IN SILICO PREDICTION OF THE CASPASE DEGRADOME

Lawrence Wee Jin Kiat

A THESIS SUBMITTED FOR THE

DEGREE OF THE DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgement

This thesis would not be possible without the following people:

- Professor Shoba Ranganathan, for her supervision and support over the entire course of my PhD candidature.
- Associate Professor Tan Tin Wee, for his insightful comments and advice on my work.
- My colleagues and friends at the Department of Biochemistry for always being there to assist me: Justin Choo, Victor Tong, Vivek Gopalan, Bernett Lee and Kong Lesheng.
- My parents, for their patience, love and support.

Table of Contents

List of figures

List of tables

Abstract

Caspases belong to a unique class of cysteine proteases which play critical roles in important processes such as cell death, differentiation and inflammation. The central feature of caspase function resides on their ability to selectively cleave cellular proteins at specific recognition motifs. The caspase degradome, or the natural repertoire of caspase substrates, spans across a multitude of functional classes, from DNA binding proteins to cell-surface receptors to viral proteins. With more than 300 substrates characterized to date and many more expected to be discovered, the proteome-wide identification of caspase substrates presents a refreshing direction for deepening our understanding of caspase biology in health and disease. In this thesis, a series of computational studies were conducted to meet this goal. Firstly, data on experimentally-verified caspase substrates was meticulously extracted from literature, cleaned and deposited into a web-accessible database (www.casbase.org/casvm/squery/index.html). Secondly, using datasets constructed from the database, a support vector machines (SVM) system was developed to predict for caspase cleavage sites on protein sequences. The SVM method was shown to be comparable, if not better than existing algorithms for predicting caspase cleavage sites. A web server, CASVM (www.casbase.org/casvm/index.html) incorporating the SVM method was developed for the scientific community. Thirdly, as a measure towards predicting caspase substrates, a two-step prediction model, incorporating the SVM method and structural factors (e.g. solvent accessibilities and secondary structures) for substrate cleavage was designed. The two-step model was shown to enhance prediction accuracy by reducing the proportion of false positives from cleavage sites prediction. Lastly, the SVM method was used to predict for potential

caspase substrates among the family of receptor tyrosine kinases. The results suggest that these receptors could be commonly regulated by caspase cleavage and implicate them as agents that mediate both cell survival and death.

Chapter 1: The Caspase Degradome

1.1 Introduction

Proteolysis is a distinctive class of mechanism for cellular control in all living organisms (Barrett *et al*., 1998). Proteases (also known as proteinases, peptidases or proteolytic enzymes) represent the proteolytic engines of the cell, cutting and dicing up cellular and extracellular proteins, though the catalysis of peptide-bond hydrolysis. Constituting 1.7% of human genes, proteases form the largest enzyme family with 566 members - larger than the kinases family and second only to the transcription factors family in size (Puente *et al*., 2003). Proteases are intimately involved in the initiation, modulation and termination of a myriad of essential biological processes such as DNA replication, cell cycle control, cell proliferation, differentiation, migration, morphogenesis, tissue remodeling, haemostasis, immunity and apoptosis. Not surprisingly, aberrant protease activity has been implicated in many pathological, life-threatening conditions such as cancer, neurological diseases and heart abnormalities.

The hallmark of a protease's activity resides in the catalysis of specific and non-reversible hydrolysis of the peptide bond between two amino acids. The protease substrate binds specifically to the protease at a uniquely structured cleft called the active site. The protease active site constitutes a set of subsites which serve as specific binding pockets for the residues on the substrate. The specific accommodation of a substrate residue to each subsite ensures that only a restricted set of sequences on the substrate is cleaved (Figure 1-1). The catalysis of the scissile bond hydrolysis is mediated by a key amino acid on the protease which serves as the catalytic

nucleophile. Consequently, five classes of proteases have been categorized in accordance to the nature of the catalytic nucleophile; namely the serine, cysteine, threonine, metallo, or aspartyl proteases.

While proteases were originally known for their role as digestive enzymes, they are increasingly being recognized as signaling molecules through specific substrate cleavage. Proteolysis have been shown to generate an eclectic range of changes in the structure and function of the target protein such as repression of the protein's function through the removal of a catalytic domain, or a severance of an inhibitory domain leading to enhanced protein activity and even a complete transformation of the protein's intended cellular role. In any case, the functional identity of a protease is often defined by the uniqueness and extent of its substrate repertoire. For example, the matrix metalloproteinase family of proteases mediates tissue re-modeling, cell migration and cancer metastasis through the cleavage of extracellular components such as collagen (Overall and Blobel, 2007). Granzymes, a class of serine proteases, serve as potent initiators of apoptosis where they cleave and activate upstream pro-apoptotic signaling molecules (such as Bid) during cytotoxic T-cell mediated cell death (Lord *et al*., 2003).

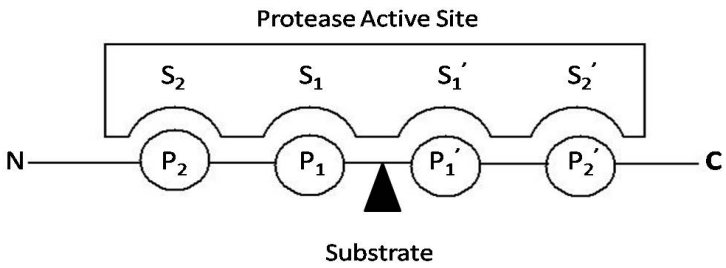


Figure 1-1 Schematic diagram of hypothetical protease-substrate interaction at protease active site as suggested by Schecter and Berger (1967). Substrate residues (P_2-P_1-P_1'-P_2') in contact with protease active site subsites (S_2-S_1-S_1'-S_2') respectively. Proteolytic cleavage (black triangle) occurs at the peptide bond (also called the scissile bond here) between P_1 and P_1' residues.

In recent years, a family of cysteine proteases, called caspases, has been a subject of great interest to researchers (Yuan *et al*., 2003). These proteases are found to cleave a bewildering array of cellular substrates, ranging from membrane structural components to signaling molecules to transcription factors. The selective cleavage of cellular proteins by these proteases implicate them as important signaling molecules in the initiation and execution of apoptosis, as well as in other important cellular processes such as inflammation and differentiation.

In an attempt to further the understanding of caspase biology, a series of computational studies were initiated on the atypical repertoire of caspase substrates. The studies revealed a much greater level of complexity in the mechanistic regulation of substrate cleavage and unraveled additional modes of regulation by caspases in signaling pathways. However, before the discourse on these findings, a comprehensive review on caspase biochemistry and their substrate repertoire will be presented.

1.2 Caspase Biology

1.2.1 Discovery

The term caspase is the short for **c**ysteinyl **asp**artate prote**ase**. The discovery of caspases was dated back to 1992 when ICE (for interleukin-1β converting enzyme), also known as caspase-1, was identified as the protease that cleaves pro-interleukin-1β to its pro-inflammatory active form, establishing the involvement of ICE in mediating inflammation (Thornberry *et al*., 1992; Cerritti *et al*., 1992) At about the same time, genetic studies on the cell death pathway in the nematode, C. elegans, identified the

ced-3 gene as being essential for programmed cell death in the worm during development (Yuan and Horvitz, 1990). Later, the cloning of the *ced-3* gene by Yuan and co-workers led to the identification of the gene as a C. elegans homologue of the mammalian ICE (Yuan *et al*., 1993). These observations suggested a conserved programmed cell death mechanism involving these cysteine proteases. Subsequent work led to the identification of other cysteine proteases with homology to the *ced-3* gene. The sequential characterization of *ced-3* homologues by various research groups took place without a controlled nomenclature but was eventually standardized with the name "caspase" in 1996 (Alnemri *et al*., 1996). At present, total of 12 caspases have been identified in mammals: caspase-1 to -10, caspase-12 and caspase-14 (Yuan *et al*., 2003; Pistritto *et al*., 2002). The protein initially named caspase-13 was later found to represent a bovine homolog of caspase-4 (Koenig *et al*., 2001), and caspase-11 is most likely the murine homolog of human caspase-4 and caspase-5 (Kang *et al*., 2000).

1.2.2 Caspase Structure and Activity

As reviewed as Nicholson (1999) and Yuan *et al*. (2003), the distinguishing trait in all members of the caspase family is the specificity for substrate cleavage after an Asp residue at P_1, a trait which is exceptional among mammalian proteases. The primary specificity pockets at P_1 in caspases are almost identical, being formed by the side chains of the strictly conserved residues, Arg-179, Arg-341 and Gln-283 (caspase-1 numbering). This deep, highly basic pocket is perfectly shaped to accommodate an Asp side chain with a much lower efficiency for a Glu. In addition, caspases are shown to preferentially recognize and cleave at unique tetrapeptide

Table 1-1 Optimal tetrapeptide specificities of caspases

Group	Member	Tetrapeptide Specificity ($P_4P_3P_2P_1$)
I	Caspase-1	WEHD
	Caspase-4	(W/L)EHD
	Caspase-5	(W/L)EHD
II	Caspase-3	DEVD
	Caspase-7	DEVD
	Caspase-2	DEHD
III	Caspase-6	VEHD
	Caspase-8	LETD
	Caspase-9	LEHD

signature motifs (P_4-P_3-P_2-P_1) on substrates. Accordingly, these proteases are grouped into three categories based on their optimal tetrapeptide cleavage sequences as characterized by Thornberry and co-workers (Thornberry *et al*., 1997) using *in vitro* combinatorial methods (Table 1-1). The differences between the optimal sequence preferences are largely attributed to the requirement of the P_4 residue. Group I caspases (caspases-1, caspase-4 and caspase-5) share a preference for residues with bulky, hydrophobic side chains at the P_4 substrate position (tryptophan or leucine), while Group II caspases (caspase-2, caspase-3 and caspase-7) prefers the negatively charged aspartic acid and Group III caspases (caspase-6, caspase-8, caspase-9) is optimized for leucine or valine. All groups, however, have an absolute preference for aspartic acid at P_2 and a hydrophobic P_3 residue.

Besides their preference for specific cleavage site sequence motifs, caspases share a number of other distinctive features. The catalysis of protein cleavage is governed by a critical cysteine residue, which is a part of a conserved QACXG (X =

C, G, Q or R) pentapeptide sequence motif. The caspase enyzme is synthesized as an inactive zymogen which contains a prodomain, a large and small subunit. Enyzme activation is activated upon proteolytic cleavage by other members of the caspase family, or by other proteases such as granzyme B. The activation process is carried out sequentially, firstly through cleavage and separation of the large and small subunits, followed by the removal of the prodomain after another cleavage event on the large subunit. The large subunit and small subunit then associate with each other to form a heterodimer which unites with another identical heterodimer, generating an active tetrameric caspase molecule. Each tetrameric caspase molecule contains two active sites, one from each heterodimer. A structural model of caspase-3 is illustrated in Figure 1-2.

Figure 1-2 Structure of caspase-3. The structure of active caspase-3 is a tetramer comprising of two copies each of the p17 (large) and p12 (small) subunits from residues 35-173 and 185-277 of each proenzyme (*green* and *blue*) respectively. The four polypeptide chains associate to form a compact (p17/p12)2 tetramer containing two active sites. The p17 and p12 subunits interact extensively with each other with the core of the enzyme being formed by a central 12-stranded β-sheet. Each p17/p12 dimer donates six strands. At the enzyme active sites, each copy of the inhibitor (Ac-DVAD-fmk) binds in a narrow cleft across the C-terminal end of the central β-sheet. Image from the Protein Data Bank [PDB ID: 1CP3, Mittal *et al*. (1997)].

