would not be relevant in this context. Using GO annotations, a total of 55 of the 93 substrates identified are found either in the cytosol or have cytosolic domains that would be accessible to p38α (**Fig. 4.5c** and **Supplementary Table 4.3**). Three of these cytosolic substrates are known to regulate myocyte fusion, and a total of fifteen more are known to regulate either myoblast differentiation or the differentiation of some cell type. Western blotting on phosphoprotein-enriched fractions was done for substrates from several different cytoplasmic compartments and revealed that certain cytosolic proteins become phosphorylated with differentiation (**Fig. 4.6**). Actin, HSP27 and MARCKS are highly enriched from day 3 of differentiation onwards relative to proliferation. Interestingly both actin and MARCKS are known to play a role in myocyte fusion. Not all of the substrates blotted for get phosphorylated with differentiation, suggesting there may be fine differences in localization between them and p38α, or that regulatory mechanisms exist to prevent phosphorylation or dephosphorylate these proteins as quickly as they become phosphorylated. Collectively, however, our results suggest that p38α could indeed be playing a major role in the cytoplasm during differentiation.

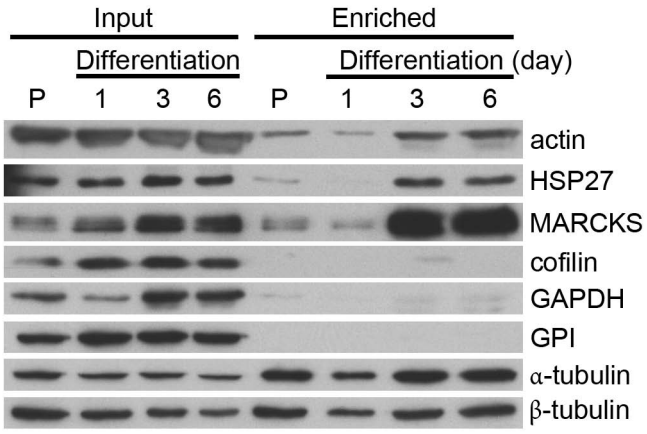



**Figure 4.6**. Certain cytosolic p38α substrates become phosphorylated with differentiation. Lysate from proliferating (P) or differentiating C2C12 cells were enriched for phosphoproteins. Western blotting was performed on the input and enriched fractions to identify if any cytosolic p38α substrates become phosphorylated with differentiation. Actin, HSP27 and MARCKS are highly enriched from day 3 onwards relative to proliferation.

**Discussion**

We have described here a simple technique for kinase substrate finding that uses whole-cell lysate, can identify sites of phosphorylation in addition to the protein phosphorylated, and can be used to compare the substrate specificity of two kinases in the same experiment. For p38α we have identified 156 phosphorylation sites on 93 different proteins with lysate derived from differentiating myoblasts. Performing the assay with p38α and p38β revealed that although differences in affinity are present, there are no obvious phosphorylations specific to p38α. Immunofluorescence has shown that a major characteristic that would distinguish p38α from p38β during myoblast differentiation is their localization, with p38α located ubiquitously throughout the cell, while p38β is found only at the cell periphery. At the same time, we have shown that p38α can phosphorylate dozens of cytosolic proteins and that this may be a critical yet previously unknown function for it. These results demonstrate the utility of the substrate screening technique we have developed, and how its use can not only find proteins phosphorylated by a kinase of interest, but also answer particular questions and propose novel avenues of research.

The technique we have developed has several advantages over existing substrate finding approaches. Whole-cell lysate can be used as a source of candidate substrates, allowing for screens to be performed on samples of specific relevance to the area of interest instead of on protein or peptide arrays. At the same time the recombinant kinase used does not require mutation or unnatural ATP analogs, making the technique applicable to any kinase that can be made and purified in an active form. Although we have used dimethyl labeling for quantitative MS, the technique is fully compatible with

iTRAQ, which would allow up to eight samples (seven kinases plus a control) to be compared in a single experiment. Radioactivity and gel work are not required but can be used to visualize phosphorylations and make qualitative comparisons between kinases on 1D gels. The lysate requirements are also relatively low; 1.5 mg was used in this assay, which may make the technique difficult to employ on some primary cells lines, but it is easily applicable to most secondary cell lines and to tissue as well. We have treated several different lysate types with FSBA and complete inhibition of endogenous kinase activity occurred in all (data not shown). It should be noted that to identify lower abundance substrates, such as transcription factors or other kinases, would require more lysate than we employed and possibly subcellular (for example nuclear) fractionations. Most known p38α substrates are low abundance proteins such as these, likely the major reason why we do not identify more previously known targets. However, these are drawbacks that will affect any cell-lysate based technique and can only be overcome with further advancements in phosphopeptide purification and mass spectrometry.

As mentioned, a major advantage of the technique presented here is that it can be used to compare the substrate profiles of two or more kinases, and we have used this to study the substrate specificity of the p38α and β isoforms. However, we were unable to identify any specific p38α phosphorylations. There are apparent differences in affinity for substrates between p38α and p38β, which could be functionally relevant. Assuming p38β had a ubiquitous localization pattern the same as p38α, a low affinity for certain critical myogenic phosphorylations could result in a difference in cell behavior if p38β were the only isoform present. However, it seems likely that differentiation would still occur, but possibly at a reduced rate or in a compromised way. This is not the case, as p38β-

containing myoblasts that lack p38α fail to differentiate at all[80, 86, 89, 168]. Rather, we believe what truly distinguishes the two isoforms is their localization. p38α has a known critical role in the nucleus, and our results suggest the same may be true for the cytoplasm, therefore p38β would be unable to compensate because it is found solely at the periphery of the cell. A similar distinction exists for two other closely related kinases, focal-adhesion kinase (FAK) and proline-rich tyrosine kinase 2 (PYK2)[176]. In fibroblasts FAK localizes to focal adhesions while PYK2 has a perinuclear distribution. PYK2, which has a kinase domain 60% identical to FAK, can compensate for the loss of FAK provided it has a focal adhesion targeting sequence. Together these results suggest that rather than substrate specificity, it is other characteristics, in these cases localization, that distinguish closely related kinases. Whether the same holds true for the entire kinase family is an intriguing question.

In addition to p38α's known role regulating gene expression at the onset of myoblast differentiation, our results suggest it is likely to have a critical cytosolic role as well. We have found that p38α is present in the cytosol, active p38 increases in the cytoplasm with differentiation, p38α can phosphorylate many cytosolic proteins, some of these proteins are already known to be critical for myoblast differentiation or fusion, and certain cytosolic proteins are getting phosphorylated with differentiation. Together these results suggest a previously unrecognized role for p38α in the cytosol, thus demonstrating the benefit of performing an unbiased screen such as that presented here and letting the results suggest direction for exploration.

**Acknowledgements**

JDRK was supported by a Vanier Canada Graduate Scholarship, a Canadian Institutes of Health Research (CIHR) Canada Graduate Scholarship and a Multiple Sclerosis Society of Canada (MSSC) Studentship. LAM holds the Mach Gaensslen Chair in Cardiac Research and was funded by CIHR. DF would like to acknowledge a Canada Research Chair in Proteomics and Systems Biology, and grants from the Natural Sciences and Engineering Research Council of Canada and the J.-Louis Lévesque Foundation. RK was supported by grants from CIHR and the MSSC.

## Methods

**Cell culture**. C2C12 cells were grown in DMEM supplemented with 10% (v/v) FBS with 100 units/ml penicillin, 100 μg/ml streptomycin, and 250 ng/ml of amphotericin B. To induce differentiation, cells were grown to 85-90% confluence and the media changed to DMEM with 2% horse serum supplemented with penicillin, streptomycin and amphotericin B as above. FLAG-tagged p38α and p38β were acquired from Addgene[177]. Constructs were transfected into C2C12 myoblasts with Lipofectamine 2000 according to the manufacturers instructions. To inhibit p38 activity, SB 202190 (Promega) solubilized in DMSO or DMSO as control was added to differentiation media at 48 hours following the induction of differentiation, and media with inhibitor was changed daily.

**Immunofluorescence**. Cells were fixed with 4% formaldehyde and stained with the following antibodies: Flag M2 (1:1000, Sigma), MyHC (1:20, Developmental Studies Hybridoma Bank), Alexa Fluor goat anti-mouse 488 (1:1000, Invitrogen), and Alexa Fluor goat anti-mouse 555 (1:1000 dilution, Invitrogen). The differentiation index was calculated as the number of MyHC-positive nuclei divided by the total number of nuclei. The fusion index was quantified as the number of nuclei per MyHC-positive cell. Five fields of view at a 20X magnification were counted and averaged per replicate with a total of three replicates.

**Statistical analysis**. Statistical analyses were performed using StatPlus. Means are shown with standard deviation, and student's t-tests were performed to determine significance for the differentiation and fusion indices.

**Western blotting**. For western blotting, cells were lysed in RIPA (50 mM Tris-HCl pH 7.5, 1% NP40, 0.1% SDS, 150 mM NaCl, 1 mM EDTA, 50 mM NaF, 200 μM $Na_3VO_4$, 1 mM PMSF, 10 μg/ml aprotinin, 10 μg/ml leupeptin and 10 μg/ml pepstatin), and 5X Laemmli buffer (300 mM Tris pH 6.8, 0.01% bromophenol blue, 10% SDS, 50% glycerol, 5% β-mercaptoethanol) was added to 25 μg of protein per sample to a final concentration of 1X for SDS-PAGE. For fractionation cells were lysed in hypotonic buffer (10 mM HEPES pH 7.5, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 50 mM NaF, 200 μM $Na_3VO_4$, 1 mM PMSF, 10 μg/ml aprotinin, 10 μg/ml leupeptin and 10 μg/ml pepstatin) and left on ice for 10 minutes. Cells were then passed through a 25 gauge needle three times and centrifuged at 500 g to pellet nuclei and unlysed cells. The supernatant was collected as whole cytoplasm. For further fractionation of the cytoplasm the supernatant was centrifuged again at 5000 g to pellet mitochondria and membrane fractions. The supernatant was then collected and centrifuged at 100000 g to pellet any remaining cell particles and the resulting supernatant collected as cytosol. RIPA was added to all fractions to a final concentration of 1X for complete lysis. Phosphoprotein enrichment was performed using a Qiagen kit according to the manufacturer's instructions. Briefly, cells were harvested in lysis buffer and 2.5 mg of protein per time point added to individual enrichment columns. The columns were washed and eluted proteins concentrated 10-fold. Equal volumes ($1/10^{th}$ of total enriched fraction = 20 μl) per time point were diluted in 5X Laemmli and electrophoresed by SDS-PAGE for western blotting. Recombinant-phosphorylated p38α (Millipore) was dephosphorylated using λ-protein phosphatase (NEB) by adding 800 units of phosphatase to 200 ng of p38α

diluted in λ-PPase buffer and the sample placed at 30°C for 1 hour. Antibodies for blotting were as follows: actin (1:500 dilution, Fitzgerald), α-actinin (1:125, Abcam), cofilin (1:500, Millipore), COX IV (1:1000, Abcam), GAPDH (1:5000, Abcam), GPI (1:500, Abcam), GRP 78 (1:500, Cell Signaling), HSP27 (1:500, Cell Signaling), lamin A/C (1:500, Abcam), MARCKS (1:50, Santa Cruz), MyHC (1:100, Developmental Studies Hybridoma Bank), MyoD (1:1000, Santa Cruz Biotechnology), myogenin (1:100, Developmental Studies Hybridoma Bank), NCAM (1:200, Abcam), p38α (1:500, Cell Signaling), phospho-p38 (1:500, Abcam), α-tubulin (1:1000, Oncogene) and β-tubulin (1:1000, Developmental Studies Hybridoma Bank). An Alpha Innotech HD2 was used to quantify phospho-p38 and tubulin expression.

**FSBA treatment and substrate labeling**. Cells were lysed in NP40 buffer (50 mM Tris-HCl pH 7.8, 150 mM NaCl, 1% (v/v) NP40, 1 mM PMSF, 10 μg/ml aprotinin, 10 μg/ml leupeptin and 10 μg/ml pepstatin). Lysate was treated at a concentration of 2 mg/ml with 20 mM 5′-4-Fluorosulfonylbenzoyladenosine (FSBA) solubilized in DMSO, and placed at 30°C for 1 hour. The sample was then diluted down 1:5 with NP40 buffer minus protease inhibitors and desalted using Millipore Amicon ultra filtration columns with a 10 kDa molecular weight cutoff. Following concentration the sample was diluted to 4 mg/ml with NP40 buffer and diluted 1:2 with 2X kinase assay buffer (40 mM MOPS pH 7.2, 50 mM β-glycerophosphate, 10 mM EGTA, 2 mM $Na_3VO_4$, 2 mM DTT, 50 mM $MgCl_2$, 400 μM cold ATP, 5 μCi $^{32}$P-γ-ATP). Recombinant p38α or p38β (Millipore) were added to a final concentration of 0.5% (w/w) total protein. Control and kinase-added samples were assayed at 30°C for 1.5 hours. For 1D SDS-PAGE, 5X Laemmli buffer was added

following the assay to 1X and the sample electrophoresed. For 2D electrophoresis, 17 cm ReadyStrip IPG strips (Biorad) were directly rehydrated with labeled lysate diluted in rehydration buffer (7 M urea, 2 M thiourea, 4% CHAPS, 1% DTT) following the manufacturer's directions. Isoelectric focusing was performed on a Protean IEF Cell (Biorad) under the following conditions: 200 V for 1 hour, 500 V for 1 hour, 5000 V ramp for 5 hours, 5000 V for 80000 VH. IPG strips were then equilibrated following the manufacturer's instructions and overlaid onto a 12% SDS-PAGE gel. Following electrophoresis, gels were dried and imaged. For 1D electrophoresis 100 μg of lysate was used per reaction. 300 μg was used for 2D electrophoresis.

For substrate identification, assays were performed as above with the following modifications. 1.5 mg of lysate was treated with FSBA, the sample desalted and 2X kinase assay buffer added (40 mM MOPS pH 7.2, 50 mM β-glycerophosphate, 10 mM EGTA, 2 mM $Na_3VO_4$, 2 mM DTT, 50 mM $MgCl_2$, 2 mM cold ATP). The sample was then split into three 500 μg aliquots, and 5 μg heat-inactivated p38α added to the control, 5 μg active p38α added to the second aliquot and 5 μg active p38β added to the third. The samples were then assayed for 3 hours at 30°C.

**Dimethyl labeling**. After assaying the samples were precipitated by methanol-chloroform, and then redissolved in 200 μL of 8 M urea, 50 mM Tris-HCl, pH 8.1 with sonication. The samples were then reduced with 20 mM DTT for 1 h at 60°C, and alkylated by 100 mM iodoacetamide for 30 minutes at room temperature in the dark. Subsequently the samples were diluted to 2 M urea with 50 mM Tris-HCl, pH 8.1, and digested with trypsin at a protein to trypsin ratio of 50:1 (w/w) for 16 h at 37°C. Finally,

the digested samples were acidified to pH 2 by 10% (v/v) formic acid. Dimethyl labeling of samples was performed as reported previously[178] and described briefly as follows. The acidified peptides were loaded onto C18 SPE columns (50 mg of packing material). After brief washing with 50 mM sodium phosphate buffer, pH 7.5, 3 mL light, intermediate and heavy labeling reagents were loaded onto C18 SPE columns trapped with control, p38α-, and p38β-labeled samples, respectively. After washing with 0.1% (v/v) formic acid, the labeled samples were eluted with 80% ACN (v/v), 0.1% (v/v) formic acid and dried by vacuum centrifugation.

**Phosphopeptide enrichment**. Phosphopeptide enrichment by $TiO_2$ was carried out as reported previously[179] with modifications. The dried samples were redissolved with 65% ACN/2% TFA/saturated glutamic acid and combined. $TiO_2$ beads suspended in 65% ACN/2% TFA/saturated glutamic acid were added into the above samples with a peptide to $TiO_2$ bead ratio of 1:4 (w/w). After vortexing for 40 minutes, the $TiO_2$ beads were recovered by centrifugation and washed thoroughly with 65% ACN/2% TFA. Finally, the enriched phosphopeptides were eluted with 10% (v/v) $NH_3 \cdot H_2O$ and dried by vacuum centrifugation.

**On-line 2D LC-MS/MS analysis**. On-line 2D LC-MS/MS analysis was performed as reported previously[180, 181] with modifications. The dried sample was redissolved with 0.1% formic acid and loaded onto a biphasic trap column (200 μm ID × 10 cm; 5 cm reversed phase column packed with ReproSil-Pur C18 resin (5 μm; 200 Å; Dr. Maisch GmbH, Ammerbuch, Germany) and a 5 cm monolith SCX column). The trapped

phosphopeptides were eluted from the trap column onto a C18 tip column (75 μm ID × 20 cm; 3 μm; 200 Å; Dr. Maisch GmbH, Ammerbuch, Germany) by a series of salt washes with increasing concentrations (0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, 100, 150, 200, 1000 mM). Each fraction was then separated by reversed phase based gradient elution and detected by LTQ-Orbitrap XL mass spectrometry. The reversed phase gradient was set as follows: 0-5% ACN for 2 minutes, 5-30% ACN for 90 minutes, 30-80% ACN for 5 minutes, and after flushing with 80% ACN for 10 minutes, the column was equilibrated with 0.1% formic acid aqueous solution for 13 minutes. The LTQ-Orbitrap XL mass spectrometer was operated in positive ionization mode. A voltage of 1.8 kV was applied. MS and MS/MS spectra were acquired in a data dependent mode, and one full MS scan was followed by ten MS/MS scans. The resolution was set at 60,000 at m/z 400 after accumulation to a target value of 500,000.

**Protein identification and quantification**. All MS/MS spectra in one acquired raw file were converted to a single *.mgf file using DTASupercharge (v2.0a7). The *.mgf file was queried against the International Protein Index (IPI) mouse database v3.52 using Mascot Version 2.1 (Matrix Science). To evaluate the false discovery rate (FDR), reversed sequences were appended to the database. Cysteine residues were searched as a static modification of +57.0215 Da, methionine residues with a variable modification of +15.9949 Da, and serine, threonine, and tyrosine residues with a variable modification of +79.9663. Light, intermediate and heavy dimethylation of peptide amino termini and lysine residues was set as a variable modification of +28.0313 Da, +32.0564 Da and +36.0757 Da, respectively. Peptides were queried using full tryptic cleavage constraints

with up to two missed cleavages sites. The mass tolerances were 7 ppm for parent masses and 0.5 Da for fragment masses. Phosphopeptides with a Mascot score ≥ 25 (rank 1, $P \leq 0.05$, bold red required) were selected and quantified.

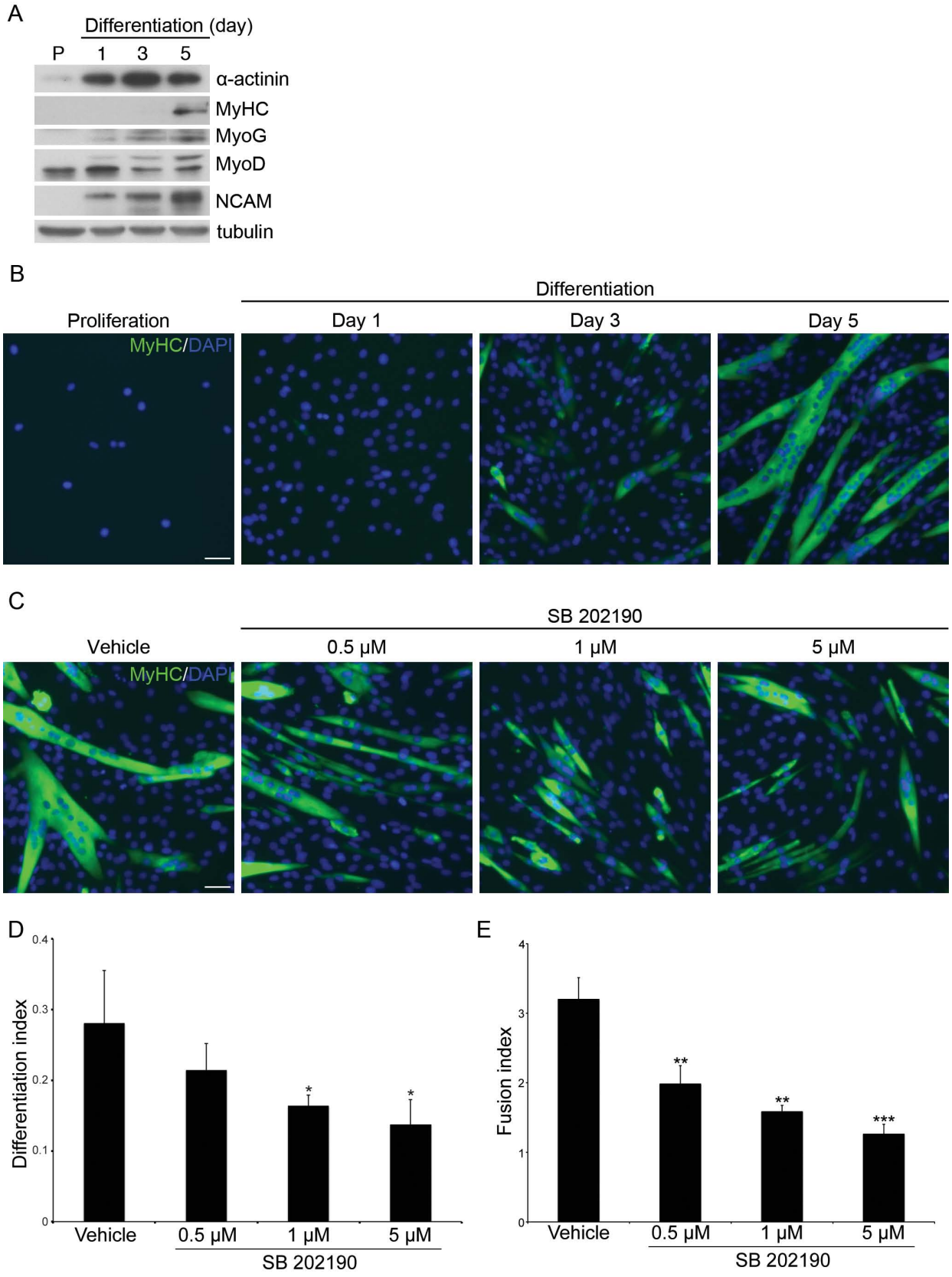
Phosphopeptide quantification was performed using a dimethyl-adapted version of MSQuant (v2.0a81). Peptide ratios were obtained by calculating the extracted ion chromatograms (XIC) of the light and heavy forms of the peptide using the monoisotopic peaks only and protein ratios were calculated from the average of the all quantified peptides. All MSQuant outputs of the same online multidimensional separation were then imported into StatQuant (v1.2.2) and the quantified phosphopeptides were sorted together and exported[181].

**Substrate Validation**. One μg of recombinant HSP27 (Stressgen/Enzo Life Sciences), recombinant Rps27 (Abnova) or recombinant Smad9 (Abnova) were diluted in 30 μl of 1X kinase assay buffer (20 mM MOPS pH 7.2, 25 mM β-glycerophosphate, 5 mM EGTA, 1 mM $Na_3VO_4$, 1 mM DTT, 25 mM $MgCl_2$, 200 μM cold ATP, 2.5 μCi $^{32}$P-γ-ATP). 500 ng of p38α or β were then added and the samples assayed at 30°C for 1 hour. 5X Laemmli buffer was added to 1X to terminate the assays, the samples electrophoresed, and the gels dried and imaged. Peptide assays were performed similarly with 5 μg of the following peptides synthesized by Biomatik: HSP27 S180-APLPKAVTQSAEITIPVTF, HSP27 A180-APLPKAVTQAAEITIPVTF, Rps27 S27-KHKKKRLVQSPNSYFMDVK, Rps27 A27-KHKKKRLVQAPNSYFMDVK, Smad9 T136-NPYHYQRVETPVLPPVLVP and Smad9 A136- NPYHYQRVEAPVLPPVLVP.

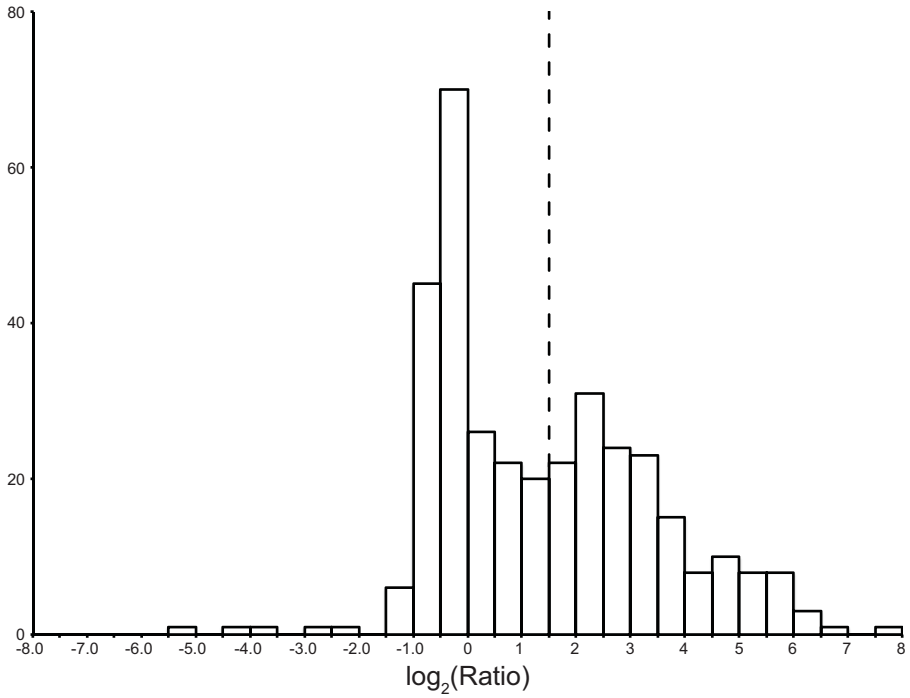
**Supplemental Data**

## A novel whole-cell lysate kinase assay identifies substrates of the p38 MAPK in differentiating myoblasts

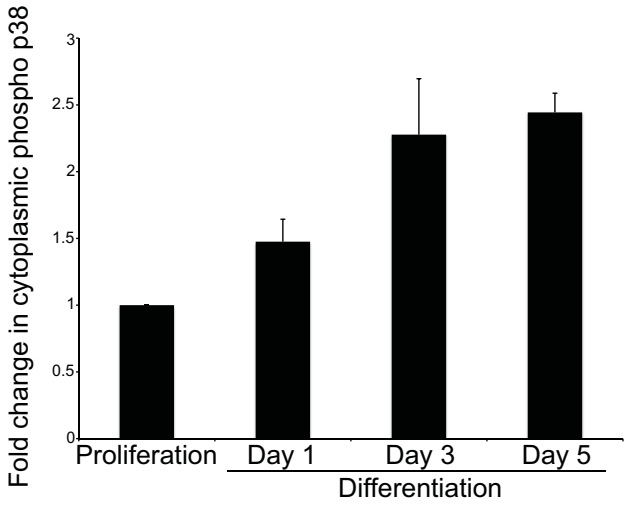
James D. R. Knight, Ruijun Tian, Robin E. C. Lee, Fangjun Wang, Hanfa Zou, Lynn A. Megeney, Daniel Figeys and Rashmi Kothary


A
Differentiation (day)
P
1
3
5
α-actinin
MyHC
MyoG
MyoD
NCAM
tubulin
B
Differentiation
Proliferation
Day 1
Day 3
Day 5
MyHC/DAPI
C
SB 202190
Vehicle
0.5 μM
1 μM
5 μM
MyHC/DAPI
D
Differentiation index
0.4
0.3
0.2
0.1
0
*
*
Vehicle
0.5 μM
1 μM
5 μM
SB 202190
E
Fusion index
4
3
2
1
0
**
**
***
Vehicle
0.5 μM
1 μM
5 μM
SB 202190

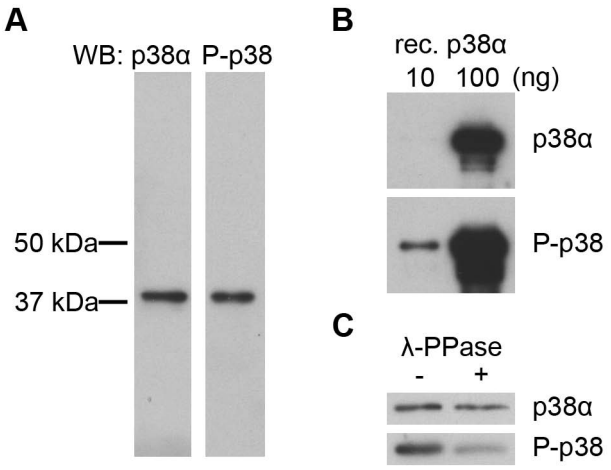
**Supplementary Figure 4.1**. p38 activity is required during the late stages of myoblast differentiation. C2C12 cells were used as a model system for myoblast differentiation. (**a**) Lysate from proliferating and differentiating C2C12 cells was subjected to western blotting. Differentiating C2C12 cells express typical muscle proteins. (**b**) Myosin-heavy chain (MyHC) staining of C2C12 cells over a differentiation time course. By day 3 the appearance of multinucleated myotubes is apparent, further augmented by day 5. (**c**) C2C12 cells were induced to differentiate and at 48 hours the media was supplemented with increasing concentrations of the p38 inhibitor SB202190. On day 5 (120 hours of differentiation), cells were stained for MyHC, a late marker of differentiation. Experiments were performed in triplicate and differentiation and fusion quantified as described in **Methods**. Inhibiting p38 at 48 hours of differentiation causes a reduction in the number of MyHC-positive/differentiated cells (**d**) and a reduction in myocyte fusion (**e**). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Scale bar = 100 μm.




**Supplementary Figure 4.2**. Distribution of relative phosphopeptide abundance values. FSBA treated lysate was labeled with active p38α (or inactive as a control), followed by dimethyl labeling and a phosphopeptide enrichment as described in the main text. The distribution of the returned abundance ratios (p38α/control) for all phosphopeptides identified is plotted with a bin size of 0.5. The vertical dashed line indicates the fold- cutoff that was used for accepted substrates.




**Supplementary Figure 4.3**. Quantification of cytoplasmic phospho-p38 levels during C2C12 differentiation from **Figure 4.5a** (normalized to tubulin expression). Western blotting was performed in triplicate and the mean is shown with standard deviation. The mean is shown in arbitrary units with proliferation set to 1.




**Supplementary Figure 4.4**. Validation of the phospho-p38 antibody. (**a**) Western blotting on C2C12 lysate was performed using p38α and phospho-p38 (P-p38) antibodies. p38α is present as a single band at 38 kDa, therefore all phosphorylated and non-phosphorylated p38α must have the same molecular weight. The phospho-p38 antibody detects a single band at 38 kDa. (**b**) The phospho-p38 antibody detects recombinant active/phosphorylated p38α (rec. p38α) with an affinity at least as high as the p38α antibody. (**c**) The phospho-p38 antibody detects p38α with less affinity following λ-phosphatase treatment of purified phosphorylated p38α, while the p38α antibody detects the purified protein with equal affinity regardless of its phosphorylation state.

**Supplementary Table 4.1**. Phosphopeptides identified using the whole-cell lysate kinase assay with p38α and p38β.

| IPI number | Sequence* | Mascot score | Abundance ratio (p38α/Control) | Abundance ratio (p38β/Control) |
|---|---|---|---|---|
| IPI00762803 | VELVPPpTPAEIPTAIQSVK | 81.03 | 21.51676023 | 22.9665184 |
| IPI00761237 | IILDLISEpSPIKGR | 59.35 | 1.73435986 | 1.296315312 |
| IPI00757768 | ITpSPL(ox)MEPSSIEK | 75.38 | 5.031611919 | 2.807758331 |
| IPI00134018 | LLPEGEEpTVESDDDKDER | 41.11 | 0.470338523 | 0.690812647 |
| IPI00407413 | RAPAAQPPAAAAPSAVGpSPAAAPR | 95.15 | 31.1769195 | 34.63834174 |
| IPI00758006 | ATpSNVFA(ox)MFDQSQIQEFK | 174.39 | 0.714270515 | 0.597432313 |
| IPI00229201 | GAAQNIIPASpTGAAK | 136.01 | 9.323296547 | 4.018772602 |
| IPI00761240 | ENPPpSPHSNSSGK | 49.23 | 0.731536199 | 0.801755379 |
| IPI00753321 | EVpSPPGAR | 46.61 | 0.646028876 | 0.639157534 |
| IPI00454179 | TNPPTQKPPpSPPVSGR | 73.55 | 0.654579043 | 0.628933787 |
| IPI00831454 | TSpSLPGYGK | 109.99 | 0.719388425 | 0.602286696 |
| IPI00154054 | FASEIpTPITISVK | 147.66 | 6.370707989 | 5.799799919 |
| IPI00911143 | LQPEQGpSPK | 65 | 0.815103352 | 0.777734995 |
| IPI00911143 | SLpSPLSGTTDTK<br>SLpSPLSGTTDTKAESPAGR | 87.55 | 0.735615318 | 0.704174133 |
| IPI00116074 | VDVpSPTSQR | 35.44 | 31.5514164 | 39.79124832 |
| IPI00473320 | AVFPpSIVGRPR | 81.51 | 5.03401947 | 6.817657471 |
| IPI00473320 | EIpTALAPST(ox)MK | 94.59 | 25.76988983 | 18.1419754 |
| IPI00473320 | VAPEEHPVLLpTEAPLNPK | 113.88 | 8.01223433 | 10.63028932 |
| IPI00221528 | SYELPDGQVIpTIGNER | 98.36 | 1.652413845 | 4.627465725 |
| IPI00653007 | VAPEEHPTLLpTEAPLNPK | 76.39 | 40.64113045 | 44.2283287 |
| IPI00653007 | VAPEEHPpTLLpTEAPLNPK | 92.17 | 4.705620766 | 11.4654274 |
| IPI00856263 | AAVVpTSPPPTTAPHK<br>AAVVTpSPPPTTAPHK | 89.76 | 1.258880615 | 0.780669987 |
| IPI00387580 | TEEVLSPDGSPpSKSPSK<br>TEEVLSPDGSPSKpSPSK | 91.57 | 0.698945701 | 0.789688528 |

| | | | | |
|---|---|---|---|---|
| IPI00553798 | ADIKpTPTVDVTVPEAELNVDSPEINIGGK<br>ADIKTPpTVDVTVPEAELNVDSPEINIGGK<br>ADIKTPTVDVpTVPEAELNVDSPEINIGGK | 64.1 | 3.218339205 | 3.755635023 |
| IPI00553798 | ADIKpTPTVDVTVPEAELNVDSPEINIGGK<br>ADIKTPpTVDVTVPEAELNVDSPEINIGGK | 78.82 | 3.160467148 | 3.318666697 |
| IPI00553798 | FKAEAPLPpSPK | 119.71 | 10.11185765 | 7.639447093 |
| IPI00553798 | AGAIpSASGPELEGAGHSK | 198.77 | 14.99047661 | 11.57382107 |
| IPI00553798 | ASLGpSLEGEVEAEASSPK | 99.69 | 0.746973574 | 0.636511564 |
| IPI00553798 | ASLGSLEGEVEAEApSSPKGK | 66.49 | 2.928720713 | 0.879008293 |
| IPI00553798 | EGVKDIDITpSPEF(ox)MIK | 75.38 | 2.27114439 | 5.111543655 |
| IPI00553798 | GGVTGpSPEASISGSKGDLK | 48.83 | 0.853488323 | 0.491832657 |
| IPI00553798 | GHYEVpTGSDDEAGKLQGSGVSLASK<br>GHYEVTGpSDDEAGKLQGSGVSLASK | 61.85 | 0.834961534 | 0.844889879 |
| IPI00553798 | GKGGVTGpSPEASISGSK | 67.76 | 1.019585371 | 0.814918041 |
| IPI00553798 | GKGGVpTGSPEASISGSK<br>GKGGVTGpSPEASISGSK | 41.92 | 0.979919016 | 0.795291126 |
| IPI00553798 | GPSFNVApSPESDFGVSLK | 174.57 | 9.354711056 | 6.457212687 |
| IPI00553798 | GPSLDIKpSPK | 28.71 | 7.68009758 | 6.782256126 |
| IPI00553798 | GPSLKGDLGApSSPS(ox)MK<br>GPSLKGDLGASpSPS(ox)MK | 39.82 | 2.007005215 | 0.767946064 |
| IPI00553798 | GSKVDIDpTPQVDVHGPDLK | 76.39 | 1.995457888 | 1.63072741 |
| IPI00553798 | IEGSITGPSVEIGpTPDVDVHGLGGK | 116.07 | 8.106399536 | 6.444357097 |
| IPI00553798 | IS(ox)MPDIDLHLKpSPK | 38.85 | 16.38875198 | 10.54151249 |
| IPI00553798 | LDIDpTPDIDIHGPEGK | 147.66 | 17.72900613 | 11.58450969 |
| IPI00553798 | LEGEIKVPDVDIpSSPGINVEAPDIH(ox)MK<br>LEGEIKVPDVDISpSPGINVEAPDIH(ox)MK | 78.04 | 6.015406132 | 3.353409767 |
| IPI00553798 | LEGEIKVPDVDISpSPGINVEAPDIH(ox)MK | 128.31 | 2.840684891 | 1.673624158 |
| IPI00553798 | LPQFGIpSTPGSDLDINIK<br>LPQFGISpTPGSDLDINIK | 64.21 | 6.302074432 | 8.361244202 |
| IPI00553798 | LPQFGISpTPGSDLDINIK | 91.91 | 3.952844381 | 5.034689426 |
| IPI00553798 | LPSGSGPApSPTTGSAVDIR | 28.02 | 1.514978886 | 1.391719699 |