1.2.3 Caspase Function

Much of the current understanding of caspase function is derived from studies on their role in apoptotic cell death – a form of programmed cell death conserved in metazoans (reviewed in Hengartner, 2000). Caspases execute downstream biochemical changes in the apoptotic cell such as shutting down of basic survival processes, termination of growth signals and dismantling cell architecture. Caspases are activated via two pathways in apoptosis: intrinsic and extrinsic (Figure 1-3). The extrinsic pathway is initiated when the CD95 ligand binds to the CD95 death receptor, leading to receptor oligomerization and the formation of the death inducing signaling complex. The multi-protein complex recruits, via adaptor protein FADD, multiple pro-caspase-8 molecules, leading to the cleavage and activation of the enzymes through an induced proximity mechanism. Upon activation, caspase-8 functions as an initiator caspase as it cleaves and activates downstream pro-caspase-3. Active caspase-3 serves as an executioner caspase by cleaving a myriad of downstream cellular proteins, generating the phenotypic changes observed in the apoptotic cell.

In contrast, the intrinsic pathway is initiated when cellular perturbations such as genotoxic stress or internal insult propagate downstream pro-apoptotic signals that converge at the mitochondria. Pro-apoptotic and anti-apoptotic members of the Bcl-2 family of apoptotic regulators meet at the surface of mitochondria, where they compete to regulate the release of the pro-apoptotic molecule such as cytochrome c and Smac. When the balance is tipped in favor of the pro-apoptotic Bcl-2 regulators, the mitochondria permeability becomes compromised, leading to the release of the pro-apoptotic molecules. Cytochrome c, upon release into the cytoplasm, associates with Apaf-1, pro-caspase-9 and other molecules to form a heptametrical protein

complex called the apoptosome. Within this multi-protein complex, pro-caspase-9 is activated and goes on to cleave and activate downstream pro-caspase-3 molecules. Both extrinsic and intrinsic pathways converge at the level of caspase-3 activation. Caspase-3 activation and activity is antagonized by the IAP molecules, which themselves are antagonized by the Smac protein released from mitochondria. Notably, cross-talk and integration between the extrinsic and intrinsic pathways is mediated by Bid, a pro-apoptotic Bcl-2 family member. Active caspase-8 cleaves Bid during extrinsic apoptotic signaling which translocates to the mitochondria, abrogates the activity of anti-apoptotic Bcl-2 proteins and mediates cytochrome c exit. Also, positive feedback regulation is mediated through caspase-3 cleavage of upstream initiator caspases; caspase-8 and caspase-9.

Interestingly, while most caspases are involved in apoptosis, either as initiator or executioner caspases, emerging evidence has implicated them in a plethora of other vital non-apoptotic processes (Table 1-2), suggesting a more disparate and complex role of these enzymes in the cell (Launay *et al*., 2005; Siegel, 2006). Caspase-1, caspase-5 and caspase-11 are thought to be inflammatory caspases as they are primarily involved in the processing of the inflammatory cytokines. Remarkably, key apoptotic caspases such as caspase-3, caspase-6 and caspase-8 have been shown also to mediate immune cell proliferation in addition to apoptotic cell death. Many inflammatory or apoptotic caspases were also shown to mediate differentiation of a wide variety of cell types such as erythroblasts, macrophages, lens epithelial cells, osteoblasts and keratinocytes.

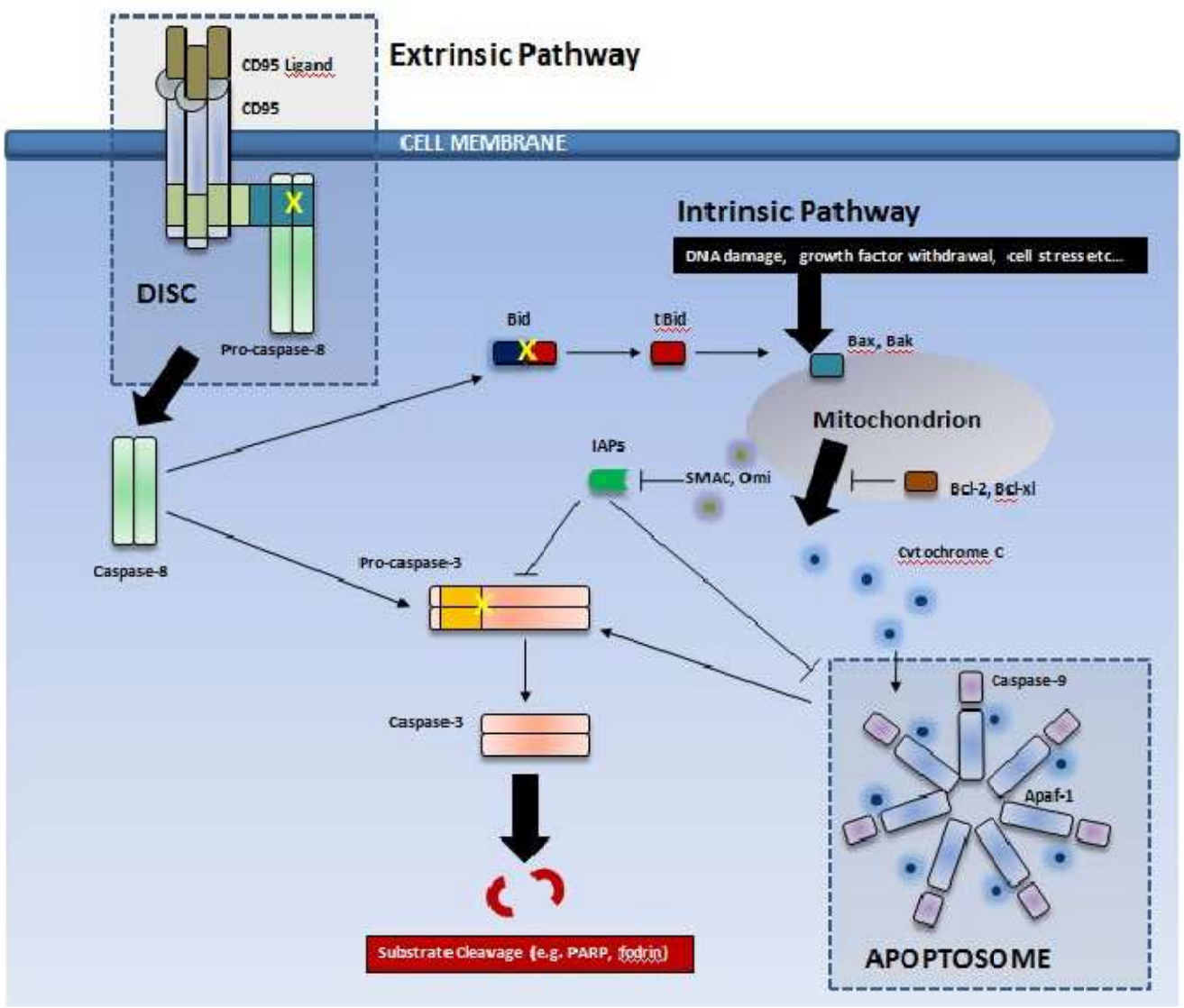


Figure 1-3 Two major pathways in apoptosis: intrinsic and extrinsic.

Table 1-2 Functional roles of caspases in biological processes

	Apoptotic Roles	Non-Apoptotic Roles
Caspase-1	Induced apoptosis when over-expressed.	Cleaves and activates pro- IL-1 and pro-IL-18 Component of inflammasome. Differentiation of skeletal muscle Cell migration
Caspase-2	Initiator of extrinsic pathway or executioner caspase	Differentiation of erythroblasts, osteoblasts and macrophages Involved in DNA repair
Caspase-3	Executioner caspase	Differentiation of erythroblasts, keratinocytes, macrophages, lens epithelial cells, sperm, skeletal muscle, osteoblasts and placental trophoblasts. Negative cell cycle control in B cells IL-16 production Platelet formation Brain development
Caspase-4	Might be involved in ER-stressed apoptosis	Might be involved in bacteria-induced cell death
Caspase-5	NA	IL-1 production, component of inflammasome that activates caspase-1
Caspase-6	Executioner caspase	Differentiation of lens epithelial cells. Positive cell cycle control in B cells
Caspase-7	Executioner caspase	Differentiation of erythroblasts

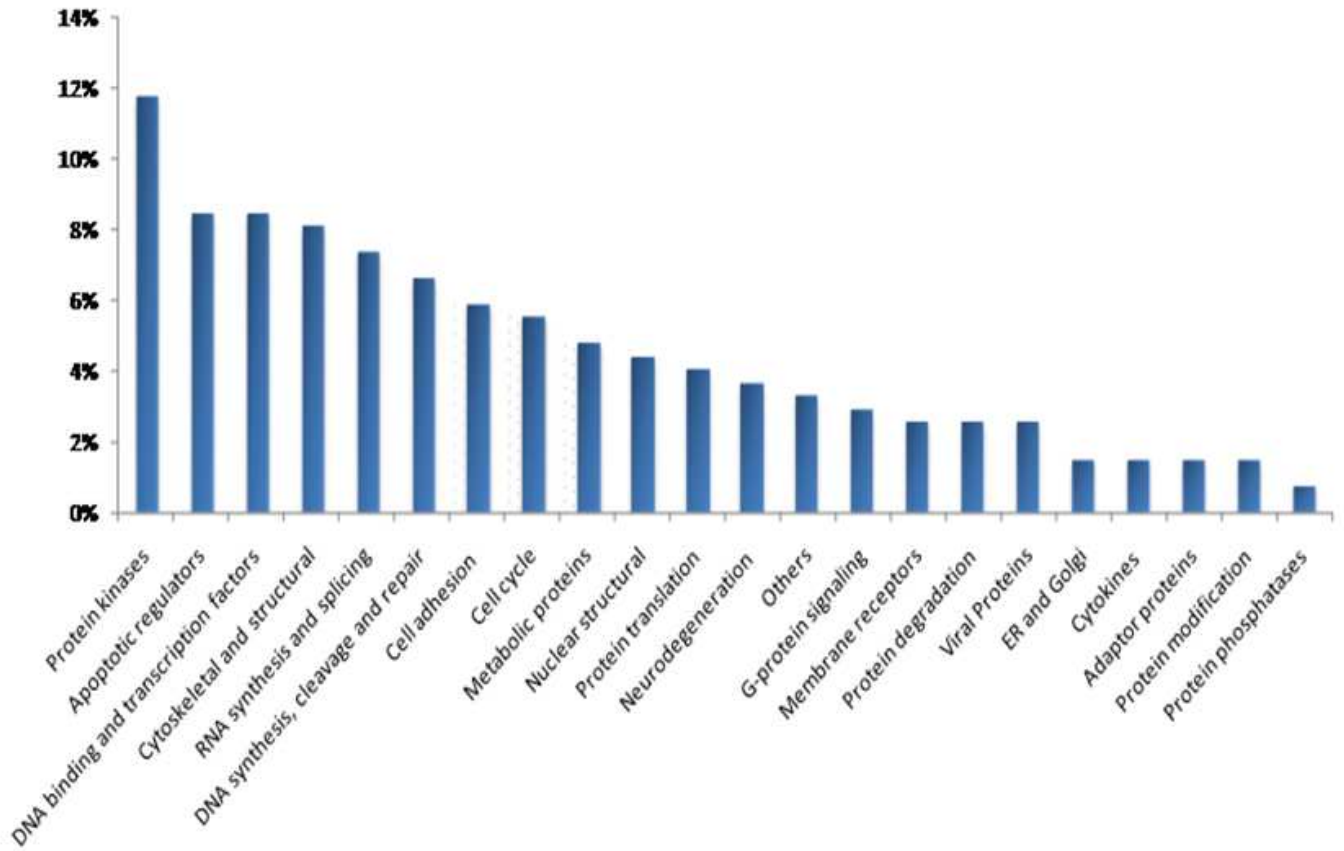
	Apoptotic Roles	Non-Apoptotic Roles
Caspase-8	Initiator caspase of extrinsic pathway	T cell proliferation and activation Positive cell cycle control in B cells Differentiation of placental trophoblasts, osteoblasts, erythroblasts, monocytes Internalization of death receptors
Caspase-9	Initiator caspase of intrinsic pathway	NA
Caspase-10	Initiator caspase of extrinsic pathway	NA
Caspase-11	Initiator caspase of extrinsic pathway	IL-1 production
Caspase-12	Initiator caspase in ER-stress induced apoptosis	Attenuates inflammation Involved in innate immune response
Caspase-14	NA	Differentiation of keratinocytes

1.2.4 Caspase Substrates

In 1998, the first list of caspase substrates was compiled in Earnshaw *et al*. Most of the substrates known at that time (from a total of 65) could be categorized into only a few functional groups, such as structural or scaffolding proteins in the cytoplasm and in the nucleus, signal transduction proteins, transcription factors, cell cycle controlling components and proteins involved in DNA replication and repair. More recently, Fischer *et al*. (2003) updated the compilation to more than 280, with proteins belonging to an even greater range of functional groups (Figure 1-4). Not surprising, transcription factors, DNA cleavage and repair proteins, RNA-associated proteins and proteins involved in cytoskeletal structures represented a large proportion of the characterized substrates. Notably, a much greater proportion of signal transduction proteins – such as protein kinases, G-protein signaling components and membrane receptors – were mentioned in the update.

The growing list of substrates from vastly different functional groups suggests a much varied range of consequences of substrate cleavage as well as a more complex role of caspases in biological processes beyond apoptosis. However, caspase cleavage remains a highly selective process where target proteins are cleaved at specific recognition sites and purposeful changes in protein function are effected. As described in the excellent review on caspase substrates by Fischer *et al*. (2003), most outcomes of caspase cleavage are implicated in apoptosis and are broadly classified into two distinct categories: gain or loss of function of protein. The following sections summarize the salient points mentioned in the review.

Figure 1-4 Functional distribution of caspase substrates. Data from Fischer *et al*. (2003).



1.2.4.1 Gain of Function

In many cases of caspase cleavage, the cleaved substrate exhibits an increased level of activity - often through the removal of regulatory or inhibitory domains - leading to the downstream enhancement or propagation of the apoptotic process. The most striking example for caspase-mediated gain of protein function is that of the caspase itself. As mentioned earlier, caspase cleavage of executioner caspases - caspases-3 and caspase-7 - by upstream initiator caspases such as caspase-9 and caspase-8 is required for complete enzyme activation. Several members of the PKC and MAP kinase pathway kinase, such as PAK2 and ROCK-1, were shown to be constitutively activated upon separation of the N-terminal regulatory and the C-terminal catalytic domains through caspase cleavage. Activation of PAK2 and ROCK-1 is important for the expression of the apoptotic phenotype such as cytoskeletal reorganization and plasma membrane blebbing. Cleavage of several MST kinases by caspase-3 also yields constitutively active molecules which are potent inducers of apoptosis.

Interestingly, caspase cleavage of certain proteins also led to the exposure of previously hidden pro-apoptotic domains on the native protein, converting these proteins into cell death effectors. Caspase cleavage of MEKK1 leads to the exposure of a kinase fragment which induces downstream caspase activation, causing a positive feedback loop for apoptosis. Bid, a BH3 domain only member of the Bcl-2 family of apoptotic regulators, is cleaved by caspase-8 and translocates to the mitochondria, inducing cytochrome c release. The cleavage of Bid exposes a previously occluded hydrophobic binding surface of its BH3 motif which is important for antagonizing the anti-apoptotic Bcl-2 regulators. Similarly, its pro-apoptotic cousin, $Bim_{EL,}$ another

BH3-domain only protein, was found to demonstrate a higher affinity for Bcl-2 and a markedly enhanced apoptotic activity after cleavage of its N-terminal region.

1.2.4.2 Loss of Function

On the flip side, caspase cleavage results in the loss of native protein function. Not surprisingly, many proteins belonging to this category belong to the scaffold and structural proteins of the cell– cleavage of these proteins is instrumental for the observed apoptotic phenotype such as nuclear fragmentation, cell shrinkage and membrane blebbing. For example, the DNase inhibitor, ICAD is cleaved by caspases, liberating active CAD nuclease that mediates DNA fragmentation. Poly-(ADP-ribose) polymerase or PARP, is an abundant nuclear protein that catalyses poly-(ADP-ribose) ligation to acceptor proteins, including itself, in response to DNA strand breaks. PARP cleavage by caspases-3 and caspase-7 bisects a bipartite nuclear localization signal, generating a form of the protein that cannot synthesize ADP-ribose polymers in response to damaged DNA. Caspases also terminates several proteins involved in maintenance of the cytoskeletal architecture such as the intermediate filaments cytokeratin-18 and vimentin. Cleavage of golgin-160 and GRASP65 was suggested to cause disassembly of the Golgi complex, and proteolysis of Bap31 disrupts the transport between the endoplasmic reticulum and the Golgi complex. Caspase cleavage of acinus and helicard was found to contribute to chromatin condensation and nuclear remodeling respectively.

Proteins directly involved in anti-apoptotic signaling pathways were found to be cleaved and inactivated as well. The inhibitors of caspase activity, c-FLIP and c-IAP, are inactivated after caspase cleavage. The cleaved fragment of c-IAP is pro-

apoptotic and leads to downstream amplification of apoptosis. Cleavage of the anti-apoptotic regulators of the Bcl-2 family, Bcl-2 and Bcl-xl results in the removal of the N-terminal BH4 domains which not only leads to a loss of their anti-apoptotic function, but also converts them to pro-apoptotic proteins.

Signaling proteins involved in anti-apoptotic pathways, such as kinases and transcription factors, are inactivated through caspase cleavage during apoptosis. In their native state, Akt and Raf - components of survival pathways in the cell - inactivate pro-apoptotic molecules such as Bad. Caspase cleavage inactivates these molecules and contributes to a positive feedback loop in apoptosis. Anti-apoptotic transcription factors such as NF-κB were shown to mediate positive feedback loops through proteolytic cleavage. The cleavage of the p65 subunit of NF-κB generates a protein that is still able to bind to DNA but lack trans-activating activity, therefore repressing the transcription of downstream regulators by functioning as a dominant negative inhibitor. Also, the NF-κB inhibitor, IκB is converted to a constitutive protein that is no longer degraded by the proteosome upon cleavage of its N-terminal by caspases.

1.2.4.3 Non-apoptotic consequences of caspase cleavage

While the majority of proteins cleaved by caspases involve modulation of apoptotic signaling and/or changes in cellular integrity, the involvement of caspases in non-apoptotic processes suggest that a notable proportion of substrates not directly implicated in apoptosis are important as well. Several negative regulators of the cell cycle are cleaved by caspases, leading to their inactivation. Wee1 is a critical component of the G2/M cell cycle checkpoint machinery and mediates cell cycle

arrest by phosphorylation of Cdc2. Caspase cleavage of Wee1 in proliferating cells was shown to inactivate the protein, leading to cell cycle progression. Inflammatory cytokines such as pro-interleukin-1-beta, pro-interleukin-16 and pro-interleukin-18 are converted into their active state via caspase cleavage. More significantly, caspases are also found to be involved in the propagation of neurodegenerative diseases such as Huntington's and Alzheimer's. In Huntington's disease, caspase cleavage of Huntingtin abrogates its native protective function and generates a toxic N-terminal byproduct that sensitizes neurons to further stressors such as excitotoxic stimulation. Caspases are implicated in the progression of Alzheimer's disease through the proteolysis of the trans-membrane APP protein at its cytosolic tail which releases the neurotoxic fragment, C31.

1.3 The Caspase Degradome

1.3.1 Emerging perspectives

In a similar fashion to the genome and the proteome, López and Overall (2000) coined the termed "degradome" to represent the complete set of proteases that are expressed at a specific time by a cell, tissue or organism. The natural substrate repertoire of an enzyme in a cell, tissue or organism is termed as the protease degradome. Elucidating the degradome will help assign proteases to biological pathways and delineate the protease's physiological and pathological roles (Overall and Blobel, 2007). Furthermore, as protease degradomes are connected with another through promiscuous partnerships of the same substrate to multiple upstream proteases, characterization of individual protease degradomes will further clarify the roles and significance of each protease and their downstream proteolytic events at the

systems level. Undoubtedly, the knowledge of protease degradomes will be useful for therapeutic research and drug discovery. Despite their potential, however, the protease degradomes of all proteases remain to be fully elucidated. To be sure, for many recently discovered proteases and even for many established proteases, no native substrates are known. Evidently, one of the most important tasks today in protease biology is defining the protease degradome.

The bewildering array of caspase substrates has brought several major questions into focus. For instance, what is the minimal set of proteins that must be cleaved in order to induce the phenotypic hallmarks of apoptosis? Are there bystander proteins which get inadvertently cleaved alongside the mandatory set of apoptotic substrates? If so, just how extensive is the "collateral damage"? In addition, the extensive array of native substrates begs the question on how caspase substrate cleavage is differentially coordinated in apoptosis and presumably unrelated events such as cell proliferation and differentiation. In any case, it is highly likely that the cleavage of caspase substrates is a tightly regulated event – only selected proteins are cleaved under particular cellular conditions – and the dysregulation of caspase substrate cleavage is expected to contribute to abnormal physiology and the progression of human diseases.

To date, many more caspase substrates are expected to be discovered and the functional consequences of several cleaved substrates remain uncharacterized. Accordingly, efforts to elucidate the caspase degradome will offer an alternative perspective for unraveling the complexities of regulating caspase substrate cleavage and its downstream consequences in cell biology.

1.3.2 Methodology challenges

Several classes of experimental methods are available to characterize protease substrates (López and Overall, 2000). The characterization of protease substrates has traditionally involved serial biochemical processes in which putative protein substrates, either purified from cell extracts or *in-vitro* translated, are incubated with proteases and analyzed for cleavage. Newer approaches involving genetics-based techniques and high-throughput proteomic tools have also been used extensively for substrate identification in recent times. In the former, proteins are analyzed for cleavage using specific gene-disrupted animal models. For example, a mutated substrate gene – expressing non-cleavable cleavage sites – can be used for detecting protease activity by observing the functional differences between the knock-out animals and controls. In other genetic-based methods, experiments could involve utilizing the yeast-two-hybrid method where the ancillary exosite domains of proteases can be used as baits to screen cDNA libraries for interacting proteins – the observation of a binding partner could represent a potential protease substrate.

More significantly, degradomics – or the application of proteomic approaches for the direct investigation of proteases and proteolytic processing in a system-wide context – has greatly advanced the field of substrate identification. In the majority of degradomics experiments, large protein sets treated with or without the protease, are separated by gel-based or liquid chromatography approaches, and individual proteins are identified by tandem mass spectrometry of tryptic peptides. Many studies use cell lysates or cell-conditioned medium as large substrate libraries for exogenously applied proteases, whereas other more ambitious approaches use cell-based systems involving protease deficient or protease over-expressing cells. Coupled with

successive validation of substrate candidates by applying an array of complementary techniques, new protease substrates are being uncovered in a high-throughput manner.

The quintessential substrate detection tool will be capable of detecting proteolytic cleavage products of natural substrates in their biological context such as in cell-based systems and in tissue samples. While much progress has been made in advancing techniques for experimental identification of substrates, individual limitations in the methods suggest that no one method will lead to the discovery of all substrates in the protease degradome. To this end, it would be perceptive that future work be focused not only on improving existing methods but also on the development of complementary approaches.

Over the past decade, the deluge of "omics" data in biology has rendered the creation of a whole generation of predictive computational algorithms and tools to assist research in many subject domains in molecular and cell biology, ranging from the detection of transcription factor binding sites to prediction of protein-protein interactions to the identification of signal peptide cleavage sites (Brazas *et al*., 2008). It is not surprising that computational methods for predicting protease substrates will serve as useful complements to experimental methods which can be cumbersome and time consuming. As part of the protease biologist's arsenal of research tools, such computational approaches will help accelerate hypothesis generation and narrow the scope of experimentation. Notably, as studies clarify the mechanisms and regulation of caspase substrate cleavage and with mounting data on caspase substrates, it is plausible that reliable tools for the computational prediction of caspase substrates can be developed to assist the experimenter.