| | | | | |
|---|---|---|---|---|
| IPI00553798 | LPpSGSGPASPTTGSAVDIR<br>LPSGpSGPASPTTGSAVDIR<br>LPSGSGPApSPTTGSAVDIR | 28.02 | 1.494694233 | 1.499192357 |
| IPI00553798 | LPpSGSGPApSPTTGSAVDIR<br>LPSGpSGPApSPTTGSAVDIR | 42.72 | 0.939601362 | 0.992946565 |
| IPI00553798 | LRpSEDGVEGDLGETQSR | 71.8 | 0.563569069 | 0.600886464 |
| IPI00553798 | (ox)MPFLSIpSSPK<br>(ox)MPFLSISpSPK | 57.59 | 203.2595215 | 112.4314575 |
| IPI00553798 | (ox)MPFLSISpSPK | 81.51 | 66.5812912 | 44.76893234 |
| IPI00553798 | SNSFSDEREFSAPpSTPTGTLEFAGGDAK<br>SNSFSDEREFSAPSpTPTGTLEFAGGDAK | 96.66 | 4.731276989 | 4.313008785 |
| IPI00553798 | pSNSFSDEREFSAPpSTPTGTLEFAGGDAK<br>SNpSFSDEREFSAPpSTPTGTLEFAGGDAK<br>SNSFpSDEREFSAPpSTPTGTLEFAGGDAK<br>pSNSFSDEREFSAPSpTPTGTLEFAGGDAK<br>SNpSFSDEREFSAPSpTPTGTLEFAGGDAK<br>SNSFpSDEREFSAPSpTPTGTLEFAGGDAK | 49.83 | 4.523502827 | 5.059371948 |
| IPI00553798 | SNpSFSDEREFSAPSTPTGTLEFAGGDAK<br>SNSFpSDEREFSAPSTPTGTLEFAGGDAK | 71.69 | 0.055511478 | 0.246567145 |
| IPI00553798 | pSNSFSDEREFSAPSTPTGTLEFAGGDAK<br>SNpSFSDEREFSAPSTPTGTLEFAGGDAK | 96.6 | 0.029090421 | 0.01622114 |
| IPI00553798 | SSEVVLpSGDDEDYQR | 133.43 | 0.697577119 | 0.696674883 |
| IPI00553798 | TVIRLPSGSGPApSPTTGSAVDIR<br>TVIRLPSGSGPASPpTTGSAVDIR | 75.25 | 1.445720315 | 0.668393075 |
| IPI00553798 | TVIRLPpSGSGPApSPTTGSAVDIR<br>TVIRLPSGpSGPApSPTTGSAVDIR | 84.91 | 0.854798615 | 0.433122545 |
| IPI00553798 | TVIRLPSGpSGPApSPTTGSAVDIR | 82.8 | 0.809538841 | 0.524245262 |
| IPI00553798 | TVIRLPpSGSGPASPTTGSAVDIR<br>TVIRLPSGpSGPASPTTGSAVDIR | 106.28 | 0.07723733 | 0.108180381 |
| IPI00553798 | VDIDpTPQVDVHGPDLK | 164.13 | 5.659213543 | 5.218640804 |
| IPI00553798 | VDLEpTPSLDVH(ox)MESPDINIEGPDVK<br>VDLETPpSLDVH(ox)MESPDINIEGPDVK | 65.7 | 5.35586977 | 4.504248857 |

| | | | | |
|---|---|---|---|---|
| IPI00553798 | VDLEpTPSLDVH(ox)MEpSPDINIEGPDVK | 76.28 | 3.907617807 | 1.853503942 |
| IPI00553798 | VDLETPSLDVH(ox)MEpSPDINIEGPDVK | 107.4 | 1.839476824 | 3.798395872 |
| IPI00553798 | VDLEpTPSLDVH(ox)MESPDINIEGPDVK | 56.79 | 1.809990406 | 2.280708075 |
| IPI00553798 | VPDVDIpSSPGINVEAPDIH(ox)MK | 138.22 | 9.175672531 | 7.833360672 |
| IPI00553798 | VPDVDIpSSPGINVEAPDIH(ox)MK<br>VPDVDISpSPGINVEAPDIH(ox)MK | 124.1 | 2.940662821 | 2.841840148 |
| IPI00553798 | VQpTPEVDVK | 106.41 | 56.81110764 | 48.3420639 |
| IPI00553798 | VSVApTPDVSLEASEGAVK | 114.49 | 5.698278785 | 3.542760086 |
| IPI00553798 | ADIDVpSGPKVDIDVPDVNIEGPDAK | 76.28 | 0.91219002 | 0.933866918 |
| IPI00553798 | AEpSPE(ox)MEVNLPK | 89.76 | 0.948591053 | 1.510591865 |
| IPI00553798 | ATIDVSGPKLDIEpTSDVSLEGPEGK<br>ATIDVSGPKLDIETpSDVSLEGPEGK | 82.8 | 6.461926937 | 4.198677063 |
| IPI00553798 | GPDINLPEVpSVKTPK | 106.25 | 21.87118721 | 23.05458641 |
| IPI00553798 | GSRVDIEpTPNLEGTLTGPK | 40.26 | 6.218453884 | 1.737978339 |
| IPI00553798 | IpSSPSGK<br>ISpSPSGK | 40.22 | 1.154801369 | 1.851094961 |
| IPI00856247 | GILAADESpTGSIAKR | 109.95 | 29.69270134 | 15.15349483 |
| IPI00172197 | pSCTKPSPSK | 21.9 | 0.939577937 | 0.804510415 |
| IPI00230395 | GGPGpSAVpSPYPSFNVSSDVAALHK | 87.24 | 3.443393469 | 2.932432175 |
| IPI00230395 | GGPGpSAVSPpYPSFNVSSDVAALHK | 91.57 | 3.356889963 | 2.932432175 |
| IPI00230395 | GGPGpSAVSPYPSFNVSSDVAALHK | 87.76 | 2.964440227 | 3.096542358 |
| IPI00310240 | DQAQEDAQEIADpTPSGDKTSLETR | 83.62 | 3.476381302 | 3.741753817 |
| IPI00754418 | TSEDTSSGpSPPKK | 60.79 | 0.68289423 | 0.894969821 |
| IPI00754418 | TSEDpTSSGSPPKKSPGGPK<br>TSEDTpSSGSPPKKSPGGPK<br>TSEDTSpSGSPPKKSPGGPK<br>TSEDTSSGpSPPKKSPGGPK | 83.47 | 0.814678729 | 0.614110351 |
| IPI00170307 | RGpSET(ox)MAGAAVK | 139.42 | 0.682991385 | 0.723003387 |
| IPI00224070 | SSpSPEPVTHLK | 65.55 | 2.056138992 | 1.987481236 |
| IPI00224070 | SpSSPEPVTHLK | 83.81 | 1.854281425 | 2.027197361 |

| | | | | |
|---|---|---|---|---|
| | SSpSPEPVTHLK | | | |
| IPI00322312 | AEEYEFLpTP(ox)MEEAPK | 161.03 | 8.238176346 | 6.729856014 |
| IPI00468481 | VLDpSGAPIKIPVGPETLGR | 136.01 | 24.97766495 | 24.40690041 |
| IPI00169500 | STSTPTpSPGPR | 81.51 | 10.68512535 | 9.208967209 |
| IPI00129519 | AEGAGTEEEGpTPKESEPQAAADATEVK | 81.42 | 1.029820204 | 2.010046482 |
| IPI00129519 | SDAAPAASDSKPpSSAEPAPSSK<br>SDAAPAASDSKPSpSAEPAPSSK | 119.94 | 1.590026975 | 7.439866543 |
| IPI00415385 | EVQpSPEQVKSEK | 81.51 | 0.634465396 | 0.453926951 |
| IPI00415385 | IDISPpSALR | 91.73 | 0.6715042 | 0.661963761 |
| IPI00282748 | GNKpSPSPPPDGSPAATPEIR | 130.95 | 0.822660387 | 0.850421548 |
| IPI00282748 | GNKSPpSPPPDGSPAATPEIR | 105.78 | 0.57112509 | 0.61547929 |
| IPI00282748 | VNHEPEPASGApSPGATIPK | 73.29 | 0.553376339 | 0.761530601 |
| IPI00311344 | GSVFSAPpSASGTPNKETAGLK<br>GSVFSAPSApSGTPNKETAGLK<br>GSVFSAPSASGpTPNKETAGLK | 89.63 | 15.6720953 | 17.13145828 |
| IPI00311344 | GSVFSAPpSASGTPNKETAGLK<br>GSVFSAPSApSGTPNKETAGLK | 101.78 | 13.53505516 | 11.50659084 |
| IPI00311344 | QSVDKVpTSPTKV<br>QSVDKVTpSPTKV | 77.31 | 1.175395727 | 0.932990015 |
| IPI00399958 | SFDQLpTPEESKER | 62.3 | 0.591723025 | 1.443731308 |
| IPI00119618 | AEEDEILNRpSPR | 73.55 | 0.758565954 | 0.909349354 |
| IPI00119618 | APVPpTGEVYFADSFDR | 180.59 | 4.525886536 | 4.148888588 |
| IPI00133349 | GNVVPpSPLPTRR | 65.55 | 0.868576044 | 0.739671483 |
| IPI00133349 | TFpSATVR | 63.74 | 0.763473213 | 0.760139287 |
| IPI00230645 | LSAAISEVVSQpTPAPSTHAAAPLPGTEQK | 130.88 | 3.589641571 | 1.996758342 |
| IPI00230645 | VTASSAApTSKSPS(ox)MSTTETK<br>VTASSAATpSKSPS(ox)MSTTETK<br>VTASSAATSKpSPS(ox)MSTTETK | 35.73 | 0.636452913 | 0.667064607 |
| IPI00381495 | LDQPVSAPPpSPR | 73.55 | 1.189889908 | 1.506361842 |
| IPI00753875 | SLpSADNFIGIQR | 154.47 | 0.721474349 | 0.772702277 |

| | | | | |
|---|---|---|---|---|
| IPI00653274 | LALVpTGGEIASTFDHPELVK | 124.1 | 7.553201675 | 10.19968319 |
| IPI00113849 | YVECpSALTQK | 28.71 | 1.808095932 | 1.343613386 |
| IPI00227808 | TAPVQAPPAPVTVTEpTPEPA(ox)MPSGVYRPPGAR | 54.09 | 2.251980066 | 4.621519566 |
| IPI00136703 | VLpTPELYAELR | 150.12 | 9.626471519 | 8.553112984 |
| IPI00828500 | EQTASAPApTPLVSK | 122.72 | 6.829474449 | 6.444790363 |
| IPI00828500 | ITVEKDPDSALGIpSDGETSPSSK | 69.06 | 0.748685696 | 0.699242297 |
| IPI00828500 | STpSVDDTDKSSSEAI(ox)MVR | 48.06 | 0.623449147 | 0.632381797 |
| IPI00109588 | GFPGPPGPDGLPGS(ox)MGPPGpTPSVDHGFLVTR | 118.15 | 1.560830474 | 2.374943972 |
| IPI00123891 | GFGFGQGAGALVHpSE | 202.39 | 0.770651281 | 0.768593609 |
| IPI00468516 | VDNARVpSPEVGSADVASIAQK | 78.04 | 0.537959574 | 0.474731256 |
| IPI00775841 | LpSSPVLHR<br>LSpSPVLHR | 56.74 | 1.061162829 | 0.739433467 |
| IPI00458127 | SAGAPRpTGEPEQEAVSR | 106.25 | 1.522797346 | 1.508806348 |
| IPI00881493 | SPSPSPpTSPGSLRK<br>SPSPSPTpSPGSLRK | 37.98 | 0.529263377 | 0.48207593 |
| IPI00130102 | TFGGAPGFSLGSPLpSSPVFPR<br>TFGGAPGFSLGSPLSpSPVFPR | 162.62 | 5.576372385 | 5.621459842 |
| IPI00130102 | TFGGAPGFSLGSPLSpSPVFPR | 123.02 | 6.713711262 | 5.781813383 |
| IPI00130102 | TFGGAPGFSLGpSPLSpSPVFPR | 110.44 | 2.479888678 | 3.487055063 |
| IPI00130102 | TFGGAPGFSLGpSPLSpSPVFPR<br>TFGGAPGFSLGSPLpSpSPVFPR | 138.17 | 2.067015171 | 2.880161524 |
| IPI00134809 | AKPAEpTPAPAHK | 52.96 | 24.51044846 | 23.23945236 |
| IPI00894804 | SLSTpSGESLYHVLGLDK | 147.4 | 0.546143889 | 1.006157994 |
| IPI00114375 | pTSPAKQQAPPVR<br>TpSPAKQQAPPVR | 61.05 | 0.526955187 | 0.374703526 |
| IPI00122349 | GSPpTRPNPPVR | 57.59 | 0.763181925 | 0.817926824 |
| IPI00153421 | SVSSNVASVpSPIPAGSKK | 95.78 | 11.63234901 | 8.854105949 |
| IPI00330066 | TRpSPDVISSASTALSQDIPEIASEALSR | 130.88 | 0.680407107 | 0.739012182 |
| IPI00307837 | VEpTGVLKPG(ox)MVVTFAPVNVTTEVK | 99.63 | 7.897317444 | 6.000155551 |
| IPI00307837 | VEpTGVLKPG(ox)MVVTFAPVNVpTTEVK | 90.42 | 5.384685516 | 4.658336163 |

| | | | | |
|---|---|---|---|---|
| | VEpTGVLKPG(ox)MVVTFAPVNVTpTEVK | | | |
| IPI00307837 | VEpTGVLKPG(ox)MVVTFAPVNVpTTEVK | 101.22 | 4.681442738 | 6.01963377 |
| IPI00620302 | GATPAEDDEDKDIDLFGpSDEEEEDKEAAR | 112.44 | 0.635066763 | 0.563856326 |
| IPI00115992 | LIpTPAVVSER | 106.41 | 11.07410049 | 9.419182777 |
| IPI00620806 | SEpSPKEPEQLR | 83.38 | 0.982602655 | 1.160811961 |
| IPI00409405 | LGGpSAVISLEGKPL | 56.67 | 15.88397408 | 14.62335873 |
| IPI00321647 | QPLLLpSEDEEDTKR | 98.01 | 0.70149942 | 0.686574039 |
| IPI00421179 | AApSLTEDR | 101.82 | 0.870357871 | 1.274669528 |
| IPI00421179 | EATLPPVpSPPK | 69.17 | 0.843857368 | 0.732594743 |
| IPI00648821 | EALELLKpTAIAK | 111.19 | 11.89870739 | 11.00475693 |
| IPI00648821 | YIpTPDQLADLYK | 174.3 | 0.869313419 | 1.290469646 |
| IPI00122684 | AAVPSGApSTGIYEALELR<br>AAVPSGASpTGIYEALELR | 191.05 | 35.18431346 | 37.44134808 |
| IPI00622968 | FAAATGApTPIAGR | 144.53 | 27.76387215 | 23.59006119 |
| IPI00173160 | LVQpSPNSYF(ox)MDVK | 119.47 | 3.612674475 | 1.551651597 |
| IPI00420187 | TVFAGAVPVLPApSPPPKDSLR | 179.98 | 0.913019776 | 0.814337611 |
| IPI00116753 | LNVAPVSDIIEIKpSPDTFVR | 76.62 | 6.699317932 | 5.57572031 |
| IPI00154109 | HGLLLPApSPVR | 100.39 | 1.859685421 | 2.273117542 |
| IPI00113223 | LFDHPEVPpTPPESApSVSR | 91.57 | 5.942027569 | 6.022670269 |
| IPI00113223 | LFDHPEVPTPPEpSASVSR<br>LFDHPEVPTPPESApSVSR | 103.96 | 2.781153917 | 4.825872898 |
| IPI00875567 | FGGEHVPNSPFQVTALAGDQPTVQpTPLR | 61.85 | 8.982500553 | 8.96617047 |
| IPI00875567 | VATVPQHApTSGPGPADVSK | 67.76 | 15.74446535 | 8.582057953 |
| IPI00663627 | VLFASQEIPApSPFR | 125.18 | 17.37866211 | 14.38440704 |
| IPI00664670 | LGpSFGSITR | 65 | 0.723294854 | 0.886812866 |
| IPI00331295 | SApTPPPAEPASLPQEPPKPR | 102.85 | 0.722537458 | 0.821812153 |
| IPI00111960 | TpSPTFFPK | 86.47 | 8.900146484 | 15.0602684 |
| IPI00111960 | YEVPLEpTPR | 91.73 | 10.62809086 | 14.5218792 |
| IPI00108811 | NFVDpSPIIVDIPK | 174.39 | 4.659234524 | 3.191150188 |

| | | | | |
|---|---|---|---|---|
| IPI00165716 | ADASSLTVDVpTSPASKVPTTVEDR<br>ADASSLTVDVTpSPASKVPTTVEDR | 124.1 | 4.119154453 | 3.174692631 |
| IPI00228633 | ELQAAGKpSPEDLEK | 70.15 | 33.58008194 | 38.31991959 |
| IPI00759948 | AALKpTASDFISK | 77.31 | 35.98693085 | 29.98003197 |
| IPI00759948 | DGGQpTAPASIR | 67.31 | 71.81306458 | 39.77776718 |
| IPI00153986 | GTSRPGpTPSAEAASTSSTLR<br>GTSRPGTPpSAEAASTSSTLR | 91.57 | 5.166966438 | 6.633099556 |
| IPI00828543 | TIpTLVKSPISVPGGSALISNLGK<br>TITLVKpSPISVPGGSALISNLGK | 71.39 | 5.665997982 | 3.815047979 |
| IPI00882349 | AGDVLEDpSPKRPK | 113.87 | 0.799336821 | 0.767371282 |
| IPI00882349 | GSAEGpSpSDEEGKLVIDEPAKEK | 65.28 | 0.958018771 | 1.259046912 |
| IPI00331556 | VLApTAFDTTLGGR | 100.39 | 1.710803151 | 1.556209803 |
| IPI00319992 | ITPSYVAFpTPEGER | 167.38 | 17.79181099 | 15.56201363 |
| IPI00323357 | IINEPpTAAAIAYGLDKK | 75.25 | 33.94918823 | 16.67408371 |
| IPI00133903 | TpTPSVVAFTADGER | 167.38 | 7.578380108 | 11.57881355 |
| IPI00468068 | AQIGGPEAGKpSEQSGAK | 44.84 | 58.96623993 | 107.8104477 |
| IPI00468068 | AQIGGPEAGKSEQpSGAK | 53.05 | 5.323335171 | 12.02662373 |
| IPI00468068 | AVTQpSAEITIPVTFEAR | 138.72 | 48.22555161 | 68.71230316 |
| IPI00468068 | QLpSSGVSEIR | 111.42 | 0.723811746 | 0.684148848 |
| IPI00224109 | IEpSPKLER | 86.47 | 0.748215318 | 0.636383057 |
| IPI00308971 | AEALSSLHGDDQDpSEDEVLTVPEVK | 78.04 | 0.768615842 | 0.734100819 |
| IPI00319956 | TEEVEVEpSEEDPILEHPPENPVK | 90.47 | 0.525107026 | 0.501573026 |
| IPI00889930 | AGGApSPAASSTTQPPAQHR | 84.26 | 0.786503077 | 0.675689697 |
| IPI00856218 | AApSALLLR | 154.78 | 0.723594725 | 0.692247629 |
| IPI00856218 | ELpSAPAR | 66.57 | 0.704663932 | 0.641427338 |
| IPI00331173 | SVpSASHEGDVK<br>SVSApSHEGDVK | 77.31 | 0.723999858 | 0.634566069 |
| IPI00751369 | VTLpTPEEEAR | 100.07 | 81.00878143 | 65.96704102 |
| IPI00112339 | ELpSVEEQIKR | 62.96 | 1.014968038 | 1.0225389 |

| | | | | |
|---|---|---|---|---|
| IPI00112339 | TSSIKpSPK | 83.38 | 2.33336401 | 2.328483582 |
| IPI00400300 | LRLpSPpSPTSQR<br>LRLSPpSPpTSQR | 67.31 | 0.811540782 | 0.797581792 |
| IPI00400300 | LSPpSPTSQR | 88.9 | 0.7242257 | 0.744147837 |
| IPI00230394 | ASAPATPLpSPTR | 85.46 | 4.817977428 | 3.001209259 |
| IPI00659447 | LAPpSPSEEPR | 83.81 | 0.614996552 | 0.736216784 |
| IPI00659447 | TSPGpSPSPR | 65 | 4.849502087 | 6.031260967 |
| IPI00757916 | AGDLGVDLpTSK | 81.51 | 12.79569435 | 12.37896442 |
| IPI00757916 | LPAVVpTADLR | 37.98 | 9.101158142 | 6.789206505 |
| IPI00849751 | IEEELGpSKAK | 83.81 | 12.49800158 | 15.97111464 |
| IPI00849165 | SApSPGLPKGEK | 105.56 | 0.706182406 | 0.570384 |
| IPI00606906 | AVLPGpSPIFSR | 131.17 | 9.220336795 | 9.256473896 |
| IPI00785324 | AENQRPAEDSALpSPGPLAGAK | 48.83 | 1.23474443 | 0.904589415 |
| IPI00319270 | TLTTAAVpSTAQPILSK<br>TLTTAAVSpTAQPILSK | 206.91 | 1.352540374 | 1.087141275 |
| IPI00830432 | QGQDVAPPPNPVPQRpTSPTGPK<br>QGQDVAPPPNPVPQRTpSPTGPK | 99.69 | 0.690261126 | 0.499007702 |
| IPI00229534 | AEDGAAPpSPSSETPK | 112.17 | 2.100413918 | 3.392316341 |
| IPI00229534 | AEDGAAPSPpSSETPKK<br>AEDGAAPSPSpSETPKK | 92.17 | 9.322688103 | 3.320444107 |
| IPI00229534 | AEDGAAPpSPSSETPKKK<br>AEDGAAPSPpSSETPKKK<br>AEDGAAPSPSpSETPKKK<br>AEDGAAPSPSSEpTPKKK | 91.91 | 10.36584091 | 3.589919329 |
| IPI00229534 | AEDGAAPpSPSSEpTPKKK | 74.81 | 5.159576893 | 3.845372438 |
| IPI00229534 | GEATAERPGEAAVApSSPSK<br>GEATAERPGEAAVASpSPSK | 103.82 | 0.772615953 | 1.101140915 |
| IPI00323820 | ISDPLpTSSPGR<br>ISDPLTpSSPGR<br>ISDPLTSpSPGR | 45.48 | 0.456634462 | 0.542350292 |
| IPI00323592 | NLGIGKIpTPFEEK | 119.47 | 16.57404327 | 9.897894859 |

| | | | | |
|---|---|---|---|---|
| IPI00323592 | VAVLGASGGIGQPLpSLLLKNSPLVSR | 156.23 | 5.287952423 | 4.215777397 |
| IPI00323592 | VAVLGASGGIGQPLSLLLKNpSPLVSR | 130.17 | 5.2212677 | 3.951768398 |
| IPI00315808 | GSTpSPDLL(ox)MHQGPPDTAEIIK | 93.44 | 3.265588999 | 1.351506114 |
| IPI00408909 | ALALVPGpTPTR | 89.76 | 7.481154442 | 5.125487804 |
| IPI00896700 | ETAAAHQApSSSPPIDAATAEPYGFR<br>ETAAAHQASpSSPPIDAATAEPYGFR<br>ETAAAHQASSpSPPIDAATAEPYGFR | 35.34 | 0.662014365 | 0.767188072 |
| IPI00896700 | RSEpSPFEGK | 103.12 | 0.78600955 | 0.6921826 |
| IPI00896700 | SDIpSPLTPR | 91.73 | 0.249109641 | 0.321611315 |
| IPI00896700 | SPSLSPSPPpSPIEK | 78.04 | 0.783955634 | 0.786213338 |
| IPI00896700 | SVpSPGVTQAVVEEHCASPEEK | 20.59 | 0.747820914 | 0.549286246 |
| IPI00896700 | TTpTTPEVK<br>TTTpTPEVK | 72.16 | 4.88103199 | 4.319426537 |
| IPI00896700 | VLpSPLRSPPLLGSESPYEDFLSADSK | 81.42 | 1.197787881 | 0.923976958 |
| IPI00896700 | VLpSPLRpSPPLLGSESPYEDFLSADSK | 91.91 | 1.015276074 | 1.005762339 |
| IPI00896700 | VLpSPLRpSPPLLGSESPYEDFLSADSK<br>VLpSPLRSPPLLGpSESPYEDFLSADSK | 93.74 | 1.005020618 | 1.075676441 |
| IPI00896700 | VLpSPLRSPPLLGSESPpYEDFLSADSK<br>VLSPLRpSPPLLGSESPpYEDFLSADSK | 71.39 | 0.907762647 | 0.704117596 |
| IPI00896700 | VLpSPLRpSPPLLGSESPYEDFLSADSK | 109.03 | 0.859545767 | 0.744838953 |
| IPI00896700 | VLSPLRpSPPLLGSESPYEDFLSADSK | 154.48 | 0.641937673 | 0.583019137 |
| IPI00408119 | AAVGVTGNDITTPPNKEPPPpSPEKK | 78.04 | 1.094480515 | 0.926650345 |
| IPI00408119 | ALET(ox)MAEQTTDVVHpSPSTDTTPGPDTEAALAK | 140.55 | 1.170022845 | 0.966159225 |
| IPI00408119 | D(ox)MpSPLPESEVTLGKDVVILPETK | 70.19 | 2.095468521 | 1.406081438 |
| IPI00310519 | AELAHpSPLPAK | 150.12 | 49.89553452 | 20.81400681 |
| IPI00123199 | LDGLVDpTPTGYIESLPK<br>LDGLVDTPpTGYIESLPK | 111.87 | 10.2134161 | 14.16672611 |
| IPI00123199 | LDGLVDpTPTGYIESLPK | 171.74 | 4.824375153 | 5.192005157 |
| IPI00845608 | APLVGpSPVHLGPSQPLK | 158.15 | 1.130340576 | 1.063784406 |
| IPI00845608 | ARpTPTLASpTPIPP(ox)MSEAPYPK | 98.36 | 6.97160244 | 2.749439478 |

| | | | | |
|---|---|---|---|---|
| IPI00845608 | ESQEFLRpSPEAEEEEEQV(ox)MVR | 121.98 | 1.00393486 | 1.102925181 |
| IPI00845608 | QEpSLKSPEEEDQQAFR<br>QESLKpSPEEEDQQAFR | 68.56 | 1.111553073 | 1.09843564 |
| IPI00845608 | VSQVSLESLEKENVQpSPR | 46.94 | 2.127660751 | 1.682767034 |
| IPI00400168 | AEIKE(ox)MLApSDDEEESSPK | 67.05 | 0.408290505 | 1.04345119 |
| IPI00127417 | V(ox)MLGETNPADpSKPGTIR | 67.05 | 0.636349976 | 1.54293561 |
| IPI00341869 | ATVTPpSPVKGK | 89.76 | 0.548258601 | 0.37350271 |
| IPI00469331 | GVLSpSPSLAFTTPIR | 62.87 | 2.772338629 | 3.383033514 |
| IPI00126313 | PLHYLpTILpSPR | 48.2 | 9.889798641 | 6.283292055 |
| IPI00282266 | EIQpTAVR | 63.74 | 11.28819752 | 17.82034302 |
| IPI00624863 | EVYELLDpTPGR | 119.71 | 12.44848633 | 18.17038918 |
| IPI00229884 | IIpSIFSGTEK | 97.24 | 0.746232986 | 0.614436567 |
| IPI00762775 | SVPGVTpSTPHSK<br>SVPGVTSpTPHSK<br>SVPGVTSTPHpSK | 81.51 | 29.33709908 | 15.29289818 |
| IPI00762775 | TLpS(ox)MIEEEIR | 130.07 | 0.489557877 | 0.452896486 |
| IPI00470003 | SPLVPKpSPTPK | 52.96 | 2.232306719 | 1.605970263 |
| IPI00470003 | SPLVPKpSPTPKpSPPSR | 99.69 | 1.364190578 | 1.398820877 |
| IPI00470003 | SPTPKpSPPSR | 71.15 | 0.139414813 | 0.046658279 |
| IPI00453818 | EPSAPSIPPPAYQSSPAAGHAAAPPpTPAPR | 73.45 | 4.204166889 | 4.657279968 |
| IPI00337893 | YG(ox)MGpTSVER<br>YG(ox)MGTpSVER | 30.65 | 0.642371833 | 0.643605232 |
| IPI00309768 | SA(ox)MPFpTASPAPSTR | 116.92 | 41.06153107 | 38.69659424 |
| IPI00153375 | FSpSLDLEEDSEVFK | 214.41 | 0.587694585 | 0.585939825 |
| IPI00828969 | EVVKPVPITpSPAVSK | 100.39 | 46.95588112 | 29.39895439 |
| IPI00405307 | ALpTPPADPPR | 61.05 | 1.889647484 | 1.495816231 |
| IPI00336400 | DIIRQPpSEEEIIK | 100.39 | 0.512478398 | 0.587568969 |
| IPI00336400 | QPpSEEEIIK | 91.73 | 0.511900408 | 0.555279012 |
| IPI00387312 | TLpSIDKGF | 72.16 | 0.661338091 | 0.634651005 |

| | | | | |
|---|---|---|---|---|
| IPI00555069 | AHpSS(ox)MVGVNLPQK<br>AHSpS(ox)MVGVNLPQK | 111.19 | 2.046526432 | 0.741505384 |
| IPI00555069 | ALEpSPERPFLAILGGAK | 90.47 | 5.652894735 | 7.445711851 |
| IPI00132080 | IVAPISDpSPKPPPQR | 62.3 | 48.8340683 | 45.37990189 |
| IPI00319973 | EGEEPTVYpSDDEEPKDETAR | 107.82 | 1.366412997 | 1.955558419 |
| IPI00319973 | LLKEGEEPpTVYSDDEEPKDETAR<br>LLKEGEEPTVYpSDDEEPKDETAR | 109.03 | 0.945469558 | 0.735155523 |
| IPI00319973 | LLKEGEEPTVYpSDDEEPKDETAR | 86.23 | 0.785713136 | 0.698052168 |
| IPI00400381 | SSpSPVLVEEEPER | 122.16 | 0.70067066 | 0.660208404 |
| IPI00407130 | EApTESFASDPILYRPVAVALDTK | 100.11 | 1.235897183 | 2.331736326 |
| IPI00407130 | LDIDSAPIpTAR | 102.9 | 12.86094761 | 15.59814262 |
| IPI00626385 | EREEGAPEpTPVVSATTVGTLAR | 81.03 | 3.078224182 | 2.917778015 |
| IPI00120546 | ALDDFVLGpSAR | 57.59 | 1.821764469 | 1.868702412 |
| IPI00831115 | SAPDFpTATAVVDGAFKEIK<br>SAPDFTApTAVVDGAFKEIK | 72.63 | 1.674755454 | 2.067908049 |
| IPI00875652 | TDSREDEIpSPPPPNPVVK | 127.75 | 0.72280697 | 0.645957748 |
| IPI00126939 | GESALEPGPVPEpTPAGGPVHAVTVVTLLEK | 58.08 | 0.668460131 | 0.959567428 |
| IPI00875405 | TASAVAGKpTPDASPEPK | 130.95 | 5.994947195 | 7.970684767 |
| IPI00381291 | AAAASAAEAGIApTPGTEGERDSDDALLK | 185.98 | 4.361035824 | 4.936915278 |
| IPI00759871 | IITGPAPVLPPAALRpTPTPAGPTI(ox)MPLIR | 68.55 | 3.563090324 | 4.916073799 |
| IPI00830159 | AGQGIPAPPEASPTAVPEPpSTPFPPVLASG(ox)MSHPPPTSR | 51.81 | 2.539334059 | 12.50415802 |
| IPI00828741 | GRLpSPVPVPR | 61.05 | 1.472668052 | 1.775638342 |
| IPI00135190 | TSpSLTHSEEK<br>TSSLpTHSEEK | 126.36 | 0.521138012 | 1.325824022 |
| IPI00649362 | QPpTPPFFGR | 54.78 | 0.816814005 | 0.618681371 |
| IPI00124826 | LEPAPLDpSSPAVSTHEGSK<br>LEPAPLDSpSPAVSTHEGSK | 149.91 | 4.544881105 | 4.156165838 |
| IPI00133185 | AAILKApSPK | 91.73 | 3.479880571 | 4.195896149 |
| IPI00742383 | LASVPAGGAVAVpSAAPGSAAPAAGSAPAAAEEKKDEK | 16.12 | 4.492766857 | 3.956872225 |
| IPI00742383 | YVASYLLAALGGNSpSPSAK | 78.04 | 1.471876264 | 0.945997953 |

| | | | | |
|---|---|---|---|---|
| IPI00762542 | NIGLGFKpTPK | 126.36 | 29.15384865 | 25.75343132 |
| IPI00323819 | LIDLHpSPSEIVK | 139.42 | 9.320585251 | 11.50042343 |
| IPI00469392 | RGSGSVDEpTLFALPAASEPVIPSSAEK | 87.23 | 0.895802438 | 1.101232767 |
| IPI00122174 | APTAALpSPEPQDSKEDVK | 49.83 | 1.04063201 | 0.9422822 |
| IPI00114733 | DQAVENILLpSPLVVASSLGLVSLGGK | 56.79 | 5.865635395 | 2.852160454 |
| IPI00895328 | RPTEAVpSPK | 54.78 | 0.713203669 | 0.70842886 |
| IPI00116331 | APDRpTPPSEEDSAEAER | 101.93 | 0.847935557 | 0.838952303 |
| IPI00116331 | SRTPpSASHEEQQE | 48.06 | 0.878124237 | 1.217524052 |
| IPI00284016 | VGGPLAVLGPSRpSSEDLAGPLPSSVPSSTTSSKPK<br>VGGPLAVLGPSRSpSEDLAGPLPSSVPSSTTSSKPK | 77.33 | 0.483411729 | 0.411961019 |
| IPI00885294 | APpSPTDLPESEIKK | 94.59 | 1.095218897 | 0.946309865 |
| IPI00222090 | LGPSpSPAHSGALDLDGVSR | 108.08 | 5.270172596 | 4.791478634 |
| IPI00454008 | SAIpTPGGLR | 94.8 | 7.610964775 | 7.379424095 |
| IPI00109311 | EALVEPASEpSPRPALAR | 138.72 | 0.782189965 | 0.804872453 |
| IPI00310561 | VEpTPVLPPVLVPR | 108.66 | 3.890271187 | 3.738552809 |
| IPI00123129 | ETDGSEpTPEPFAAEAK | 195.31 | 2.279402971 | 1.93349576 |
| IPI00900438 | SFISSSPpSSPSR<br>SFISSSPSpSPSR | 85.46 | 1.04462719 | 1.257072449 |
| IPI00900438 | TpSPGRADLPGSSSTFTK | 125.18 | 0.713234663 | 0.778364092 |
| IPI00319830 | RPPpSPDPNTK | 46.8 | 0.907872021 | 1.141511083 |
| IPI00621617 | VTFVDpTPGIIENR | 133.28 | 14.18328571 | 15.11098194 |
| IPI00648313 | TApSPPPPPK | 94.8 | 0.656178534 | 0.534967542 |
| IPI00648313 | VSVpSPGR | 55.28 | 0.742977858 | 0.747279286 |
| IPI00785240 | RVPSPpTPVPK | 39.19 | 1.10208106 | 1.101135373 |
| IPI00785240 | RVPpSPTPVPK | 71.15 | 0.738718808 | 0.584268093 |
| IPI00785240 | TAVAPSAVNLADPRpTPAASAVNLAGAR | 121.36 | 6.590248585 | 10.06458187 |
| IPI00785240 | TPAAAAA(ox)MNLApSPR | 55.41 | 4.672698021 | 2.757165194 |
| IPI00649157 | ASGQAFELILpSPR | 198.81 | 10.85896397 | 6.697172483 |
| IPI00649157 | ESVPDFPLpSPPK | 98.01 | 0.638173095 | 0.777118827 |