1.4 Thesis Objectives

The content of this thesis is centered on two primary objectives:

1. Elucidate the caspase degradome by identifying known and hitherto undiscovered caspase substrates.
2. Explore the application of predictive computational methods for the above purpose.

The following summarizes the studies described in each chapter:

In Chapter 2, the problems related to data for the computational prediction of caspase substrates was discussed. Data on caspase substrates was retrieved from literature and a database was developed to store and manage the data.

In Chapter 3, a novel approach to predict for caspase cleavage sites was developed using the Support Vector Machines (SVM) algorithm. It was trained and tested with experimental data from Chapter 1 and was shown to perform better than existing tools. A web server for predicting caspase cleavage sites using the developed algorithm was constructed.

In Chapter 4, a multi-factor model comprising of the SVM algorithm for cleavage site prediction and quantitative measures of substrate structural properties was developed to improve accuracy of caspase substrate prediction.

In Chapter 5, the receptor tyrosine kinase family (RTK), an important class of survival and growth signaling molecules, was predicted for potential caspase substrates. Prediction results suggest a novel mechanism of RTK regulation by caspases and implications in apoptosis.

In Chapter 6, the thesis concludes with a summary of the previous studies and discusses these implications in the context of predicting the caspase degradome.

Chapter 2: Data

2.1 The Data Challenge

Data integrity is of paramount importance to the entire cycle of research and development of computational prediction systems. Intuitively, the use of data for computational prediction underscores two challenges: quality and quantity. As the adage goes 'garbage in, garbage out', non-precise and inaccurate data may lead to spurious conclusions, while insufficient data will undermine statistical significance of predictive patterns, leading to less robust models. Accordingly, most computational prediction tools for biological problems have been developed using expertly curated data found in public or proprietary databases. For example, MHC-BPS (Cui *et al*., 2006) and POPI (Tung and Ho, 2007), which predict for MHC binding peptides using protein sequences, rely on curated datasets of MHC binding peptide sequences derived from databases such as MHCBN (Bhasin *et al*., 2003), MHCPEP (Brusic *et al*., 1997) and SYFPEITHI (Rammensee *et al*., 1999). Similarly, signal peptide prediction tools such as SignalP (Bendtsen *et al*., 2004) and Signal-3L (Shen and Chou, 2007) utilizes data on signal sequences deposited in Uniprot (The Uniprot Consortium, 2008).

To develop prediction tools for caspase substrates, it is obligatory for analyses to be carried on data on experimentally verified caspase-cleaved proteins. However, unlike the prediction tools mentioned earlier, there are no specialist resources or references where sufficient quantity of the required data can be extracted for use in this case. All existing tools for prediction of caspase substrates such as PeptideCutter (Gasteiger *et al*., 2005), GraBCas (Backes *et al*., 2005) and CasPredictor (Garay-

Malpartida *et al*., 2005) are based on *in vitro* cleavage site tetrapeptide specificities reported previously by Thornberry and co-workers (1997) and not data from legitimate caspase-cleaved substrates. While *in vitro* data is undoubtedly an important component for analysis of caspase substrate cleavage, it is not sufficient for creating robust prediction models. Factors affecting substrate cleavage such as the adequate exposure of the cleavage site region to protease, presence of exosites on substrate or other post-translational regulation need to be accounted for in the prediction model as well – the failure to do so will greatly mitigate prediction accuracy (this subject will be discussed further in Chapter 4 and Chapter 6). Moreover, inadequate data were used for testing the validity of these tools as well. For the dataset of 280 cleavage site sequences used in the development of CasPredictor, no description on data retrieval or data cleaning was reported and the final dataset was unavailable for download on the website. The bio-basis function neural network-based prediction algorithm by Yang was created based on a limited dataset of 18 cleavage site sequences which was not also reported (Yang, 2005). In addition, caspase sequences in general protein databases such as Uniprot and GenBank (Benson *et al*., 2008), despite being well annotated with post-translational modifications such as signal sequences and phosphorylation sites are yet to be supplemented with data on their natural substrate repertoire or preferred substrate cleavage sequences.

Evidently, the absence of reliable sources of high quality data on experimentally-defined caspase substrates is likely to limit the development of predictive algorithms and tools. This suggests that a concerted effort to extract relevant data from literature and making it easily accessible for researchers would be helpful. In this chapter, a two-step approach was carried to address these issues

(Figure 2-1). Firstly, a systematic process was carried to retrieve, clean and construct datasets for data analysis and algorithm construction from experimentally verified caspase substrates reported in literature. Secondly, a web-accessible relational database was developed to store and manage the datasets. The database was supplemented with tools for efficient data retrieval and knowledge discovery in accordance to data warehousing principles,

2.2 Data Retrieval

2.2.1 Literature Search

The absence of a definitive source of caspase substrates data necessitates a thorough search for *bona fide* caspase substrates. Much of the currently known caspase substrate repertoire has been comprehensively reviewed by Fischer *et al*. (2003) and, to a lesser extent, in Earnshaw *et al*. (1999). As caspase substrates discussed in earlier reviews were included in Fischer *et al*., it is assumed that the latter had a reasonably extensive coverage of reported substrates up till the time of its publication in 2003. However, it is likely that many more substrates would have been discovered and reported in original research papers since then. Therefore, the currently known caspase substrate repertoire would constitute entries compiled in Fischer *et al*. (henceforth termed as Fischer dataset) as well as those separately reported in original papers thereafter (or Post-Fischer dataset).

To construct the Post-Fischer dataset, a comprehensive search on journal articles indexed in PubMed was carried using several permutations of keywords related to caspase-mediated substrate cleavage (e.g. “SUBSTRATES”, “CLEAVAGE”, “CASPASES”) during the period from 1 Jan 2002 through 31 May

2008. The keyword searches were restricted to the title of publications since a more inclusive search category (e.g. searches on abstracts or the entire paper) would be overly time consuming. While it is probable that some relevant journal articles would be filtered out due to the absence of the targeted keywords, such instances are not expected to greatly affect the final dataset. The stated start date was selected to extract journal articles which overlapped with those compiled in the Fischer dataset to ensure that all recently reported substrates were covered. To ensure that only experimentally -verified caspase substrates are selected, all caspase substrates suggested by authors were shown to be cleaved under experimental conditions (e.g. using either serial approaches such as in vitro protease-substrate cleavage assays or through proteomic methods). Substrates implied by authors with no direct supporting experiments for caspase cleavage were eliminated. The data retrieval process extracted a total of 53 caspase substrates for the Post-Fischer dataset. For the Fischer dataset, 260 caspase substrates reported in Fischer *et al*. were extracted.

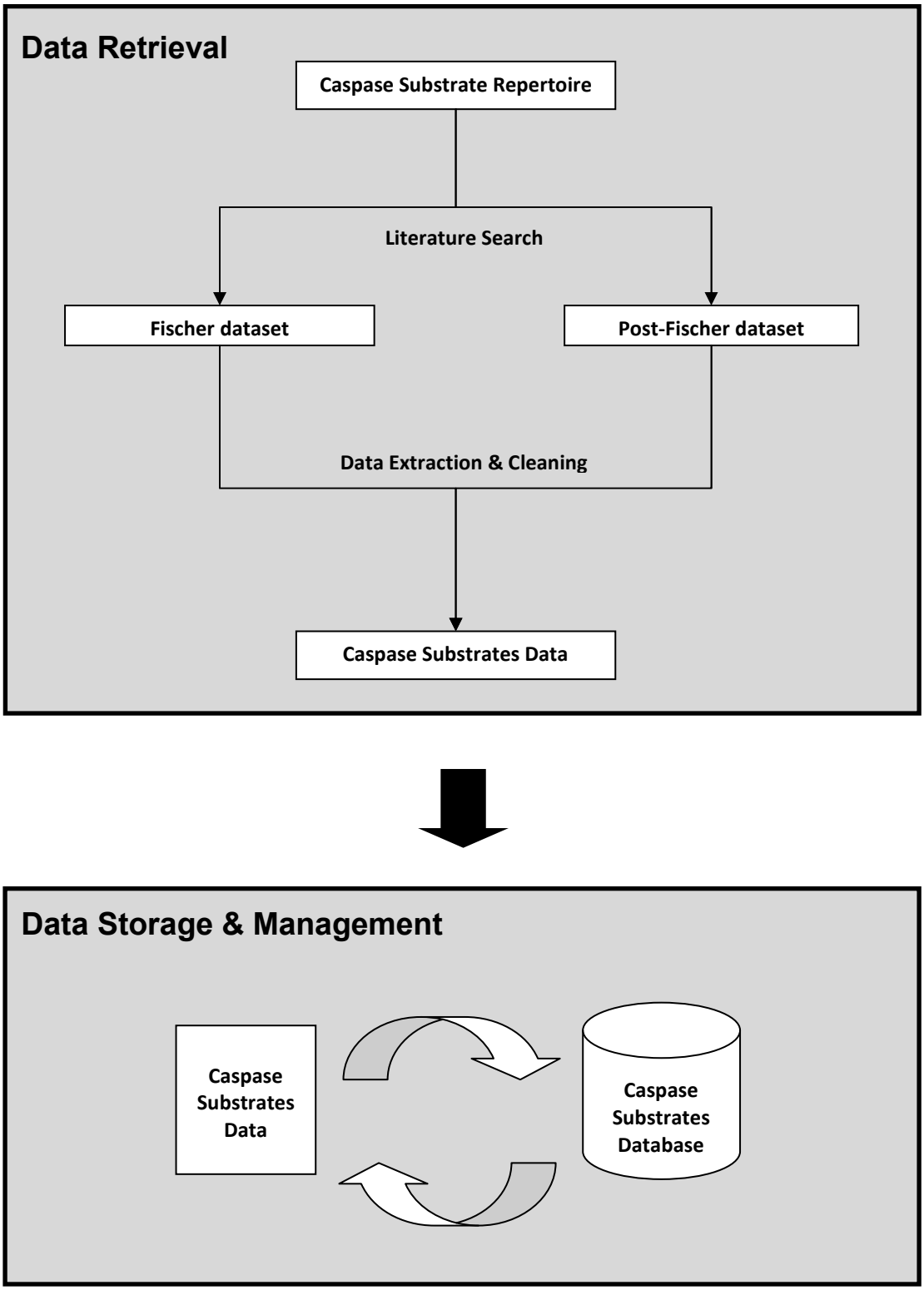


Figure 2-1 Schematic diagram depicting the processes and output involved in data retrieval, storage and management of caspase substrates.

2.2.2 Data Extraction and Cleaning

The amino acid sequences and location of experimentally-verified cleavage sites on substrates were extracted from literature. Cleavage sites implied or suggested by authors with no supporting experiments were noted as "putative". Also, the protein sequences of substrates were obtained from Uniprot databases (Swiss-Prot and TREMBL) through keyword searches. All substrates data were checked for ambiguities, typos, and other forms of errors through cross-referencing with the original literature and online databases. For example, the α-tubulin protein was erroneously reported as a caspase substrate in Fischer *et al*. with an unknown cleavage site. The cleavage site was updated to $LEKD^{431}$ when cross-referenced with the original reporting paper. In another example, the cleavage site location of DCC protein was reported at Asp^{794} in Fischer *et al*., but was corrected to Asp^{1290} after a confirmation with the original reporting paper.

The final datasets of caspase substrates are listed in Appendix A (Table A-1 and Table A-2).