| | | | | |
|---|---|---|---|---|
| IPI00649157 | RApSGQAFELILpSPR | 85.46 | 1.067178369 | 1.163950205 |
| IPI00471361 | LGTGGGGpSPDKSPSAQELK | 36.43 | 0.818988562 | 0.881100774 |
| IPI00626106 | NFpSDNQLQEGK | 166.42 | 0.806705952 | 0.641806185 |
| IPI00226205 | GHLLLAApTPGLAGR | 170.79 | 0.68735075 | 0.55401063 |
| IPI00469012 | ASAGPVPGAVVIAEGLHPSLPSPpTGNSTPLGTSK | 37.14 | 0.732137561 | 2.496587038 |
| IPI00469012 | AVGGAPpSPPPPVRR | 50.46 | 0.703671455 | 0.679569304 |
| IPI00127008 | IDIpSPSTFR | 46.61 | 0.686862171 | 0.794312477 |
| IPI00649283 | ASVSDLpSPR | 106.41 | 0.878090203 | 0.842149854 |
| IPI00874522 | AVPVpSPSAVEEDEDEDGHTVVATAR | 65.28 | 2.167450011 | 1.521011115 |
| IPI00874522 | AVPVpSPSAVEEDEDEDGHTVVATAR<br>AVPVSPpSAVEEDEDEDGHTVVATAR | 138.22 | 2.054923415 | 2.129760623 |
| IPI00896574 | GPPDFpSSDEEREPTPVLGSGASVGR<br>GPPDFSpSDEEREPTPVLGSGASVGR | 142.65 | 0.635078788 | 0.711198151 |
| IPI00459443 | SSGSLpSPGLETEDPLEAR | 73.92 | 1.54477489 | 1.255101442 |
| IPI00459443 | VSGAGLpSPSRK | 26.52 | 0.880207896 | 0.679489434 |
| IPI00378438 | AVNPT(ox)MAAPGpSPSLSHR | 112.17 | 2.11969614 | 1.946917772 |
| IPI00378438 | QGpSPTPALPEKR | 100.39 | 0.702841759 | 0.673131227 |
| IPI00652758 | GDLSQHApTPLPTPAVLPGDSPITPTPEQIGK<br>GDLSQHATPLPpTPAVLPGDSPITPTPEQIGK | 82.58 | 3.229072332 | 3.116264582 |
| IPI00652758 | GDLSQHATPLPpTPAVLPGDSPITPTPEQIGK | 89.72 | 2.943491697 | 2.671375036 |
| IPI00831423 | AISEELDHALND(ox)MTpSI | 141.67 | 0.970936147 | 0.480093483 |
| IPI00880644 | TDGFAEAIHpSPQVAGVPR | 90.47 | 0.896591127 | 1.482336044 |
| IPI00312128 | LDLDLpTSDSQPPVFK | 144.27 | 6.443159103 | 5.571760178 |
| IPI00877238 | SEDRPpSSPQVSVAAVETK<br>SEDRPSpSPQVSVAAVETK | 89.63 | 0.765558768 | 0.669963344 |
| IPI00110753 | AVFVDLEPpTVIDEVRpTGTYR | 144.98 | 4.930887222 | 9.076258659 |
| IPI00110753 | AVFVDLEPpTVIDEVRTGTpYR | 88.83 | 4.930887222 | 9.076258659 |
| IPI00110753 | AVFVDLEPpTVIDEVRTGTYR | 60.78 | 3.473395824 | 3.943846226 |
| IPI00169463 | I(ox)MNTFSVVPpSPK | 62.3 | 14.58024406 | 7.520189285 |

| | | | | |
|---|---|---|---|---|
| IPI00889248 | QLQpSPFILDEDQAR | 164.13 | 5.961415768 | 4.067269802 |
| IPI00123313 | ATLPpSPDKLPGFK | 106.25 | 55.15907478 | 34.80770683 |
| IPI00123313 | SDpTAAAAVR | 126.13 | 47.94510269 | 33.9345932 |
| IPI00123589 | SSPPATDPGPVPpSSPSQEPPTKR<br>SSPPATDPGPVPSpSPSQEPPTKR | 95.15 | 5.447118282 | 5.730722427 |
| IPI00881557 | IPYpTPGEIPK | 49.68 | 10.09185886 | 7.122915268 |
| IPI00881557 | VLIGGDEpTPEGQK | 103.96 | 20.14777565 | 16.13555527 |
| IPI00404693 | NLLEDDpSDEEEDFFLR | 226.15 | 0.709050298 | 0.96781987 |
| IPI00308187 | SLpTSPLDDTEVKK<br>SLTpSPLDDTEVKK | 100.39 | 43.05687332 | 77.48855591 |
| IPI00308187 | SLTpSPLDDTEVKK | 75.38 | 12.88979244 | 18.96881676 |
| IPI00126072 | LPPLPVpTPG(ox)MEGAGVVVAVGEGVGDR | 64.66 | 3.031758547 | 2.699586749 |
| IPI00227299 | ISLPLPpTFSSLNLR | 101.78 | 35.70420074 | 51.64565659 |
| IPI00227299 | ISLPLPTFpSSLNLR | 125.18 | 2.674481869 | 2.110260725 |
| IPI00227299 | LLQDpSVDFSLADAINTEFK | 74.81 | 1.879904628 | 1.739447713 |
| IPI00227299 | LRpSSVPGVR | 91.73 | 92.36534119 | 57.50434875 |
| IPI00227299 | SLYSpSSPGGAYVTR<br>SLYSSpSPGGAYVTR | 152.78 | 11.71202087 | 9.97661972 |
| IPI00108989 | QASTDAGpTAGALTPQHVR | 84.26 | 0.463960841 | 0.618706474 |
| IPI00387422 | GPLSQAPpTPAPK | 158.19 | 24.32490158 | 12.10877228 |

* Phosphorylated residues are indicated with a "p" and oxidated methiones with "(ox)".

**Supplementary Table 4.2**. Identified p38α substrates and their phosphorylation sites.

| UniProt* | Recommended Name | Site(s) |
|---|---|---|
| Q3TXS7 | 26S proteasome non-ATPase regulatory subunit 1 | T311 |
| O35226-2 | 26S proteasome non-ATPase regulatory subunit 4 | T250 |
| Q9CQX8 | 28S ribosomal protein S36, mitochondrial | S60 |
| P62281 | 40S ribosomal protein S11 | T46 |
| P60867 | 40S ribosomal protein S20 | S93 |
| P62852 | 40S ribosomal protein S25 | T69 |
| Q6ZWU9<br>Q6ZWY3 | 40S ribosomal protein S27<br>40S ribosomal protein S27-like | S27 |
| P14206 | 40S ribosomal protein SA | T97 |
| Q9CQ60 | 6-phosphogluconolactonase | S178 |
| Q9CR57 | 60S ribosomal protein L14 | S139 |
| P20029 | 78 kDa glucose-regulated protein | T70 |
| O54931 | A-kinase anchor protein 2 | S18, T19, S22 |
| Q8QZT1 | Acetyl-CoA acetyltransferase, mitochondrial | T233 |
| Q99KI0 | Aconitate hydratase, mitochondrial | S559 |
| P68134<br>P68033<br>P62737 | Actin, alpha skeletal muscle<br>Actin, alpha cardiac muscle 1<br>Actin, aortic smooth muscle | S35, T105, T108, T320 |
| Q9QZ83 | Actin-like protein Gamma | T110, T322 |
| P17182 | Alpha-enolase | S40, T41, T229, S419 |
| P10107 | Annexin A1 | S37 |
| IPI00310240 † | Annexin A6 isoform b | T529 |
| Q7TQH0 | Ataxin-2-like protein | S687 |
| P56480 | ATP synthase subunit beta, mitochondrial | S128 |
| Q9CPQ8 | ATP synthase subunit g, mitochondrial | T42 |
| Q8VCQ8 | Caldesmon 1 | S463, S465, T467 |
| P35564 | Calnexin | T67 |
| P51125 | Calpastatin | T624 |
| P18760 | Cofilin-1 | S156 |
| Q9D1L0<br><br>Q8CEW7 | Coiled-coil-helix-coiled-coil-helix domain-containing protein 2, mitochondrial<br>Putative uncharacterized protein | S45 |
| Q3UMF0 | Cordon-bleu protein-like 1 | T304 |
| Q04447 | Creatine kinase B-type | T35 |
| Q8R1Q8 | Cytoplasmic dynein 1 light intermediate chain 1 | S421 |
| Q9D2G2 | Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex, mitochondrial | T159 |
| Q99LC5 | Electron transfer flavoprotein subunit alpha, mitochondrial | S140 |
| Q9DCW4 | Electron transfer flavoprotein subunit beta | T182, T219 |

| | | |
|---|---|---|
| P10126 | Elongation factor 1-alpha 1 | T269, T286, T287 |
| P19096 | Fatty acid synthase | T976, S982 |
| Q8BTM8 | Filamin-A | T1750, T2549 |
| Q80X90 | Filamin-B | S833 |
| P05064 | Fructose-bisphosphate aldolase A | T37 |
| P13020 | Gelsolin | T359, T556 |
| Q8R5B7 | General transcription factor IIF, polypeptide 1 | T389, S391 |
| P06745 | Glucose-6-phosphate isomerase | S455 |
| P17439 | Glucosylceramidase | S418 |
| P16858 | Glyceraldehyde-3-phosphate dehydrogenase | T209 |
| Q99JX3 | Golgi reassembly-stacking protein 2 | T417, S418 |
| P14602 | Heat shock 27 kDa protein | S180, S203, S206 |
| P17156<br>P63017 | Heat shock-related 70 kDa protein 2<br>Heat shock cognate 71 kDa protein | T178/T177 |
| P70696<br>Q64475<br>Q6ZWY9<br>P10853<br>Q64478<br>P10854<br>Q64525 | Histone H2B type 1-A<br>Histone H2B type 1-B<br>Histone H2B type 1-C/E/G<br>Histone H2B type 1-F/J/L<br>Histone H2B type 1-H<br>Histone H2B type 1-M<br>Histone H2B type 2-B | T97/T98 |
| Q61191 | Host cell factor 1 | T662, S666 |
| P06151 | L-lactate dehydrogenase A chain | T309 |
| P14733 | Lamin-B1 | S24 |
| A0T1J8 | LIM domain only 7 | S1602 |
| P70699 | Lysosomal alpha-glucosidase | S156, T197 |
| P08249 | Malate dehydrogenase, mitochondrial | S41, S47, T309 |
| Q9DBV4 | Matrix-remodeling-associated protein 8 | S423 |
| Q9QYR6 | Microtubule-associated protein 1A | T2182 |
| P14873 | Microtubule-associated protein 1B | T2300, T2301 |
| Q62432<br>Q8BUN5<br>Q9JIW5 | Mothers against decapentaplegic homolog 2<br>Mothers against decapentaplegic homolog 3<br>Mothers against decapentaplegic homolog 9 | T172/T132/T136 |
| Q9DCL9 | Multifunctional protein ADE2 | T27 |
| P26645 | Myristoylated alanine-rich C-kinase substrate | S140, S141, T143 |
| P82343 | N-acylglucosamine 2-epimerase | S419, S420 |
| Q6P5H2 | Nestin | T383, T389 |
| IPI00553798 † | Neuroblast differentiation-associated protein AHNAK | S232, T423, T551, S692, T736, T1165, S1166, S2381, T3094, S3139, S3140, T4342, T4773, T4775, |

| | | |
|---|---|---|
| | | T4779, S4890, S4905, T5169, S5194, S5195, S5325, S5566, T5567 |
| P28656 | Nucleosome assembly protein 1-like 1 | T62, T64 |
| Q8VG12 | Olfactory receptor MOR245-1 | T134, S137 |
| O70400 | PDZ and LIM domain protein 1 | T128 |
| Q8CI51 | PDZ and LIM domain protein 5 | S111 |
| P09411 | Phosphoglycerate kinase 1 | S203 |
| Q9QXS1 | Plectin | T158 |
| Q9WU78 | Programmed cell death 6-interacting protein | T741 |
| Q9Z2U1 | Proteasome subunit alpha type-5 | S56 |
| Q9QYS9 | Protein quaking | T243 |
| P52480 | Pyruvate kinase isozymes M1/M2 | T41 |
| Q5I1X5 | RelA-associated inhibitor | S394 |
| Q99PT1 | Rho GDP-dissociation inhibitor 1 | T160 |
| Q7TQ48 | Sarcalumenin | T628 |
| Q9CZN7 | Serine hydroxymethyltransferase | T420 |
| Q8BTI8 | Serine/arginine repetitive matrix protein 2 | S2224, T2241 |
| P19324 | Serpin H1 | S69 |
| A2AAY5 | SH3 and PX domain-containing protein 2B | S291 |
| P38647 | Stress-70 protein, mitochondrial | T87 |
| Q62465 | Synaptic vesicle membrane protein VAT-1 homolog | T122 |
| P80314 | T-complex protein 1 subunit beta | T327 |
| O88746 | Target of Myb protein 1 | T196 |
| Q62318 | Transcription intermediary factor 1-beta | T498 |
| P68369<br>P05213<br>P68373 | Tubulin alpha-1A chain<br>Tubulin alpha-1B chain<br>Tubulin alpha-1C chain | T73, T80 |
| P99024<br>P68372<br>Q9ERD7<br>Q9D6F9 | Tubulin beta-2C chain<br>Tubulin beta-3 chain<br>Tubulin beta-4 chain<br>Tubulin beta-5 chain | S172 |
| Q02053 | Ubiquitin-like modifier-activating enzyme 1 | T531, S835 |
| Q922Y1 | UBX domain-containing protein 1 | S199, S200 |
| O70475 | UDP-glucose 6-dehydrogenase | T185, T474 |
| Q3TW96 | UDP-N-acetylhexosamine pyrophosphorylase-like protein 1 | S490 |
| Q9QY76 | Vesicle-associated membrane protein-associated protein B | T158, S159 |
| P20152 | Vimentin | S55, S56, S72, T417 |
| Q62523 | Zyxin | T252 |

*Certain phosphopeptides matched multiple proteins and therefore all matching proteins are listed with their UniProt accession numbers and recommended names. If the site of phosphorylation varies, this is indicated by a slash (for example T12/T13).

[†]The IPI accession number is listed for annexin A6 isoform b and neuroblast differentiation-associated protein AHNAK as they are not found in the UniProt database.

**Supplementary Table 4.3**. Cytosolic p38α substrates. The 55 cytosolic p38α substrates are listed below by UniProt ID and gene name as in **Figure 4.5c**. If a substrate is known to be critical for fusion or differentiation it is indicated in the fourth column with the associated reference.

| UniProt | Gene | Full name | Function |
|---|---|---|---|
| **P68134** | Acta1 | Actin, alpha skeletal muscle | fusion (myoblast)[182, 183] |
| **Q9QZ83** | Actg1 | Gamma actin-like protein | - |
| **P10107** | Anxa1 | Annexin A1 | differentiation (myoblast)[184] |
| **IPI00310240*** | Anxa6b | Annexin A6 isoform b | - |
| **Q99PT1** | Arhgdia | Rho GDP-dissociation inhibitor 1 | - |
| **Q7TQH0** | Atxn2l | Ataxin-2-like protein | - |
| **Q8VCQ8** | Cald1 | Caldesmon 1 | - |
| **P51125** | Cast | Calpastatin | fusion (myoblast)[185] |
| **P80314** | Cct2 | T-complex protein 1 subunit beta | - |
| **P18760** | Cfl1 | Cofilin-1 | differentiation (myofibroblast)[186] |
| **Q9D2G2** | Dlst | Dihydrolipoyllysine-residue succinyltransferase component of 2-oxoglutarate dehydrogenase complex, mitochondrial | - |
| **Q8R1Q8** | Dync1li1 | Cytoplasmic dynein 1 light intermediate chain 1 | differentiation (germ cells *C. elegans*)[187] |
| **P10126** | Eef1a1 | Elongation factor 1-alpha 1 | - |
| **P17182** | Eno1 | Alpha-enolase | differentiation (myoblast)[188] |
| **Q8BTM8** | Flna | Filamin-A | - |
| **Q80X90** | Flnb | Filamin-B | neg. regulation of differentiation (myoblast)[189, 190] |
| **P16858** | Gapdh | Glyceraldehyde-3-phosphate dehydrogenase | - |
| **P06745** | Gpi | Glucose-6-phosphate isomerase | - |
| **P13020** | Gsn | Gelsolin | differentiation (adipocyte)[191] |
| **Q61191** | Hcfc1 | Host cell factor 1 | differentiation (myoblast)[192] |
| **P14602** | Hspb1 | Heat shock protein beta-1 | differentiation (neuronal, erythroid)[193, 194] |
| **P06151** | Ldha | L-lactate dehydrogenase A chain | - |
| **A0T1J8** | Lmo7 | LIM domain only 7 | - |
| **Q9QYR6** | Map1a | Microtubule-associated protein 1A | - |
| **P14873** | Map1b | Microtubule-associated protein 1B | differentiation (neuron)[195] |
| **P26645** | Marcks | Myristoylated alanine-rich C-kinase substrate | fusion (myoblast)[196, 197] |
| **Q9DBV4** | Mxra8 | Matrix-remodeling-associated protein 8 | - |
| **Q6P5H2** | Nes | Nestin | neg. regulation of differentiation (myoblast)[198] |
| **Q8VG12** | Olfr1316 | MCG141903 | - |
| **Q9DCL9** | Paics | Multifunctional protein ADE2 | - |
| **Q9WU78** | Pdcd6ip | Programmed cell death 6-interacting protein | - |
| **O70400** | Pdlim1 | PDZ and LIM domain protein 1 | - |

| | | | |
|---|---|---|---|
| **Q8CI51** | Pdlim5 | PDZ and LIM domain protein 5 | - |
| **P09411** | Pgk1 | Phosphoglycerate kinase 1 | - |
| **Q9CQ60** | Pgls | 6-phosphogluconolactonase | - |
| **Q9QXS1** | Plec | Plectin | - |
| **Q5I1X5** | Ppp1r13l | RelA-associated inhibitor | differentiation (trophoblast)[199] |
| **Q9Z2U1** | Psma5 | Proteasome subunit alpha type-5 | - |
| **Q3TXS7** | Psmd1 | 26S proteasome non-ATPase regulatory subunit 1 | - |
| **O35226** | Psmd4 | 26S proteasome non-ATPase regulatory subunit 4 | - |
| **Q9QYS9** | Qki | Protein quaking | differentiation (smooth muscle)[200] |
| **P82343** | Renbp | N-acylglucosamine 2-epimerase | - |
| **P14206** | Rpsa | 40S ribosomal protein SA | - |
| **A2AAY5** | Sh3pxd2b | SH3 and PX domain-containing protein 2B | differentiation (adipocyte)[201] |
| **O88746** | Tom1 | Target of Myb protein 1 | - |
| **P68369** | Tuba1a | Tubulin alpha-1A chain | differentiation (neuron)[202] |
| **P99024** | Tubb5 | Tubulin beta-5 chain | - |
| **Q3TW96** | Uap1l1 | UDP-N-acetylhexosamine pyrophosphorylase-like protein 1 | - |
| **Q02053** | Uba1 | Ubiquitin-like modifier-activating enzyme 1 | - |
| **Q922Y1** | Ubxn1 | UBX domain-containing protein 1 | - |
| **O70475** | Ugdh | UDP-glucose 6-dehydrogenase | - |
| **Q9QY76** | Vapb | Vesicle-associated membrane protein-associated protein B | - |
| **Q62465** | Vat1 | Synaptic vesicle membrane protein VAT-1 homolog | - |
| **P20152** | Vim | Vimentin | neg. regulation of differentiation (myoblast)[198] |
| **Q62523** | Zyx | Zyxin | - |

*The IPI accession number is listed for Anxa6b as it is not found in the UniProt database.