2.3 Data Storage and Management

2.3.1 The Biological Data Warehouse

One of the most pertinent changes in biological research in the past decade is the burgeoning usage of databases for biological data (Galperin, 2008). The deluge of data from "omics"-based research - a result of the ubiquitous advances in high throughput technology - presents unique challenges in data storage and management. These challenges motivate the need to create robust and scalable database

architectures for data storage as well as to develop integrative workflows and tools for data management.

The issue of data storage has been tackled by the collaborative efforts of primary databases such as GenBank (Benson *et al*., 2008), DDBJ (Tateno *et al*., 2002) and EMBL-Bank (Kulikova *at el*., 2006). These databases serve as general one-stop shops for the deposition and retrieval of biological sequence data and annotations. While these repositories excel at data storage, the very nature of their size and architecture presents limited usage of these tools as knowledge platforms for biological research of a specific field. Large databases necessarily present a cumbersome and tedious process of retrieval of specific datasets where redundancies and errors are commonplace. Also meta-data for specific biological domains, while important for research, cannot be conveniently integrated or retrieved within these databases.

One answer to these constraints may reside on the use of boutique biological databases or data-warehouses (Schönbach, 2000). A biological data warehouse is a subject-oriented, expert collection of biological data designed for supporting biological data analysis and knowledge discovery. In contrast with a general-purpose database (such as GenBank), the biological data warehouse appears to suit the needs of niche research. While the general purpose databases focus on expansion and dissemination of information and provide basic annotation, specialized biological data warehouses integrates relevant information from these underlying data sources and merges them with expert curation and annotation. It is not surprising that an increasing number of specialized biological databases are being developed for a plethora of research domains - a collection which totals more than a thousand to date,

as reported in the *2008 Nucleic Acid Research Database Issue* (Galperin, 2008). These databases address the data challenges of subject fields ranging from animal model genomes to cell proteomes to human diseases. Clearly, in the biologist's arsenal of research tools, biological data-warehouses are fast becoming an indispensable addition to the wet laboratory.

2.3.2 The Caspase Substrates Database

To address the challenges of storing and managing data, the Caspase Substrates Database was developed. Based on the conceptual framework of the data warehouse, it aims to be a central resource for expertly curated data on caspase substrates with tools for data retrieval and knowledge discovery. The Caspase Substrates Database is deployed using the MySQL database system (www.mysql.com) which is based on the architecture of the relational databases introduced by E.F. Codd (1970). In a relational database model, data is stored in a collection of inter-related tables - consisting of sets of rows and columns - each assigned to specific categories of data. The relational structure facilitates the extraction of multi-dimensional datasets with user-defined queries and enables efficient expert curation of data. Each database entry in the Caspase Substrates Database describes an experimentally verified caspase substrate, with annotations on sequence, structure and function made available through direct database links to Uniprot and PubMed, as well as indirect links to other useful public databases via Uniprot (See Databases Interconnectivity Chart in Figure 2-2). The database is hosted on a UNIX web server and web interfaces to the database was created with Perl CGI scripts.

As shown in Figure 2-3, the primary interface to the database is a web form where users can query and retrieve data using one of three options. Database entries can be queried using the substrate's database accession ID (also termed as the CASVM Accession ID), or using the substrate's Uniprot Accession ID. Users can also execute queries through the input the keywords of protein or gene names of prospective substrates. Alternatively, users can submit a protein sequence to an integrated BLASTP search tool (Altschul *et al*., 1990) and retrieve a list of structurally similar sequences from the database. Once a query has been made, a list of results will be presented with corresponding links to a page containing details on the entry. Every database entry is annotated with the following fields on the details page (a screenshot is shown in Figure 2-4):

CASVM Accession ID: The unique identifier for all substrates in the database.

Substrate Name: The name of the caspase substrate in the database entry. Each substrate's name is checked for consistency by cross-referencing with the Uniprot protein name and gene name fields. In ambiguous cases, the name as mentioned in literature is selected instead.

Organism Type: The organism(s) from which the substrate was found to be cleaved in. Recent work have suggested that cleavage of certain caspase substrates were not consistent across organisms - orthologs were found to be cleaved in certain organism are not in others (Ussat *et al*., 2000). The disparity of substrate cleavage across organism type is likely to influence the interpretation of the functional role of the substrate in processes mediated by caspase activity. This is particularly important in therapeutic and translation research where observations of substrate cleavage in

model organisms are often extrapolated to human studies. These annotations are obtained from the original literature reporting the caspase substrate.

Uniprot ID: The Uniprot Accession ID for the substrate. A link to the protein entry in Uniprot databases is provided.

Cleavage Site(s): The caspase cleavage site(s) on the substrate as reported in literature. In cases of erroneous or ambiguous data, the Uniprot sequence is stated instead. Sequence of cleavage site is reported in format suggested by Schecter and Berger (2000), followed by the location of scissile bond cleavage (numbering indicates the position of P_1 residue).

Cleavage Effect(s): A brief description of the consequence of substrate cleavage as reported in literature. For substrates derived from the Fischer dataset, the comments from Fischer *et al*. are used.

Caspase(s) Involved: The caspase(s) which is (are) shown to be responsible for cleavage of the substrate as reported in literature.

Comments: Comments related to the cleavage of substrate. Cleavage of substrates assigned as "putative" is noted here.

Sequence: The link to retrieve a protein sequence of substrate in FASTA format.

References: The list of references concerning cleavage of substrate.

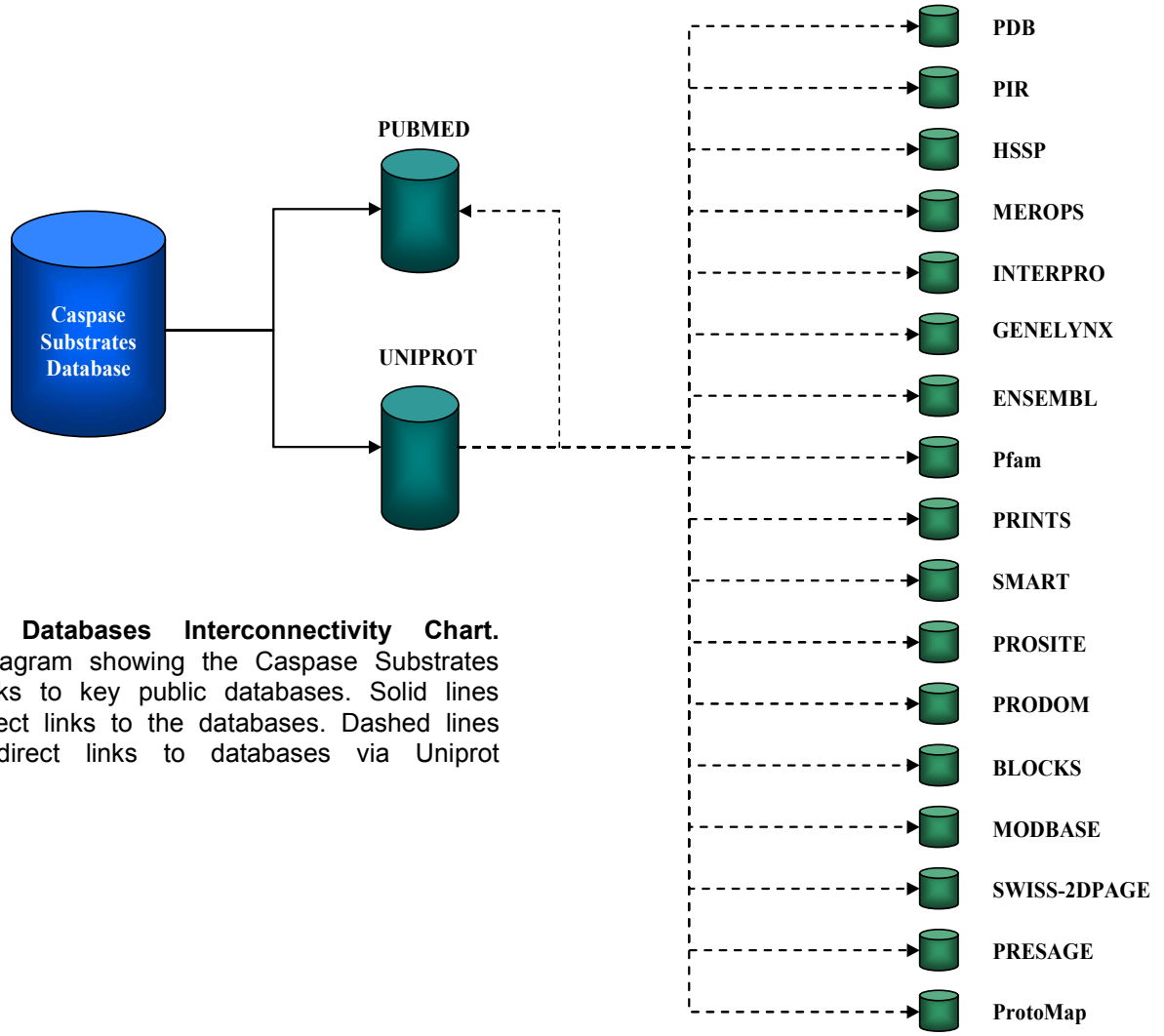


Figure 2-2 Databases Interconnectivity Chart. Schematic diagram showing the Caspase Substrates Database links to key public databases. Solid lines represent direct links to the databases. Dashed lines represent indirect links to databases via Uniprot database

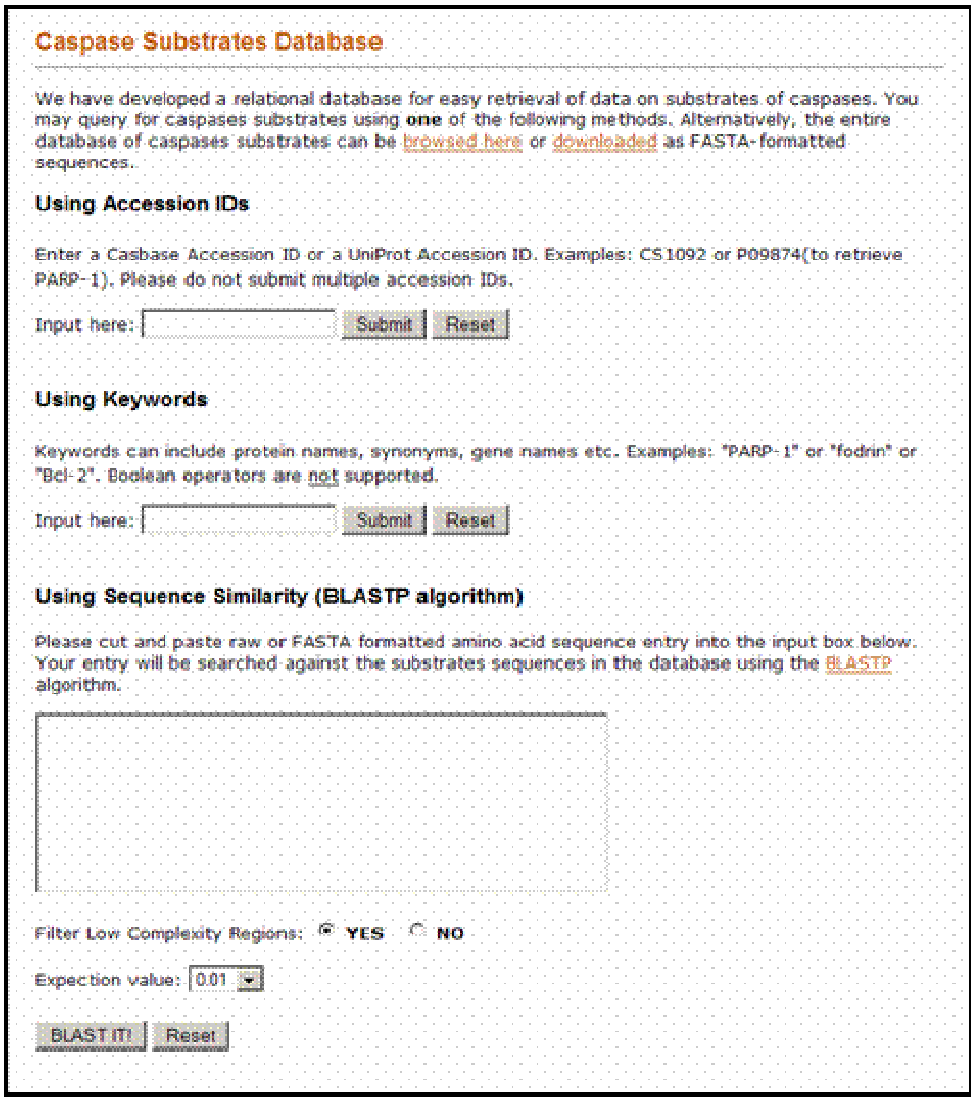
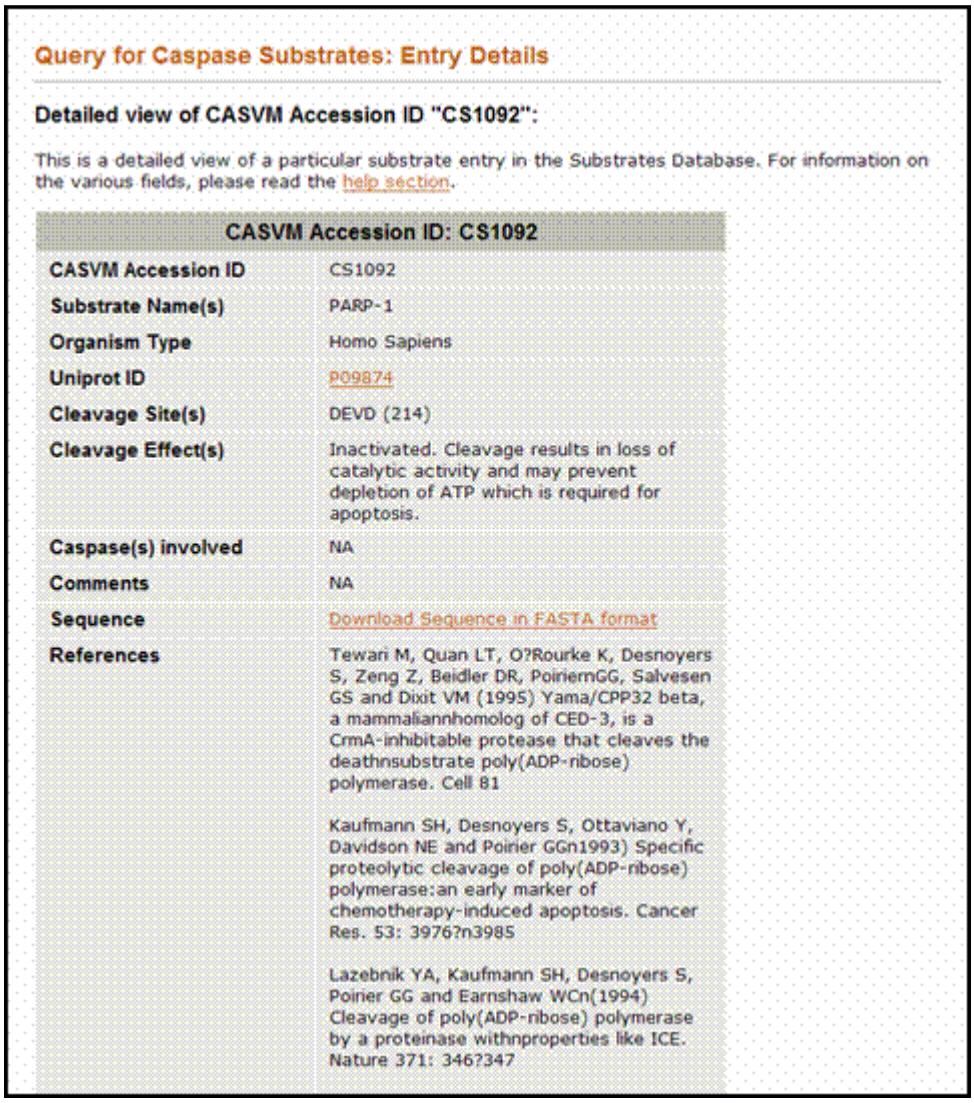


Figure 2-3 The Caspase Substrates Database Query Page. Page displays options for querying the database. Three search options are available: (i) using Casbase Accession ID or Uniprot Accession ID, (ii) using keywords such as protein names or gene names, and (iii) using similarity searches via BLASTP.



Query for Caspase Substrates: Entry Details

Detailed view of CASVM Accession ID "CS1092":

This is a detailed view of a particular substrate entry in the Substrates Database. For information on the various fields, please read the help section.

CASVM Accession ID: CS1092	
CASVM Accession ID	CS1092
Substrate Name(s)	PARP-1
Organism Type	Homo Sapiens
Uniprot ID	P09874
Cleavage Site(s)	DEVD (214)
Cleavage Effect(s)	Inactivated. Cleavage results in loss of catalytic activity and may prevent depletion of ATP which is required for apoptosis.
Caspase(s) involved	NA
Comments	NA
Sequence	Download Sequence in FASTA format
References	Tewari M, Quan LT, O?Rourke K, Desnoyers S, Zeng Z, Beidler DR, PoirierGG, Salvesen GS and Dixit VM (1995) Yama/CPP32 beta, a mammaliannhomolog of CED-3, is a CrmA-inhibitable protease that cleaves the deathnsubstrate poly(ADP-ribose) polymerase. Cell 81 Kaufmann SH, Desnoyers S, Ottaviano Y, Davidson NE and Poirier GGn1993) Specific proteolytic cleavage of poly(ADP-ribose) polymerase:an early marker of chemotherapy-induced apoptosis. Cancer Res. 53: 3976?n3985 Lazebnik YA, Kaufmann SH, Desnoyers S, Poirier GG and Earnshaw WCn(1994) Cleavage of poly(ADP-ribose) polymerase by a proteinase withnproperties like ICE. Nature 371: 346?347

Figure 2-4 The Caspase Substrates Database Details Page. Page displays the data on a database entry (e.g. PARP-1 protein). Page is populated with fields containing useful annotations on the substrate.

2.4 Conclusion

In this chapter, the problem of data for prediction of caspase substrates was discussed. Specifically, the lack of accurate and readily available data on bona fide caspase substrates is a limiting factor for studies on caspase-mediated substrate cleavage as well as the development of appropriate prediction tools. Consequently, a systematic approach was carried to extract the relevant data from published literature. A dataset of 313 experimentally verified caspase substrates (updated to 31 May 2008) was constructed after data retrieval. To facilitate efficient storage and retrieval of data, a relational database was development and is accessible at http://www.casbase.org/casvm/squery/index.html. The content of chapter has been published in **Wee LJ, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. Bioinformatics. 2007, 23:3241-3243.**

Chapter 3: Prediction of caspase cleavage sites

3.1 Introduction

As discussed in Chapter 1, substrates of caspases belong to a myriad of protein classes such as structural elements of cytoplasm and nucleus, components of the DNA repair machinery, protein kinases, GTPases and viral structural proteins. Although more than 280 caspase substrates have been discovered to date, it is possible that several more remain undetected. The identification and characterization of caspase substrates is critical for deepening our understanding of the role of these enzymes in the various cellular pathways. However, the accurate detection of caspase cleavage sites in target proteins requires complex and time consuming *in vivo* and *in vitro* experiments. Given the readily available sequence data in public databases, a useful alternative is to conduct *in silico* screening for potential cleavage sites among proteins. While the preferential cleavage specificities may be useful here, recently identified substrates have shown significant variation in their cleavage sites (Fischer *et al*., 2003). Therefore, the development of computational tools to accurately capture complex sequence patterns and to automate the identification of new cleavage sites would be valuable.

A number of caspase substrate cleavage prediction methods currently exist. All methods are based on the detection of caspase cleavage sites on potential substrates (primary features of these methods are summarized in Table 3-1). The pioneering work began with PeptideCutter, a proteases substrates cleavage prediction server for various families of proteases (Gasteiger *et al*., 2003). Due to the scarcity of experimental data, PeptideCutter was based only on the preferential cleavage specificities of certain caspases (Thornberry *et al*., 1997). Lohmuller *et al*. (2003)

developed the peptidase substrate prediction tool (PEPS) based on position specific scoring matrices (PSSM) for cathepsin B, cathepsin L and caspase-3 substrates. While useful, the utility of these tools is limited as they were built on a small dataset of cleavage sites and the cleavage specificities are confined to certain caspases alone, rather the entire family. In recent years, the exponential discovery and characterization of new substrates and their cleavage sites enabled the development of more effective algorithmic tools. Garay-Malpartida *et al*. (2005) developed the CasPredictor software which exhibited an improvement over previous methods with an accuracy of 81% on a dataset of 137 experimentally verified cleavage sites. The CasPredictor software uses an algorithm which analyzes the cleavage sites for amino acid substitution, amino acid frequency and the presence of 'PEST' sequences in the vicinity of the cleavage site (flanking 10-15 residues). The GraBCas software by Backes *et al*. (2005) advanced the previous PSSM-based methods by including an updated set of caspase cleavage specificities based on the work by Thornberry *et al*. (1997), and observing conservation at P_1' and even P_2' positions. Yang (2005) experimented with different neural networks for predicting cleavage sites such as single-layer perceptrons, multi-layer perceptrons and the Bayesian bio-basis function neural networks. They achieved an accuracy of 97% using the Bayesian bio-basis function neural network with two Gaussian distributions. In the same study, the support vector machines (SVM) method was tested and was found to give excellent results. However, Yang used a small dataset of 13 sequences and the method is not available for testing.

Table 3-1 Comparison of caspase cleavage sites prediction tools and algorithms

Tool	Algorithm	Dataset	Structural Features	Specificity	Accuracy	Availability
PeptideCutter	Consensus Motifs	NA	P_4 to P_1	Caspases 1-9	NA	Online
PEPS	Position Specific Scoring Matrices	11 sequences	P_4 to P_2'	Caspase-3 only	NA	By request from authors
CasPredictor	Position Specific Scoring Matrices	137 sequences	P_4 to P_1', PEST sequences	All	81%	By request from authors
GraBCas	Position Specific Scoring Matrices	NA	P_4 to P_1'	Caspases 1-9	87%	Online
BBFNN	Artificial Neural Networks	13 sequences	P_4 to P_1	All	96%	By request from authors

In this chapter, a support vector machine (SVM) system was developed to predict for potential caspase substrate cleavage sites on protein sequences. First introduced by Cortes and Vapnik (1995), the SVM method is a relatively new sub-branch of the machine learning algorithms. SVM has been shown to perform well in diverse computational biology applications such as the prediction of protein secondary structure (Hua and Sun, 2001; Ward *et al*., 2003; Nguyen and Rajapakse, 2005); protein fold (Ding and Dubchak, 2001); protein quaternary structure (Zhang *et al*., 2003); protein homology (Busuttil *et al*., 2004); protein-protein interaction sites (Bradford and Westhead, 2005); protein domains (Vlahovicek *et al*., 2005), HIV protease cleavage sites (Cai *et al*., 2002) and T-cell epitopes (Zhao *et al*., 2003). It is also used in the classification and validation of cancer tissue samples (Furey *et al*., 2000) and microarray expression data (Brown *et al*., 2000). Other applications of SVMs in biology have been reviewed by Byvatov and Schneider (2003), and Yang (2004). An extensive dataset of unique (non-redundant) cleavage sites extracted from the datasets of experimentally verified caspase substrates (from Fischer and Post-Fischer datasets) was constructed to validate the SVM method and to further the development of other computational tools. Using various statistical metrics, the SVM method was shown to be a rigorous and effective approach for predicting cleavage sites of caspase substrates.