# CHAPTER 5: General discussion

**Identity derived from the kinase catalytic domain**

The protein kinases analyzed in **Chapter 2** represent six of the seven major groups (no TKL kinases in active conformations were available), two members of the atypical group, as well as a bacterial and a yeast kinase with no metazoan orthologs. Our analyses demonstrate the high degree of structural conservation that exists within the protein kinase superfamily, at least for those regions involved in the conserved catalytic mechanism. ATP binding and phosphotransfer are functions all active kinases perform and so it is natural to expect high similarity in the areas involved. However, these functions occupy a large part of the catalytic domain, leaving only small well-defined regions where variability can be incorporated without disrupting the phosphotransfer mechanism. By contrast, surface residues on the exterior of the kinase domain show little conservation. These residues play prominent roles in binding other proteins, and the variability in these areas indicates the potential for different protein interactions between kinases. Some of these interactions will not result in phosphorylation of the bound protein, but instead may affect kinase activation, localization, etc., though many interactions will be for the purpose of phosphorylation. As shown in **Chapter 2**, a great deal of variability is found in surface residues located within the substrate-binding groove, reflecting the potential for this area to be a major determinant of target specificity and hence kinase identity. It might be suggested the variability present in this region does not reflect the evolution of substrate specificity but instead reflects the lack of functional importance for this region, and hence a lack of evolutionary constraint. However, this area is well conserved between orthologous kinases, suggesting it is functionally important. It cannot be ruled out that this region of the catalytic domain may be

optimized for functions besides or in addition to substrate binding; too little is currently known about the mechanisms of substrate binding to be definitive. Whatever role this region plays, it is likely a critical element determining any identity a catalytic domain has in isolation from its other parts, i.e. non-kinase domain sequence.

The conservation found within the catalytic domain extends beyond the amino acids contained in the protein sequence to functionally important water molecules located near the active site (**Chapter 3**). Although we have studied water molecule conservation across the superfamily, we have not examined conservation that may be particular to a group, family or subfamily. There may be important water conservation within a division of the superfamily that confers a unique functional feature upon certain related kinases, or even individual members for that matter. The liquid environment a kinase is found in could be essential to the establishment of its identity. In this regard, other components found within the environment of the cell could be equally important across the kinase superfamily or amongst particular members. The concentration of sodium and potassium ions can change rapidly and drastically within a cell, and just as water molecules are functionally important to kinases, it is likely these ions are as well. It is natural to assume so as ions such as sodium and potassium can be present at very high concentrations and would be something a kinase would come into constant access with. The incorporation of these ions - or other ubiquitous molecules for that matter – as part of the functional gestalt of the kinase domain is to be expected. We do not know whether ions or water do in fact aid in determining identity, but it is certainly conceivable, and it is interesting to consider the possibility that fluctuations in the availability of these components may differentially affect kinase stability, activity or substrate specificity.

**The catalytic identity of the p38 isoforms**

The kinase domains of the p38α and β isoforms are 75% identical, with most variability found in the large lobe (~residues 190-310 in **Figure 1.2**). Since this region is generally thought to be involved in substrate binding for all kinases, this would suggest that a distinguishing characteristic between p38α and β would be substrate specificity. Instead as we show in **Chapter 4**, the substrate profiles for these two isoforms overlap completely. Interestingly, p38γ, which has a kinase domain ~65% identical to that of p38α or β, displays a quite different substrate profile (**Appendix Figure A.1**). p38γ autophosphorylates to the same degree as either p38α or β, and shows a high affinity for a few substrates, but overall is not as promiscuous as the other two isoforms. A drop in sequence identity of 10% corresponds with a severely limited ability for p38γ's to phosphorylate several targets, likely restricting the potential roles the p38γ kinase domain could have. From this it is not hard to see why p38γ is unable to compensate for the loss of α in differentiating myoblasts and in other cell types[81, 82, 86]. In contrast, our results suggest that the kinase domain of p38β may possess the same phosphorylation functionality as α and have the inherent potential to compensate for its loss. This compensation does not occur. In myoblasts these two isoforms localize differently during differentiation, and differences in localization may explain why endogenous p38β cannot compensate for the loss of α in other cell types. We do not know what causes the difference in observed localization. Both p38α and β have an additional ~25 amino acids at their N-terminus outside of the kinase domain and ~60-65 additional amino acids at their C-terminus. These regions are conserved in known orthologous sequences but vary between p38α and β. There are no canonical signal motifs in these areas, but they would

be suitable regions for such motifs, or these regions may contain important protein-protein interaction sequences that indirectly affect localization. It would be interesting to determine if the kinase domain of p38β plus the N- and C-terminal sequences of p38α has a different localization pattern in myoblasts than wild type p38β, and if so whether it can now compensate for α.

**Non-catalytic p38 functions**

In the absence of any clear differences in substrate specificity, a difference in localization is at least one characteristic that would distinguish p38α and β *in vivo*. An alternative explanation could be that these kinases have distinct non-catalytic roles. Several papers have reported roles for p38α that are independent of its kinase activity. p38 has a well-known role mediating the stress response in many systems[83], and these may be partly due to kinase-independent mechanisms. Work from the lab of Steven Pelech has shown that p38α can bind to CK2 and ERK1/2 in response to stress stimuli. When bound to CK2, p38α activates this kinase through what appears to be an allosteric mechanism[203]. By comparison, p38α binds to and inhibits ERK1/2 activation, possibly by blocking access to a phosphorylation site for the upstream ERK-activating kinase MEK1[204]. This inhibitory interaction with ERK1/2 is worth noting as ERK1/2 activity suppresses myoblast differentiation[205]. If this were a function specific to p38α it would suggest another distinguishing characteristic between it and the other p38 isoforms. Similar to p38α's role in blocking ERK activation, p38α and p38β can bind to the dual specificity tyrosine-phosphorylation-regulated kinase 1B (DYRK1B) and prevent it from associating with its upstream activating kinase MKK3[206]. By preventing this association, p38α and β block the activation of DYRK1B. This role is specific to these two isoforms as p38γ and δ

cannot perform the same function. Finally a role for p38α during the cell cycle has been described that does not require its kinase activity[207]. The mechanism is completely unknown, but together with the previous papers described, it suggests that the p38 isoforms could be required for several processes in ways that do not involve phosphorylation of downstream targets. There are no non-catalytic roles that are known to be specific for p38α versus β, but certainly the potential is there, and could be another distinguishing feature between these two kinases (a summary of p38 isoform characteristic is shown in **Figure 5.1**).

**The catalytic identity of the protein kinase superfamily**

The importance of phosphorylation and its consequences for protein function cannot be questioned, but what is uncertain is how this process is directed. It is natural to assume that two different kinases will phosphorylate different substrates, or at least have substrate profiles that do not completely overlap. Substrate specificity would then be a major identifying feature for each kinase. As we showed in **Chapter 4**, this may not be the case with p38α and β, and there is evidence from other kinases to suggest that substrate specificity may not be a major determinate of uniqueness/identity, at least for kinases with a certain degree of sequence similarity. As mentioned in the **Discussion** of **Chapter 4**, the kinase domain of PYK2 can compensate for that of FAK provided it has an appropriate localization signal[176]. Another example of a kinase domain that appears to lack uniqueness is the muscle-specific receptor tyrosine kinase (MuSK). MuSK plays a critical role in acetylcholine receptor clustering in muscle and requires kinase activity to do so[208]. Its kinase domain does not appear to be a defining feature however. It can be

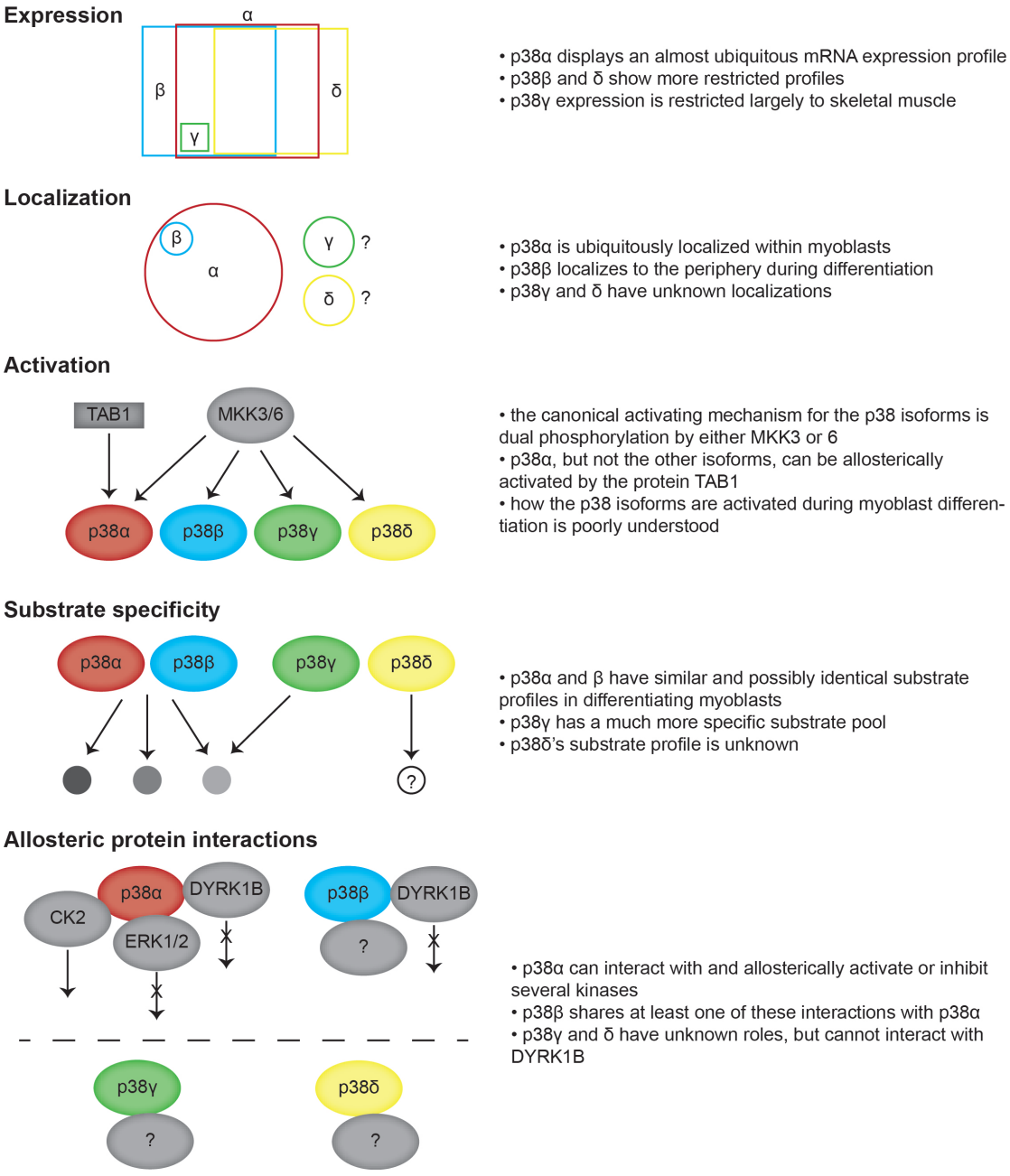



**Figure 5.1**. Properties of the p38 isoforms with reference to myoblast differentiation.

replaced with the kinase domain from either the platelet-derived growth factor β receptor (PDGFβ) or that from the high-affinity nerve growth factor receptor (TrkA), and acetylcholine receptor clustering occurs normally[209-211]. It seems that the juxtamembrane region of MuSK and certain phosphorylation sites within its kinase domain that are common to other tyrosine kinase receptors are the critical features required for MuSK or MuSK chimeric proteins to cluster acetylcholine receptors. This is very surprising as the kinase domain of MuSK is only 51% identical to that of TrkA, and only 37% identical to that of PDGFβ, differences that might otherwise suggest the existence of MuSK specific substrates. From differences in substrate specificity between p38α and γ, we can see how 65% sequence identity can result in quite drastic differences in the proteins phosphorylated. On the other hand, the 37% difference between MuSK and PDGFβ does not appear to result in the same degree of difference (or any degree). Certainly substrate specificity is a feature that will distinguish many kinases, particularly those distantly related, but to what degree it distinguishes closely related kinases, how much variability is required to create such a difference, and where does that variability need to be in the kinase domain, are questions that need further examination.

Assuming that what distinguishes two kinases is substrate specificity is obviously quite naïve, although it is a common assumption amongst those studying cell signaling. For many kinases it may be other factors that distinguish them, such as localization in the case of p38α and β. When considering what distinguishes two closely related proteins, kinase or otherwise, the entirety of their potential functions needs to be considered: expression, localization, method of activation, substrate specificity and protein-protein interactions. The p38 isoforms are relatively simple constructs that contain a kinase

domain plus short N- and C-terminal sequences. Much larger, multi-domain protein kinases exist, and the added complexity that these domains confer may contextualize the activity of a kinase domain to a very restricted localization, complex of proteins, cellular process or time frame. Additional domains and sequence motifs may take two catalytically identical kinase domains and change them into two very distinct and non-redundant proteins. The signaling complexity that exists within phosphorylation pathways likely owes much to these accessory domains and motifs.

**Implications for targeting protein kinases in disease**

The importance of protein kinases for cell signaling, and the potential for adverse deviations in cell behavior as a consequence of misregulated kinase activity, has led to the emergence of protein kinases as major drug targets[212]. While only a few inhibitors are approved for drug treatment, nearly 200 protein kinases have been implicated in disease and efforts are underway to design inhibitors for many of these, or test existing inhibitors in clinical trial[a]. The fact that protein kinases are so similar at the structural level creates a major challenge for designing specific inhibitors with limited off-target effects. Screens of different inhibitors against panels of kinases has revealed how many popular drugs used in research for years, as well as drugs clinically approved for disease therapy, inhibit more than just their intended targets[98, 213-215]. Many of these off-target effects could not have been predicted, as common kinases targeted by an inhibitor are often very distantly related. SP600125, designed as an inhibitor of JNK[216], and used by many groups to study the functions of this kinase, inhibited 13 of 28 kinases tested with equal or greater potency than JNK[214]. Sunitinib (also known as SU11248) is an FDA approved drug for

[a]A detailed list of kinases implicated in disease can be found at http://www.cellsignal.com/reference/kinase_disease.html

the treatment of renal cell carcinoma and gastrointestinal stromal tumors. It was designed to inhibit members of the split kinase family of tyrosine kinase receptors that include the receptors for VEGF and PDGF[217], but this compound was shown by Fabien et al.[98] to bind to at least 73 kinases in total (out of 119 tested), giving the potential for a vast array of adverse effects.

In addition to the inherent difficulties in designing a specific inhibitor due to high similarity between kinases, inhibitor design may have to take into consideration the need to preserve critical non-catalytic functions. As described in both the **General Introduction** and above in the **General Discussion**, protein kinases, including the kinase domain itself, can fulfill functions that are unrelated to catalysis. These other functions may require the kinase domain be in a particular conformation, which is important to consider as kinase inhibitors will often lock a kinase into a certain conformation (activated, inactivated, partially activated). To correctly study the catalytic function of a kinase *in vivo* would require any of its other functions to remain intact, therefore it may need to maintain conformational flexibility. The same may or may not be a necessary consideration when targeting a kinase for disease therapy. If it is only the consequences of overactive kinase activity that trigger disease and not other non-catalytic functions, a correctly designed inhibitor that can block one function and not the other may be necessary to prevent unwanted side effects. The design of such an inhibitor will not be easy. In **Chapter 2** we suggested that identifying kinases substrates and understanding substrate binding in detail might offer a solution to the challenges of inhibitor design. If a kinase has a specific substrate, and an inhibitor can be designed to mimic that substrate, then kinase-specific inhibitors could potentially be created. However, as we have shown

in **Chapter 4** and discussed above, identifying specific substrates, or more correctly specific phosphorylation sites, that are completely exclusive to one kinase may rarely be possible. A specific inhibitor that mimics a substrate may also have to be designed to target a specific tissue or cellular localization in order to inhibit its target without additional unwanted effects. Currently that is an impossible task. Whether it will ever be possible, or whether inhibitor design will always be plagued by the complications of similarity within the protein kinase superfamily, raises serious concerns about the utility of inhibitors for drawing conclusions about the details of cell biology, and for their application to disease therapy.

## REFERENCES


1. Adams, J.A. Kinetic and catalytic mechanisms of protein kinases. *Chem Rev* **101**, 2271-2290 (2001).
2. Besant, P.G., Attwood, P.V. & Piggott, M.J. Focus on phosphoarginine and phospholysine. *Curr Protein Pept Sci* **10**, 536-550 (2009).
3. Parkinson, J.S. Signal transduction schemes of bacteria. *Cell* **73**, 857-871 (1993).
4. Wagner, P.D. & Vu, N.D. Histidine to aspartate phosphotransferase activity of nm23 proteins: phosphorylation of aldolase C on Asp-319. *Biochem J* **346 Pt 3**, 623-630 (2000).
5. Wong, C., Faiola, B., Wu, W. & Kennelly, P.J. Phosphohistidine and phospholysine phosphatase activities in the rat: potential protein-lysine and protein-histidine phosphatases? *Biochem J* **296 ( Pt 2)**, 293-296 (1993).
6. Levy-Favatier, F., Delpech, M. & Kruh, J. Characterization of an arginine-specific protein kinase tightly bound to rat liver DNA. *Eur J Biochem* **166**, 617-621 (1987).
7. Smith, D.L. et al. Characterization of protein kinases forming acid-labile histone phosphates in Walker-256 carcinosarcoma cell nuclei. *Biochemistry* **13**, 3780-3785 (1974).
8. Ku, N.O. & Omary, M.B. A disease- and phosphorylation-related nonmechanical function for keratin 8. *J Cell Biol* **174**, 115-125 (2006).
9. Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T. & Manning, G. The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci U S A* **101**, 11707-11712 (2004).
10. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912-1934 (2002).
11. Scheeff, E.D. & Bourne, P.E. Structural evolution of the protein kinase-like superfamily. *PLoS Comput Biol* **1**, e49 (2005).
12. Kannan, N. & Neuwald, A.F. Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J Mol Biol* **351**, 956-972 (2005).
13. Boudeau, J., Miranda-Saavedra, D., Barton, G.J. & Alessi, D.R. Emerging roles of pseudokinases. *Trends Cell Biol* **16**, 443-452 (2006).
14. Mukherjee, K. et al. CASK Functions as a Mg2+-independent neurexin kinase. *Cell* **133**, 328-339 (2008).
15. Abe, Y. et al. Cloning and characterization of a p53-related protein kinase expressed in interleukin-2-activated cytotoxic T-cells, epithelial tumor cell lines, and the testes. *J Biol Chem* **276**, 44003-44011 (2001).
16. Xu, B. et al. WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II. *J Biol Chem* **275**, 16795-16801 (2000).
17. Zeqiraj, E. & van Aalten, D.M. Pseudokinases-remnants of evolution or key allosteric regulators? *Curr Opin Struct Biol* **20**, 772-781 (2010).