3.2 Results and Discussion

A cleavage sites dataset containing 195 and 24 unique caspase cleavage site sequences were constructed from the Fischer and Post-Fischer datasets respectively. The 195 sequences were used for training the SVM classifier while the 24 sequences were used for testing the effectiveness of the SVM method. As there were no experimentally reported non-cleavage sites for caspases, tetrapeptide sequences were extracted at random positions (not including the cleavage sites) on experimentally determined caspase substrates. One non-cleavage site was extracted for every cleavage site on the same substrate. The assumption that an intuitively large proportion of tetrapeptide sequences other than the cleavage sites(s) on the same substrate should not be recognized and cleaved by caspases justify the use of these sequences as non-cleavage sites. An equal number of these non-cleavage sites were extracted to match the cleavage sites. Together, a primary dataset consisting of the tetrapeptide cleavage sites (positive sequences) and non-cleavage sites (negative sequences) was constructed and designated as the P_4P_1 dataset (Figure 3-1).

Previously, Backes *et al*. (2005) and Garay-Malpartida *et al*. (2005) suggested that residues adjacent to the cleavage site may influence substrate cleavage. Backes *et al.* reported the high occurrence of specific amino acids at P_1' for caspase-3 and P_1' and P_2' for granzyme B, a serine protease involved in apoptosis and in immune response (Lord *et al*., 2003). Garay-Malpartida *et al*. reported that a sizeable proportion of cleavage sites are localized within 'PEST' regions, of which have been suggested to label proteins for protease degradation. PEST regions are defined as sequence segments enriched with

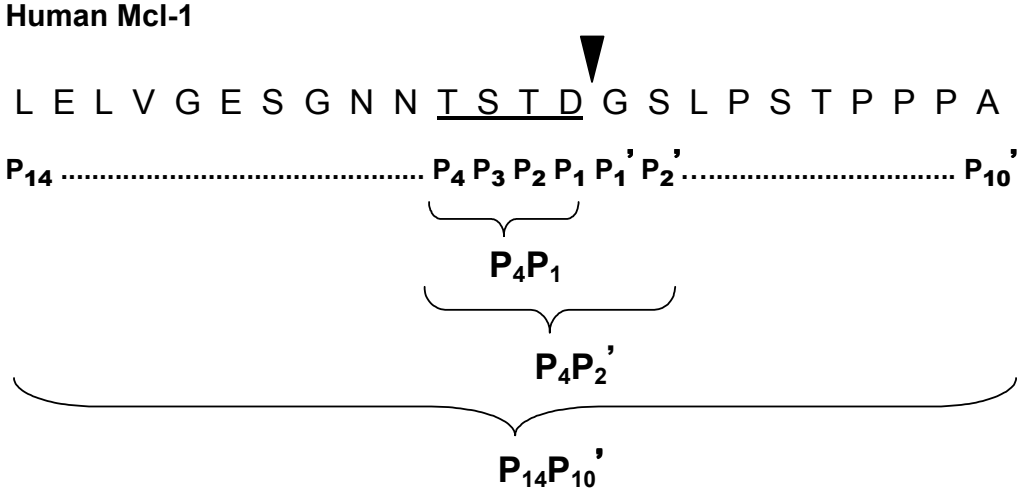


Figure 3-1 Different subsequence segments for SVM training and testing. For human Mcl-1 [Swiss-Prot: Q07820], a sequence window of 24 amino acids in length centred on the tetrapeptide cleavage site, TSTD (underlined) is shown. Amino acids to the left of the scissile bond (shown by the inverted triangle) are labelled from P_1 (D) to P_{14} (L). Amino acids to the right of the scissile bond are labelled from P_1' (G) to P_{10}' (A). Curly brackets indicate the subsequence segments extracted for SVM implementation. The sequences spanning P_4 to P_1 (TSTD), P_4 to P_2' (TSTDGS) and P_{14} to P_{10}' (LELVGEGSNNTSTDGSLPSTPPPA) are labelled as P_4P_1, P_4P_2' and $P_{14}P_{10}$' respectively.

proline (P), glutamate (E), aspartate (D), serine (S) and threonine (T) residues (Rogers *et al*., 1986; Rechsteiner and Rogers, 1996). Therefore, to investigate the influence of the adjacent sequences on substrate cleavage, a dataset containing tetrapeptide sequences with the P_1' and P_2' residues and a dataset containing tetrapeptide sequences flanked by ten residues on either side of the cleavage site were constructed. These datasets were designated as P_4P_2' and $P_{14}P_{14}$' respectively (Figure 3-1). The longer sequence segments would encapsulate the information contained in the critical tetrapeptide sequences as well as the P_1' and P_2' amino acids and other residues adjacent to the cleavage sites.

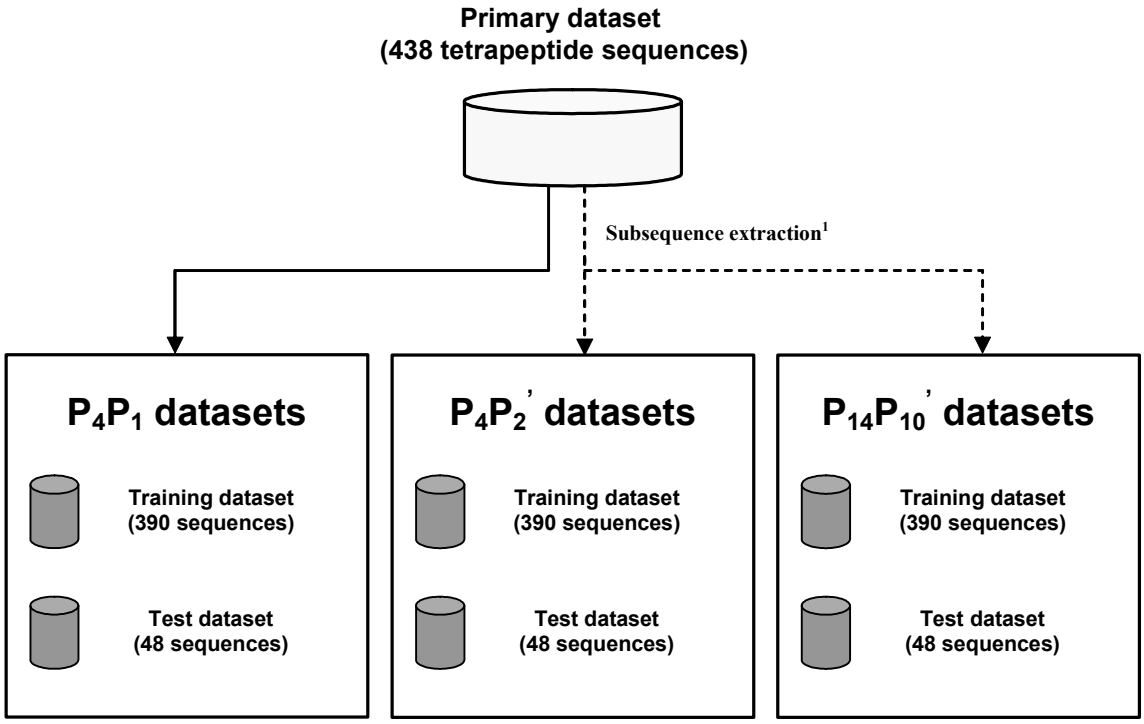


Figure 3-2 Schematic layout of the datasets used for SVM training and testing. The primary dataset consist of non-redundant tetrapeptide caspase substrate cleavage sites obtained from literature and an equal number of non-cleavage sites. [1]The P_4P_1 sequences consist of all the sequences in the primary tetrapeptide cleavage site dataset. P_4P_2' and $P_{14}P_{10}$' datasets were derived by extracting subsequence segments from the parent protein chains in the vicinity of the tetrapeptide cleavage sites, as shown in Figure 3-1. All datasets contain equal number of positive and negative examples.

Next, the P_4P_1, P_4P_2' and $P_{14}P_{14}$' datasets were divided into training and test datasets (Figure 3-2). The training datasets were used for optimizing the SVM parameters and for training the SVM classifier, while the test datasets were used for evaluating the SVM method. The RBF kernel, with parameters γ and C, was chosen for SVM implementation. Using 10-fold cross-validation, the parameters γ and C were optimized at 0.01 and 100 (for P_4P_1 training dataset) and 0.1 and 100 (for both P_4P_2' and $P_{14}P_{10}$' training datasets). For each of P_4P_1, P_4P_2' and $P_{14}P_{10}$' training datasets, an overall accuracy of 98.97% was obtained during the cross-validation.

While the reported accuracy on the training datasets may indicate the effectiveness of a prediction method, it may not accurately portray how the method will perform on novel, hitherto undiscovered cleavage sites. Therefore, testing the SVM methodology on independent out-of-sample datasets not used in the cross-validation is critical. The SVM classifiers, trained separately using the entire training datasets from the P_4P_1, P_4P_2' and $P_{14}P_{10}$' datasets with the optimized γ and C parameters, were applied on the respective test datasets and results were evaluated. As shown in Table 3-2, for the P_4P_1 test dataset, the SVM method obtained an accuracy of 95.83% using the RBF kernel with γ=0.01 and C=100. For both the P_4P_2' and $P_{14}P_{10}$' test datasets, the SVM method obtained an accuracy of 97.92% using the RBF kernel with γ=0.1 and C=100.

An analysis on the training and test datasets indicated a large percentage of cleavage sites with the XXXD motif (~98%) and a very small percentage of cleavage sites with a non-canonical XXXE motif (~2%). While experimental cleavage site specificities reported in Thornberry *et al*. (2007) suggest most, if not all, sequences to conform to the XXXD motif, the inclusion of a large proportion of these sequences in the development of the SVM system could lead to over-training of the classifier and confound the results obtained with different sequence representations. To mitigate this possibility, datasets identical to P_4P_1, P_4P_2' and $P_{14}P_{10}$' datasets but with the P_1 residue removed in all the sequences (labeled as P_4P_1(-D), P_4P_2'(-D) and $P_{14}P_{10}$'(-D) datasets respectively) were constructed. These datasets were further divided into training and test sets and SVM parameters were optimized in the manner as reported for the original P_4P_1,

Table 3-2 Results of SVM prediction for various test datasets

Test datasets	γ[1]	C[1]	Performance Evaluation			
			AC (%)	SE (%)	SP (%)	MCC
P_4P_1	0.01	100	95.83	95.83	95.83	0.92
P_4P_2'	0.1	100	97.92	95.83	100.00	0.96
$P_{14}P_{10}'$	0.1	100	97.92	95.83	100.00	0.96
P_4P_1(-D)	0.01	1	81.25	62.50	100.00	0.67
P_4P_2'(-D)	1	100	89.58	79.17	100.00	0.81
$P_{14}P_{10}'$(-D)	0.1	1	93.75	87.50	100.00	0.88

1. The SVM parameters (γ and C) were obtained from the cross-validation conducted on the training datasets.

Table 3-3 GraBCas prediction on the P_4P_1 training dataset

GraBCas Cut-off	SE (%)
0.1	87.43
1.0	69.46
5.0	40.72
10.0	28.14
20.0	19.76

P_4P_2' and $P_{14}P_{10}$' datasets. The trained SVM classifiers were tested on the respective test datasets. As shown in Table 3-2, the SVM method obtained an accuracy of 81.25% for the P_4P_1(-D) test dataset. The performance of the SVM improved significantly when tested on P_4P_2'(-D) and $P_{14}P_{10}$'(-D) datasets as accuracy readings of 89.58% and 93.75% were obtained respectively. While the accuracy on all (-D) test datasets were lower compared to the corresponding original datasets, a larger degree of improvement was observed when the longer sequence representations were used, as evidenced by the greater spread in both the accuracy and sensitivity readings for the P_4P_1(-D), P_4P_2'(-D) and $P_{14}P_{10}$'(-D) datasets. An analysis of the misclassified sequences showed that cleavage sites such as $CLLD^{2193}$ from Notch1 [Swiss-Prot:P46531] and $PEVD^{142}$ from p23 co-chaperone [Swiss-Prot:Q15185], which differ markedly from reported tetrapeptide specificities, were misclassified by the P_4P_1(-D)-trained SVM, but were correctly predicted when the P_4P_2'(-D) and P14P10'(-D) datasets were used. Also, the SVM trained with the P_4P_1(-D) and P_4P_2'(-D) datasets failed to correctly classify the non-canonical cleavage site $VQPE^{205}$ from DIAP1 [Swiss-Prot:Q24306], but correctly predicted the cleavage site when trained with the $P_{14}P_{10}$'(-D) dataset. These results suggest that the SVM trained with the (-D) datasets may be useful for identifying hitherto undiscovered cleavage sites while circumventing the problem of overtraining due to the high percentage of "XXXD" cleavage sites in the training datasets. The results also provided further evidence for the suggestion that the P_1', P_2' and residues further upstream and downstream of the cleavage site may influence substrate cleavage, and by accounting for these flanking sequences, the SVM performance can be improved. It was also shown that the SVM method can be extended to predict cleavage sites with residues other than the

canonical aspartate (D) at P_1. While the occurrence of the non-canonical cleavage sites remains proportionately small, it does imply that the sampling space is not limited to the XXXD motif for cleavage sites. Consequently, the ability to predict these non-canonical cleavage sites will be a useful complement to existing computational methods which assumes the consensus XXXD motif as the basis for their algorithms.