18. Fukuda, K., Knight, J., Piszczek, G., Kothary, R. & Qin, J. Biochemical, proteomic, structural, and thermodynamic characterizations of ILK: cross-validation of the pseudokinase. *J Biol Chem* (2011).
19. Fukuda, K., Gupta, S., Chen, K., Wu, C. & Qin, J. The pseudoactive site of ILK is essential for its binding to alpha-Parvin and localization to focal adhesions. *Mol Cell* **36**, 819-830 (2009).
20. Lange, A. et al. Integrin-linked kinase is an adaptor with essential functions during mouse development. *Nature* **461**, 1002-1006 (2009).
21. Wickstrom, S.A., Lange, A., Montanez, E. & Fassler, R. The ILK/PINCH/parvin complex: the kinase is dead, long live the pseudokinase! *EMBO J* **29**, 281-291 (2010).
22. Legate, K.R., Montanez, E., Kudlacek, O. & Fassler, R. ILK, PINCH and parvin: the tIPP of integrin signalling. *Nat Rev Mol Cell Biol* **7**, 20-31 (2006).
23. Grashoff, C., Thievessen, I., Lorenz, K., Ussar, S. & Fassler, R. Integrin-linked kinase: integrin's mysterious partner. *Curr Opin Cell Biol* **16**, 565-571 (2004).
24. Zeqiraj, E., Filippi, B.M., Deak, M., Alessi, D.R. & van Aalten, D.M. Structure of the LKB1-STRAD-MO25 complex reveals an allosteric mechanism of kinase activation. *Science* **326**, 1707-1711 (2009).
25. Zeqiraj, E. et al. ATP and MO25alpha regulate the conformational state of the STRADalpha pseudokinase and activation of the LKB1 tumour suppressor. *PLoS Biol* **7**, e1000126 (2009).
26. Lu, C. et al. Participation of Rip2 in lipopolysaccharide signaling is independent of its kinase activity. *J Biol Chem* **280**, 16278-16283 (2005).
27. Thome, M. et al. Identification of CARDIAK, a RIP-like kinase that associates with caspase-1. *Curr Biol* **8**, 885-888 (1998).
28. McCarthy, J.V., Ni, J. & Dixit, V.M. RIP2 is a novel NF-kappaB-activating and cell death-inducing kinase. *J Biol Chem* **273**, 16968-16975 (1998).
29. Chong, Y.P., Mulhern, T.D. & Cheng, H.C. C-terminal Src kinase (CSK) and CSK-homologous kinase (CHK)--endogenous negative regulators of Src-family protein kinases. *Growth Factors* **23**, 233-244 (2005).
30. Chong, Y.P. et al. A novel non-catalytic mechanism employed by the C-terminal Src-homologous kinase to inhibit Src-family kinase activity. *J Biol Chem* **279**, 20752-20766 (2004).
31. Cicero, S. & Herrup, K. Cyclin-dependent kinase 5 is essential for neuronal cell cycle arrest and differentiation. *J Neurosci* **25**, 9658-9668 (2005).
32. Zhang, J. & Herrup, K. Cdk5 and the non-catalytic arrest of the neuronal cell cycle. *Cell Cycle* **7**, 3487-3490 (2008).
33. Gnad, F. et al. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* **8**, R250 (2007).
34. Alonso, A. et al. Protein tyrosine phosphatases in the human genome. *Cell* **117**, 699-711 (2004).
35. Shi, Y. Serine/threonine phosphatases: mechanism through structure. *Cell* **139**, 468-484 (2009).
36. Chang, L. & Karin, M. Mammalian MAP kinase signalling cascades. *Nature* **410**, 37-40 (2001).

37. Gallego, M. & Virshup, D.M. Protein serine/threonine phosphatases: life, death, and sleeping. *Curr Opin Cell Biol* **17**, 197-202 (2005).
38. Moorhead, G.B., Trinkle-Mulcahy, L. & Ulke-Lemee, A. Emerging roles of nuclear protein phosphatases. *Nat Rev Mol Cell Biol* **8**, 234-244 (2007).
39. Kilpinen, S., Ojala, K. & Kallioniemi, O. Analysis of kinase gene expression patterns across 5681 human tissue samples reveals functional genomic taxonomy of the kinome. *PLoS One* **5**, e15068 (2010).
40. Jin, J. et al. Eukaryotic protein domains as functional units of cellular evolution. *Sci Signal* **2**, ra76 (2009).
41. Pawson, T., Gish, G.D. & Nash, P. SH2 domains, interaction modules and cellular wiring. *Trends Cell Biol* **11**, 504-511 (2001).
42. Johnson, L.N., Lowe, E.D., Noble, M.E. & Owen, D.J. The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. *FEBS Lett* **430**, 1-11 (1998).
43. Kemp, B.E., Parker, M.W., Hu, S., Tiganis, T. & House, C. Substrate and pseudosubstrate interactions with protein kinases: determinants of specificity. *Trends Biochem Sci* **19**, 440-444 (1994).
44. Knighton, D.R. et al. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 414-420 (1991).
45. Knighton, D.R. et al. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 407-414 (1991).
46. Biondi, R.M. & Nebreda, A.R. Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions. *Biochem J* **372**, 1-13 (2003).
47. Manning, B.D. & Cantley, L.C. AKT/PKB signaling: navigating downstream. *Cell* **129**, 1261-1274 (2007).
48. Shabb, J.B. Physiological substrates of cAMP-dependent protein kinase. *Chem Rev* **101**, 2381-2411 (2001).
49. Ptacek, J. et al. Global analysis of protein phosphorylation in yeast. *Nature* **438**, 679-684 (2005).
50. Kolch, W. Coordinating ERK/MAPK signalling through scaffolds and inhibitors. *Nat Rev Mol Cell Biol* **6**, 827-837 (2005).
51. Harrison, P.M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res* **30**, 1083-1090 (2002).
52. Knebel, A., Morrice, N. & Cohen, P. A novel method to identify protein kinase substrates: eEF2 kinase is phosphorylated and inhibited by SAPK4/p38delta. *EMBO J* **20**, 4360-4369 (2001).
53. Jaleel, M. et al. LRRK2 phosphorylates moesin at threonine-558: characterization of how Parkinson's disease mutants affect kinase activity. *Biochem J* **405**, 307-317 (2007).
54. Peng, C. et al. Pim kinase substrate identification and specificity. *J Biochem* **141**, 353-362 (2007).
55. Cuomo, M.E. et al. Regulation of microfilament organization by Kaposi sarcoma-associated herpes virus-cyclin.CDK6 phosphorylation of caldesmon. *J Biol Chem* **280**, 35844-35858 (2005).

56. Auld, G.C., Campbell, D.G., Morrice, N. & Cohen, P. Identification of calcium-regulated heat-stable protein of 24 kDa (CRHSP24) as a physiological substrate for PKB and RSK using KESTREL. *Biochem J* **389**, 775-783 (2005).
57. Cartlidge, R.A. et al. The tRNA methylase METTL1 is phosphorylated and inactivated by PKB and RSK in vitro and in cells. *EMBO J* **24**, 1696-1705 (2005).
58. Eyers, C.E. et al. The phosphorylation of CapZ-interacting protein (CapZIP) by stress-activated protein kinases triggers its dissociation from CapZ. *Biochem J* **389**, 127-135 (2005).
59. Cole, A.R. et al. GSK-3 phosphorylation of the Alzheimer epitope within collapsin response mediator proteins regulates axon elongation in primary neurons. *J Biol Chem* **279**, 50176-50180 (2004).
60. Murray, J.T. et al. Exploitation of KESTREL to identify NDRG family members as physiological substrates for SGK1 and GSK3. *Biochem J* **384**, 477-488 (2004).
61. Murray, J.T., Campbell, D.G., Peggie, M., Mora, A. & Cohen, P. Identification of filamin C as a new physiological substrate of PKBalpha using KESTREL. *Biochem J* **384**, 489-494 (2004).
62. McNeill, H., Knebel, A., Arthur, J.S., Cuenda, A. & Cohen, P. A novel UBA and UBX domain protein that binds polyubiquitin and VCP and is a substrate for SAPKs. *Biochem J* **384**, 391-400 (2004).
63. Rousseau, S. et al. Inhibition of SAPK2a/p38 prevents hnRNP A0 phosphorylation by MAPKAP-K2 and its interaction with cytokine mRNAs. *EMBO J* **21**, 6505-6514 (2002).
64. Cohen, P. & Knebel, A. KESTREL: a powerful method for identifying the physiological substrates of protein kinases. *Biochem J* **393**, 1-6 (2006).
65. Troiani, S. et al. Searching for biomarkers of Aurora-A kinase activity: identification of in vitro substrates through a modified KESTREL approach. *J Proteome Res* **4**, 1296-1303 (2005).
66. Jiang, W. et al. PRC1: a human mitotic spindle-associated CDK substrate protein required for cytokinesis. *Mol Cell* **2**, 877-885 (1998).
67. Fukunaga, R. & Hunter, T. MNK1, a new MAP kinase-activated protein kinase, isolated by a novel expression screening method for identifying protein kinase substrates. *EMBO J* **16**, 1921-1933 (1997).
68. Shah, K. & Shokat, K.M. A chemical genetic screen for direct v-Src substrates reveals ordered assembly of a retrograde signaling pathway. *Chem Biol* **9**, 35-47 (2002).
69. Shah, K., Liu, Y., Deirmengian, C. & Shokat, K.M. Engineering unnatural nucleotide specificity for Rous sarcoma virus tyrosine kinase to uniquely label its direct substrates. *Proc Natl Acad Sci U S A* **94**, 3565-3570 (1997).
70. Dephoure, N., Howson, R.W., Blethrow, J.D., Shokat, K.M. & O'Shea, E.K. Combining chemical genetics and proteomics to identify protein kinase substrates. *Proc Natl Acad Sci U S A* **102**, 17940-17945 (2005).
71. Ubersax, J.A. et al. Targets of the cyclin-dependent kinase Cdk1. *Nature* **425**, 859-864 (2003).
72. Blethrow, J.D., Glavy, J.S., Morgan, D.O. & Shokat, K.M. Covalent capture of kinase-specific phosphopeptides reveals Cdk1-cyclin B substrates. *Proc Natl Acad Sci U S A* **105**, 1442-1447 (2008).

73. Jiang, Y. et al. Characterization of the structure and function of a new mitogen-activated protein kinase (p38beta). *J Biol Chem* **271**, 17920-17926 (1996).
74. Wang, X.S. et al. Molecular cloning and characterization of a novel p38 mitogen-activated protein kinase. *J Biol Chem* **272**, 23668-23674 (1997).
75. Beardmore, V.A. et al. Generation and characterization of p38beta (MAPK11) gene-targeted mice. *Mol Cell Biol* **25**, 10454-10464 (2005).
76. Sabio, G. et al. p38gamma regulates the localisation of SAP97 in the cytoskeleton by modulating its interaction with GKAP. *EMBO J* **24**, 1134-1145 (2005).
77. Greenblatt, M.B. et al. The p38 MAPK pathway is essential for skeletogenesis and bone homeostasis in mice. *J Clin Invest* **120**, 2457-2473 (2010).
78. Gillespie, M.A. et al. p38-{gamma}-dependent gene silencing restricts entry into the myogenic differentiation program. *J Cell Biol* **187**, 991-1005 (2009).
79. Efimova, T. p38delta mitogen-activated protein kinase regulates skin homeostasis and tumorigenesis. *Cell Cycle* **9**, 498-405 (2010).
80. Perdiguero, E. et al. Genetic analysis of p38 MAP kinases in myogenesis: fundamental role of p38alpha in abrogating myoblast proliferation. *EMBO J* **26**, 1245-1256 (2007).
81. Adams, R.H. et al. Essential role of p38alpha MAP kinase in placental but not embryonic cardiovascular development. *Mol Cell* **6**, 109-116 (2000).
82. Tamura, K. et al. Requirement for p38alpha in erythropoietin expression: a role for stress kinases in erythropoiesis. *Cell* **102**, 221-231 (2000).
83. Cuenda, A. & Rousseau, S. p38 MAP-kinases pathway regulation, function and role in human diseases. *Biochim Biophys Acta* **1773**, 1358-1375 (2007).
84. Zarubin, T. & Han, J. Activation and signaling of the p38 MAP kinase pathway. *Cell Res* **15**, 11-18 (2005).
85. Li, Z., Jiang, Y., Ulevitch, R.J. & Han, J. The primary structure of p38 gamma: a new member of p38 group of MAP kinases. *Biochem Biophys Res Commun* **228**, 334-340 (1996).
86. Ruiz-Bonilla, V. et al. Efficient adult skeletal muscle regeneration in mice deficient in p38beta, p38gamma and p38delta MAP kinases. *Cell Cycle* **7**, 2208-2214 (2008).
87. Perdiguero, E., Ruiz-Bonilla, V., Serrano, A.L. & Munoz-Canoves, P. Genetic deficiency of p38alpha reveals its critical role in myoblast cell cycle exit: the p38alpha-JNK connection. *Cell Cycle* **6**, 1298-1303 (2007).
88. Lluis, F., Perdiguero, E., Nebreda, A.R. & Munoz-Canoves, P. Regulation of skeletal muscle gene expression by p38 MAP kinases. *Trends Cell Biol* **16**, 36-44 (2006).
89. Li, Y., Jiang, B., Ensign, W.Y., Vogt, P.K. & Han, J. Myogenic differentiation requires signalling through both phosphatidylinositol 3-kinase and p38 MAP kinase. *Cell Signal* **12**, 751-757 (2000).
90. LaRonde-LeBlanc, N. & Wlodawer, A. A family portrait of the RIO kinases. *J Biol Chem* **280**, 37297-37300 (2005).
91. Yamaguchi, H., Matsushita, M., Nairn, A.C. & Kuriyan, J. Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity. *Mol Cell* **7**, 1047-1057 (2001).

92. Dedhar, S., Williams, B. & Hannigan, G. Integrin-linked kinase (ILK): a regulator of integrin and growth-factor signalling. *Trends Cell Biol* **9**, 319-323 (1999).
93. Slamon, D.J. et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177-182 (1987).
94. Sawyers, C.L. Chronic myeloid leukemia. *N Engl J Med* **340**, 1330-1340 (1999).
95. Hommes, D. et al. Inhibition of stress-activated MAP kinases induces clinical improvement in moderate to severe Crohn's disease. *Gastroenterology* **122**, 7-14 (2002).
96. Ono-Saito, N., Niki, I. & Hidaka, H. H-series protein kinase inhibitors and potential clinical applications. *Pharmacol Ther* **82**, 123-131 (1999).
97. Sebolt-Leopold, J.S. & English, J.M. Mechanisms of drug inhibition of signalling molecules. *Nature* **441**, 457-462 (2006).
98. Fabian, M.A. et al. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol* **23**, 329-336 (2005).
99. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93 (2004).
100. Lougheed, J.C., Chen, R.H., Mak, P. & Stout, T.J. Crystal structures of the phosphorylated and unphosphorylated kinase domains of the Cdc42-associated tyrosine kinase ACK1. *J Biol Chem* **279**, 44039-44045 (2004).
101. Yang, J. et al. Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP-PNP. *Nat Struct Biol* **9**, 940-944 (2002).
102. Russo, A.A., Jeffrey, P.D. & Pavletich, N.P. Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nat Struct Biol* **3**, 696-700 (1996).
103. Xu, R.M., Carmel, G., Sweet, R.M., Kuret, J. & Cheng, X. Crystal structure of casein kinase-1, a phosphate-directed protein kinase. *EMBO J* **14**, 1015-1023 (1995).
104. Yde, C.W., Ermakova, I., Issinger, O.G. & Niefind, K. Inclining the purine base binding plane in protein kinase CK2 by exchanging the flanking side-chains generates a preference for ATP as a cosubstrate. *J Mol Biol* **347**, 399-414 (2005).
105. Tereshko, V., Teplova, M., Brunzelle, J., Watterson, D.M. & Egli, M. Crystal structures of the catalytic domain of human protein kinase associated with apoptosis and tumor suppression. *Nat Struct Biol* **8**, 899-907 (2001).
106. Hubbard, S.R. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J* **16**, 5572-5581 (1997).
107. Bellon, S., Fitzgibbon, M.J., Fox, T., Hsiao, H.M. & Wilson, K.P. The structure of phosphorylated p38gamma is monomeric and reveals a conserved activation-loop conformation. *Structure* **7**, 1057-1065 (1999).
108. Owen, D.J., Noble, M.E., Garman, E.F., Papageorgiou, A.C. & Johnson, L.N. Two structures of the catalytic domain of phosphorylase kinase: an active protein kinase complexed with substrate analogue and product. *Structure* **3**, 467-482 (1995).
109. Qian, K.C. et al. Structural basis of constitutive activity and a unique nucleotide binding mode of human Pim-1 kinase. *J Biol Chem* **280**, 6130-6137 (2005).
110. Zheng, J. et al. 2.2 A refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor. *Acta Crystallogr D Biol Crystallogr* **49**, 362-365 (1993).

111. Young, T.A., Delagoutte, B., Endrizzi, J.A., Falick, A.M. & Alber, T. Structure of Mycobacterium tuberculosis PknB supports a universal activation mechanism for Ser/Thr protein kinases. *Nat Struct Biol* **10**, 168-174 (2003).
112. LaRonde-LeBlanc, N., Guszczynski, T., Copeland, T. & Wlodawer, A. Autophosphorylation of Archaeoglobus fulgidus Rio2 and crystal structures of its nucleotide-metal ion complexes. *FEBS J* **272**, 2800-2810 (2005).
113. Nolen, B. et al. Nucleotide-induced conformational changes in the Saccharomyces cerevisiae SR protein kinase, Sky1p, revealed by X-ray crystallography. *Biochemistry* **42**, 9575-9585 (2003).
114. Zhou, T. et al. Crystal structure of the TAO2 kinase domain: activation and specificity of a Ste20p MAP3K. *Structure* **12**, 1891-1900 (2004).
115. Nolen, B. et al. The structure of Sky1p reveals a novel mechanism for constitutive activity. *Nat Struct Biol* **8**, 176-183 (2001).
116. LaRonde-LeBlanc, N. & Wlodawer, A. Crystal structure of A. fulgidus Rio2 defines a new family of serine protein kinases. *Structure* **12**, 1585-1594 (2004).
117. Todd, A.E., Orengo, C.A. & Thornton, J.M. Plasticity of enzyme active sites. *Trends Biochem Sci* **27**, 419-426 (2002).
118. Grishin, N.V. Fold change in evolution of protein structures. *J Struct Biol* **134**, 167-185 (2001).
119. Iyer, G.H., Garrod, S., Woods, V.L., Jr. & Taylor, S.S. Catalytic independent functions of a protein kinase as revealed by a kinase-dead mutant: study of the Lys72His mutant of cAMP-dependent kinase. *J Mol Biol* **351**, 1110-1122 (2005).
120. Robinson, M.J. et al. Mutation of position 52 in ERK2 creates a nonproductive binding mode for adenosine 5'-triphosphate. *Biochemistry* **35**, 5641-5646 (1996).
121. Carrera, A.C., Alexandrov, K. & Roberts, T.M. The conserved lysine of the catalytic domain of protein kinases is actively involved in the phosphotransfer reaction and not required for anchoring ATP. *Proc Natl Acad Sci U S A* **90**, 442-446 (1993).
122. Drennan, D. & Ryazanov, A.G. Alpha-kinases: analysis of the family and comparison with conventional protein kinases. *Prog Biophys Mol Biol* **85**, 1-32 (2004).
123. Min, X., Lee, B.H., Cobb, M.H. & Goldsmith, E.J. Crystal structure of the kinase domain of WNK1, a kinase that causes a hereditary form of hypertension. *Structure* **12**, 1303-1311 (2004).
124. Kim, C., Xuong, N.H. & Taylor, S.S. Crystal structure of a complex between the catalytic and regulatory (RIalpha) subunits of PKA. *Science* **307**, 690-696 (2005).
125. Felberg, J. et al. Subdomain X of the kinase domain of Lck binds CD45 and facilitates dephosphorylation. *J Biol Chem* **279**, 3455-3462 (2004).
126. Huang, J., Tu, Z. & Lee, F.S. Mutations in protein kinase subdomain X differentially affect MEKK2 and MEKK1 activity. *Biochem Biophys Res Commun* **303**, 532-540 (2003).
127. Tu, Z., Mooney, S.M. & Lee, F.S. A subdomain of MEKK1 that is critical for binding to MKK4. *Cell Signal* **15**, 65-77 (2003).
128. Lei, M. et al. Structure of PAK1 in an autoinhibited conformation reveals a multistage activation switch. *Cell* **102**, 387-397 (2000).

129. Das, R. et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69 Suppl 8**, 118-128 (2007).
130. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680 (1994).
131. Biondi, R.M. et al. High resolution crystal structure of the human PDK1 catalytic domain defines the regulatory phosphopeptide docking site. *EMBO J* **21**, 4219-4228 (2002).
132. Hawkins, J., Zheng, S., Frantz, B. & LoGrasso, P. p38 map kinase substrate specificity differs greatly for protein and peptide substrates. *Arch Biochem Biophys* **382**, 310-313 (2000).
133. Blencke, S. et al. Characterization of a conserved structural determinant controlling protein kinase sensitivity to selective inhibitors. *Chem Biol* **11**, 691-701 (2004).
134. Liu, Y. et al. Structural basis for selective inhibition of Src family kinases by PP1. *Chem Biol* **6**, 671-678 (1999).
135. Bogoyevitch, M.A., Barr, R.K. & Ketterman, A.J. Peptide inhibitors of protein kinases-discovery, characterisation and use. *Biochim Biophys Acta* **1754**, 79-99 (2005).
136. Perea, S.E. et al. Antitumor effect of a novel proapoptotic peptide that impairs the phosphorylation by the protein kinase 2 (casein kinase 2). *Cancer Res* **64**, 7127-7129 (2004).
137. Shah, K. & Shokat, K.M. A chemical genetic approach for the identification of direct substrates of protein kinases. *Methods Mol Biol* **233**, 253-271 (2003).
138. Dryden, I.L. & Mardia, K.V. Statistical Shape Analysis. (Wiley, Chichester; 1998).
139. Goodall, C. Procrustes methods in the statistical analysis of shape. *J R Statist Soc* **53**, 285-339 (1991).
140. Claude, J., Pritchard, P., Tong, H., Paradis, E. & Auffray, J.C. Ecological correlates and evolutionary divergence in the skull of turtles: a geometric morphometric assessment. *Syst Biol* **53**, 933-948 (2004).
141. Kent, J.T. in The art of statistical science. (ed. K.V. Mardia) 115-127 (Wiley, Chichester; 1992).
142. Gutteridge, A. & Thornton, J.M. Understanding nature's catalytic toolkit. *Trends Biochem Sci* **30**, 622-629 (2005).
143. Shaltiel, S., Cox, S. & Taylor, S.S. Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase A. *Proc Natl Acad Sci U S A* **95**, 484-491 (1998).
144. Levy, Y. & Onuchic, J.N. Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct* **35**, 389-415 (2006).
145. Loris, R. et al. Conserved water molecules in a large family of microbial ribonucleases. *Proteins* **36**, 117-134 (1999).
146. Malin, R., Zielenkiewicz, P. & Saenger, W. Structurally conserved water molecules in ribonuclease T1. *J Biol Chem* **266**, 4848-4852 (1991).

147. Sreenivasan, U. & Axelsen, P.H. Buried water in homologous serine proteases. *Biochemistry* **31**, 12785-12791 (1992).
148. Bottoms, C.A., Smith, P.E. & Tanner, J.J. A structurally conserved water molecule in Rossmann dinucleotide-binding domains. *Protein Sci* **11**, 2125-2137 (2002).
149. Ogata, K. & Wodak, S.J. Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. *Protein Eng* **15**, 697-705 (2002).
150. Prasad, B.V. & Suguna, K. Role of water molecules in the structure and function of aspartic proteinases. *Acta Crystallogr D Biol Crystallogr* **58**, 250-259 (2002).
151. Rodriguez-Almazan, C. et al. Structural basis of human triosephosphate isomerase deficiency: mutation E104D is related to alterations of a conserved water network at the dimer interface. *J Biol Chem* **283**, 23254-23263 (2008).
152. Knight, J.D., Qian, B., Baker, D. & Kothary, R. Conservation, variability and the modeling of active protein kinases. *PLoS One* **2**, e982 (2007).
153. Hanks, S.K. & Hunter, T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J* **9**, 576-596 (1995).
154. Gibbs, C.S. & Zoller, M.J. Rational scanning mutagenesis of a protein kinase identifies functional regions involved in catalysis and substrate interactions. *J Biol Chem* **266**, 8923-8931 (1991).
155. Fraczkiewicz, R. & Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comput Chem* **19**, 319-333 (1998).
156. Dundas, J. et al. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* **34**, W116-118 (2006).
157. Case, D.A. et al. The Amber biomolecular simulation programs. *J Comput Chem* **26**, 1668-1688 (2005).
158. Meagher, K.L., Redman, L.T. & Carlson, H.A. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem* **24**, 1016-1025 (2003).
159. Jorgensen, W.L., Chandrasekhar, J. & Madura, J.D. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **79**, 926 (1983).
160. Ryckaert, J.P., Ciccotti, G. & Berendsen, H.J.C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comp Phys* **23**, 327-341 (1977).
161. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald. An Nlog(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089-10092 (1993).
162. Hamelberg, D. & McCammon, J.A. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J Am Chem Soc* **126**, 7683-7689 (2004).
163. Denisov, V.P., Venu, K., Peters, J., Hörlein, H.D. & Halle, B. Orientational disorder and entropy of water in protein cavities. *J Phys Chem B* **101**, 9380-9389 (1997).
164. Denisov, V.P., Peters, J., Horlein, H.D. & Halle, B. Using buried water molecules to explore the energy landscape of proteins. *Nat Struct Biol* **3**, 505-509 (1996).

165. Berwick, D.C. & Tavare, J.M. Identifying protein kinase substrates: hunting for the organ-grinder's monkeys. *Trends Biochem Sci* **29**, 227-232 (2004).
166. Johnson, S.A. & Hunter, T. Kinomics: methods for deciphering the kinome. *Nat Methods* **2**, 17-25 (2005).
167. Manning, B.D. & Cantley, L.C. Hitting the target: emerging technologies in the search for kinase substrates. *Sci STKE* **2002**, pe49 (2002).
168. Wang, H., Xu, Q., Xiao, F., Jiang, Y. & Wu, Z. Involvement of the p38 mitogen-activated protein kinase alpha, beta, and gamma isoforms in myogenic differentiation. *Mol Biol Cell* **19**, 1519-1528 (2008).
169. Hashimoto, E., Takio, K. & Krebs, E.G. Amino acid sequence at the ATP-binding site of cGMP-dependent protein kinase. *J Biol Chem* **257**, 727-733 (1982).
170. Kamps, M.P., Taylor, S.S. & Sefton, B.M. Direct evidence that oncogenic tyrosine kinases and cyclic AMP-dependent protein kinase have homologous ATP-binding sites. *Nature* **310**, 589-592 (1984).
171. Russo, M.W., Lukas, T.J., Cohen, S. & Staros, J.V. Identification of residues in the nucleotide binding site of the epidermal growth factor receptor/kinase. *J Biol Chem* **260**, 5205-5208 (1985).
172. Zoller, M.J., Nelson, N.C. & Taylor, S.S. Affinity labeling of cAMP-dependent protein kinase with p-fluorosulfonylbenzoyl adenosine. Covalent modification of lysine 71. *J Biol Chem* **256**, 10837-10842 (1981).
173. Andres, V. & Walsh, K. Myogenin expression, cell cycle withdrawal, and phenotypic differentiation are temporally separable events that precede cell fusion upon myogenesis. *J Cell Biol* **132**, 657-666 (1996).
174. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190 (2004).
175. Schwartz, D. & Gygi, S.P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* **23**, 1391-1398 (2005).
176. Klingbeil, C.K. et al. Targeting Pyk2 to beta 1-integrin-containing focal contacts rescues fibronectin-stimulated signaling and haptotactic motility defects of focal adhesion kinase-null cells. *J Cell Biol* **152**, 97-110 (2001).
177. Enslen, H., Raingeaud, J. & Davis, R.J. Selective activation of p38 mitogen-activated protein (MAP) kinase isoforms by the MAP kinase kinases MKK3 and MKK6. *J Biol Chem* **273**, 1741-1748 (1998).
178. Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A.J. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* **4**, 484-494 (2009).
179. Li, Q.R., Ning, Z.B., Tang, J.S., Nie, S. & Zeng, R. Effect of peptide-to-TiO2 beads ratio on phosphopeptide enrichment selectivity. *J Proteome Res* **8**, 5375-5381 (2009).
180. Wang, F., Dong, J., Jiang, X., Ye, M. & Zou, H. Capillary trap column with strong cation-exchange monolith for automated shotgun proteome analysis. *Anal Chem* **79**, 6599-6606 (2007).
181. Wang, F. et al. A fully automated system with online sample loading, isotope dimethyl labeling and multidimensional separation for high-throughput quantitative proteome analysis. *Anal Chem* **82**, 3007-3015 (2010).

182. Sens, K.L. et al. An invasive podosome-like structure promotes fusion pore formation during myoblast fusion. *J Cell Biol* **191**, 1013-1027 (2010).
183. Peckham, M. Engineering a multi-nucleated myotube, the role of the actin cytoskeleton. *J Microsc* **231**, 486-493 (2008).
184. Bizzarro, V. et al. Role of Annexin A1 in mouse myoblast cell differentiation. *J Cell Physiol* **224**, 757-765 (2010).
185. Barnoy, S., Maki, M. & Kosower, N.S. Overexpression of calpastatin inhibits L8 myoblast fusion. *Biochem Biophys Res Commun* **332**, 697-701 (2005).
186. Pho, M. et al. Cofilin is a marker of myofibroblast differentiation in cells from porcine aortic cardiac valves. *Am J Physiol Heart Circ Physiol* **294**, H1767-1778 (2008).
187. Dorsett, M. & Schedl, T. A role for dynein in the inhibition of germ cell proliferative fate. *Mol Cell Biol* **29**, 6128-6139 (2009).
188. Lopez-Alemany, R., Suelves, M., Diaz-Ramos, A., Vidal, B. & Munoz-Canoves, P. Alpha-enolase plasminogen receptor in myogenesis. *Front Biosci* **10**, 30-36 (2005).
189. Bello, N.F. et al. The E3 ubiquitin ligase specificity subunit ASB2beta is a novel regulator of muscle differentiation that targets filamin B to proteasomal degradation. *Cell Death Differ* **16**, 921-932 (2009).
190. van der Flier, A. et al. Different splice variants of filamin-B affect myogenesis, subcellular distribution, and determine binding to integrin [beta] subunits. *J Cell Biol* **156**, 361-376 (2002).
191. Kawaji, A., Ohnaka, Y., Osada, S., Nishizuka, M. & Imagawa, M. Gelsolin, an actin regulatory protein, is required for differentiation of mouse 3T3-L1 cells into adipocytes. *Biol Pharm Bull* **33**, 773-779 (2010).
192. Delehouzee, S. et al. GABP, HCF-1 and YY1 are involved in Rb gene expression during myogenesis. *Genes Cells* **10**, 717-731 (2005).
193. de Thonel, A. et al. HSP27 controls GATA-1 protein level during erythroid cell differentiation. *Blood* **116**, 85-96 (2010).
194. Shi, G.X., Jin, L. & Andres, D.A. Pituitary adenylate cyclase-activating polypeptide 38-mediated Rin activation requires Src and contributes to the regulation of HSP27 signaling during neuronal differentiation. *Mol Cell Biol* **28**, 4940-4951 (2008).
195. Riederer, B.M. Microtubule-associated protein 1B, a growth-associated and phosphorylated scaffold protein. *Brain Res Bull* **71**, 541-558 (2007).
196. Dulong, S. et al. Myristoylated alanine-rich C kinase substrate (MARCKS) is involved in myoblast fusion through its regulation by protein kinase Calpha and calpain proteolytic cleavage. *Biochem J* **382**, 1015-1023 (2004).
197. Kim, S.S. et al. Involvement of protein phosphatase-1-mediated MARCKS translocation in myogenic differentiation of embryonic muscle cells. *J Cell Sci* **115**, 2465-2473 (2002).
198. Pallari, H.M. et al. Nestin as a regulator of Cdk5 in differentiating myoblasts. *Mol Biol Cell* **22**, 1539-1549 (2011).
199. Minekawa, R. et al. Involvement of RelA-associated inhibitor in regulation of trophoblast differentiation via interaction with transcriptional factor specificity protein-1. *Endocrinology* **148**, 5803-5810 (2007).

200. Li, Z. et al. Defective smooth muscle development in qkI-deficient mice. *Dev Growth Differ* **45**, 449-462 (2003).
201. Hishida, T., Eguchi, T., Osada, S., Nishizuka, M. & Imagawa, M. A novel gene, fad49, plays a crucial role in the immediate early stage of adipocyte differentiation via involvement in mitotic clonal expansion. *FEBS J* **275**, 5576-5588 (2008).
202. Creppe, C. et al. Elongator controls the migration and differentiation of cortical neurons through acetylation of alpha-tubulin. *Cell* **136**, 551-564 (2009).
203. Sayed, M., Kim, S.O., Salh, B.S., Issinger, O.G. & Pelech, S.L. Stress-induced activation of protein kinase CK2 by direct interaction with p38 mitogen-activated protein kinase. *J Biol Chem* **275**, 16569-16573 (2000).
204. Zhang, H., Shi, X., Hampong, M., Blanis, L. & Pelech, S. Stress-induced inhibition of ERK1 and ERK2 by direct interaction with p38 MAP kinase. *J Biol Chem* **276**, 6905-6908 (2001).
205. Wu, Z. et al. p38 and extracellular signal-regulated kinases regulate the myogenic program at multiple steps. *Mol Cell Biol* **20**, 3951-3964 (2000).
206. Lim, S., Zou, Y. & Friedman, E. The transcriptional activator Mirk/Dyrk1B is sequestered by p38alpha/beta MAP kinase. *J Biol Chem* **277**, 49438-49445 (2002).
207. Fan, L. et al. A novel role of p38 alpha MAPK in mitotic progression independent of its kinase activity. *Cell Cycle* **4**, 1616-1624 (2005).
208. Sanes, J.R. & Lichtman, J.W. Development of the vertebrate neuromuscular junction. *Annu Rev Neurosci* **22**, 389-442 (1999).
209. Cheusova, T. et al. Casein kinase 2-dependent serine phosphorylation of MuSK regulates acetylcholine receptor aggregation at the neuromuscular junction. *Genes Dev* **20**, 1800-1816 (2006).
210. Herbst, R., Avetisova, E. & Burden, S.J. Restoration of synapse formation in Musk mutant mice expressing a Musk/Trk chimeric receptor. *Development* **129**, 5449-5460 (2002).
211. Herbst, R. & Burden, S.J. The juxtamembrane region of MuSK has a critical role in agrin-mediated signaling. *EMBO J* **19**, 67-77 (2000).
212. Cohen, P. Protein kinases--the major drug targets of the twenty-first century? *Nat Rev Drug Discov* **1**, 309-315 (2002).
213. Bain, J. et al. The selectivity of protein kinase inhibitors: a further update. *Biochem J* **408**, 297-315 (2007).
214. Bain, J., McLauchlan, H., Elliott, M. & Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem J* **371**, 199-204 (2003).
215. Davies, S.P., Reddy, H., Caivano, M. & Cohen, P. Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem J* **351**, 95-105 (2000).
216. Bennett, B.L. et al. SP600125, an anthrapyrazolone inhibitor of Jun N-terminal kinase. *Proc Natl Acad Sci U S A* **98**, 13681-13686 (2001).
217. Mendel, D.B. et al. In vivo antitumor activity of SU11248, a novel tyrosine kinase inhibitor targeting vascular endothelial growth factor and platelet-derived growth factor receptors: determination of a pharmacokinetic/pharmacodynamic relationship. *Clin Cancer Res* **9**, 327-337 (2003).

# APPENDIX

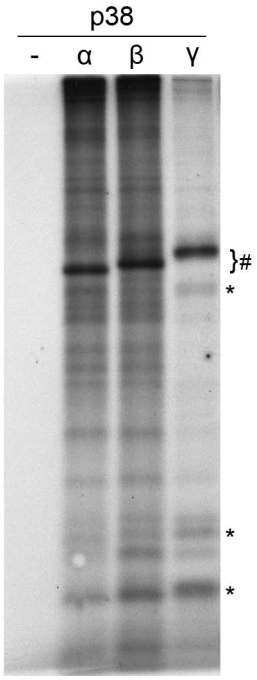



**Figure A.1**. Substrate profiles of p38α, β and γ. FSBA treated C2C12 cell lysate was incubated with a kinase assay buffer and p38α, β, γ or no kinase as a control. The prominent band in each of the p38α, β and γ lanes marked by the # sign represents autophosphorylation of the added recombinant kinase. Asterisks indicate substrates p38γ displays a high affinity for relative to the other two isoforms.