As other methods were not readily accessible, only GraBCas could be used for direct comparison with the SVM method using the present datasets. Since the GraBCas method primarily focuses on the tetrapeptide motif, it was only applied to the P_4P_1 training dataset. As the GraBCas method can only be applied to potential cleavage sites with aspartate (D) at the P_1 position, the positive sequences in the P_4P_1 training dataset were scored with the GraBCas matrix values for the different caspases then selected for the highest score and checked for the percentage of correctly predicted cleavage sites (or Sensitivity, SE) against a series of cut-off scores. As shown in Table 3-3, the sensitivity values declined steadily from 87.43% to 19.76% as the cutoff values were progressively increased (0.1, 1, 5, 10, 20). The GraBCas method was also tested on the positive sequences in the P_4P_1 test dataset. As there were no recommended cut-off scores for predicting the cleavage sites, the cut-off score of 0.1, as used for the granzyme B cleavage sites prediction in Backes *et al*. (2005), was chosen. At the cut-off score of 0.1, GraBCas predicted only 16 out of 24 cleavage sites correctly (SE = 66.67%).

Finally, to investigate how the SVM method can complement experimental work on caspase substrate cleavage, the SVM method was used to predict the caspase-mediated cleavage of an anti-apoptotic protein, Livin [Swiss-Prot: Q96CA5] and its mutant

Table 3-4 SVM prediction of caspase substrate cleavage sites in Livin and mutants

Substrate[1]	Experimental Results[2]	SVM Prediction[3]
Wild type Livin	Cleaved	Cleaved
LE Δ52-61	Not Cleaved	Cleaved
Δ53-55	Cleaved	Cleaved
Δ55-57	Cleaved	Cleaved
Δ57-59	Cleaved	Cleaved
Δ60-62	Cleaved	Cleaved
Δ52-61	Not Cleaved	Not Cleaved
Δ53-61	Not Cleaved	Cleaved
Δ52	Not Cleaved	Not Cleaved
Δ51-53	Not Cleaved	Cleaved

1. Wild type Livin and various deletion mutants as reported in Yan et al.
2. Experimentally verified cleavage (cleaved) or non-cleavage (not cleaved) of Livin and deletion mutants.
3. SVM prediction of caspase cleavage sites on Livin and deletion mutants (Cleaved – presence of cleavage site; Not Cleaved –absence of cleavage site).

sequences as reported in Yan *et al*. (2006), based on the prediction of the caspase cleavage sites. As shown in Table 3-4, the experimental cleavage of wild type human Livin and its deletion mutants were compared to the results predicted by the SVM trained with the $P_{14}P_{10}$'(-D) dataset. With the exception of the LE Δ52-61, Δ51-53 and Δ53-61 mutants, all other sequences were correctly predicted to be cleaved or not cleaved by caspases as indicated. For the LE Δ52-61 and Δ51-53 mutants, the flanking sequences upstream and downstream of the cleavage site were likely to have influenced cleavage of the substrates, as predicted by the SVM. However, cleavage of substrates was prevented due to the absence of the Asp at P_1 ($DHVD^{52}$). While the

SVM predicted the cleavage of Δ53-61 mutant, it was proposed by Yan *et al*. (2006) that the deleted residues might have led to the distortion of the structure of a neighboring domain or affected its signaling function, which subsequently inhibited the substrate cleavage through downstream signaling. These findings suggest that the SVM-based prediction of caspase substrate cleavage sites might be helpful in identifying potential caspase substrates.

3.3 Methods

3.3.1 Datasets

The primary dataset contains 438 unique sequences (219 cleavage sites and 219 non-cleavage sites). Of the 219 cleavage sites, 195 were obtained from the Fischer dataset and 24 from the Post-Fischer dataset. Besides the tetrapeptide cleavage site sequences, subsequence segments of varying lengths centred on the tetrapeptide cleavage sites were extracted as shown in Figure 3-1. In total, three groups of sequences were obtained: tetrapeptide cleavage sequences (henceforth termed as the P_4P_1 sequences), tetrapeptide cleavage sequences with the next two residues, P_1' and P_2' residues (P_4P_2'sequences), and tetrapeptide sequences with upstream residues up to P_{14} and downstream residues up to P_{10}' ($P_{14}P_{10}$' sequences). The cleavage sites and the corresponding subsequences were designated as positive sequences for the SVM training and testing.

The 219 non-cleavage sites were obtained by extracting tetrapeptide sequences at random positions (not including the cleavage sites) on caspase substrates. One non-cleavage site was extracted for every cleavage site on the same substrate. Subsequence segments centred on these non-cleavage sites were also extracted in the

manner reported earlier. The non-cleavage sites and the corresponding subsequences were designated as positive sequences for SVM training and testing. Together, the positive and negative sequences in the three group of sequences were designated as the P_4P_1, P_4P_2' and $P_{14}P_{10}$' datasets respectively. Each of these datasets was further divided in the following manner (Figure 3-2):

Training datasets: Training datasets were used for optimizing the SVM parameters and for training the SVM classifier to predict unseen test examples. Each training dataset contain 390 sequences (195 positives and 195 negatives). The positive sequences were obtained from the Fischer dataset and are available in Appendix B (Table B-1).

Test datasets: Test datasets were used for evaluating the performance of the SVM method. Each test dataset contains 48 sequences (24 positives and 24 negatives). The positive sequences were obtained from the Post-Fischer dataset and are available in Appendix B (Table B-2).

Datasets containing sequences identical to the P_4P_1, P_4P_2' and $P_{14}P_{10}$' datasets but without the P_1 residue were also constructed (designated as P_4P_1(-D), P_4P_2'(-D) and $P_{14}P_{10}$'(-D) respectively). These datasets were divided into training and test datasets as mentioned earlier.

3.3.2 Vector encoding schemes

To encapsulate the sequence information into a format suitable for SVM training and testing, the sequences were transformed into *n*-dimensional vectors using an orthonormal encoding scheme. Each amino acid is represented by a 20-dimensional vector, composed of either zero or one as elements. For example, alanine was represented as [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1] and cysteine as

[0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0]. Therefore, for the P_4P_1 dataset, each sequence was represented by an 80-dimensional vector. Sequences in the P_4P_2' and $P_{14}P_{10}$' datasets were represented by 120 and 480 dimensional vectors respectively.

3.3.3 SVM implementation

For SVM implementation, the freely downloadable LIBSVM package by Chang and Lin (2001) was used. Details of the SVM methodology can be obtained from the article by Burges (1998). Briefly, SVM is based on the structural risk minimization principle from statistical learning theory. A set of positively and negatively examples can be represented by the feature vectors x_i (i = 1, 2,…, N) with corresponding labels $y_i \in \{+1,-1\}$. To classify the data, the SVM trains a classifier by mapping the input samples, using a kernel function in most cases, onto a high-dimensional space, and then seeking a separating hyperplane that differentiates the two classes with maximal margin and minimal error. The decision function for new predictions on unseen examples is given as:

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i . x_j) + b\right)$$

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i . x_j) + b\right)$$

where $K(x_i . x_j)$ is the kernel function, and the parameters are determined by maximizing the following:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, \alpha_j \, y_i \, y_j K(x_i . x_j)$$

under the conditions,

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \text{ and } 0 \le \alpha_i \le C$$

The variable *C* serves as the regularization parameter that controls the trade-off between margin and classification error.

As the effectiveness of the SVM prediction system is dependent on the type of kernel used, various kernels (linear, sigmoid, polynomial and the radial basis function) commonly implemented in biological problems was explored for this problem. Each of the kernel functions were tested on the training datasets and the widely used radial basis function (RBF) kernel as it was found to be most effective (data not shown). The RBF kernel function is given as:

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\gamma^2}\right)$$

With the RBF kernel, two parameters are required for optimizing the SVM classifier; γ, which determines the capacity of the RBF kernel and the regularization parameter *C*.

3.3.4 SVM optimization

To optimize the SVM parameters γ and *C*, 10-fold cross-validation was applied on each of the training datasets using various combinations of γ and *C*. In 10-fold cross-validation, the training dataset was spilt into 10 subsets where one of the subsets was used as the test set while the other subsets were used for training the classifier. The trained classifier was tested using the test set. The process is repeated 10 times using a different subset for testing, hence ensuring that all subsets are used for both training and testing. SVM parameters γ and *C* were stepped through combinations of 0.01, 0.1, 1, 10, 100 for γ, and 1, 10, 100 and 1000 for *C* in a grid-based manner.

3.3.5 SVM training and testing

The best combinations of γ and *C* obtained from the optimization process were used for training the SVM classifier using the entire training dataset. The SVM classifier was subsequently used to predict the test datasets. Various quantitative variables were obtained to measure the effectiveness of the SVM method:

(i) *TP*, true positives – the number of correctly classified cleavage sites.

(ii) *FP*, false positives – the number of incorrectly classified non-cleavage sites.

(iii) *TN*, true negatives – the number of correctly classified non-cleavage sites.

(iv) *FN*, false negatives – the number of incorrectly classified cleavage sites.

Using the variables above, a series of statistical metrics were computed to measure the effectiveness of the SVM method. *Sensitivity (SE)* and *Specificity (SP)*, which indicates the ability of the prediction system to correctly classify the cleavage and non-cleavage sites respectively, were calculated:

$$SE\ (\%) = \frac{TP}{TP + FN}\ x\ 100$$

$$SP\ (\%) = \frac{TN}{TN + FP}\ x\ 100$$

To provide an indication of the overall performance of the system, we computed *Accuracy (AC),* for the percentage of correctly classified sites, and the *Matthews Correlation Coefficient (MCC)*.

$$AC\ (\%) = \frac{TP + TN}{TP + FN + TN + FP}\ x\ 100$$

$$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3.3.6 Prediction of caspase cleavage of Livin and mutants

The SVM trained using the $P_{14}P_{10}{}'$ (-D) dataset (using RBF kernel, γ =0.1, *C*=100) was used to predict the cleavage of Livin [Swiss-Prot:Q96CA5] and the various deletion mutants, based on the prediction of the caspase cleavage sites, as reported in Yan *et al*. (2006). 24 amino acids subsequence segments centred on the P_1 residue of the reported Livin cleavage site ($DHVD^{52}$) were extracted from both wild type and mutant Livin sequences. Mutants used in this study are: LE Δ52-61, Δ53-55, Δ55-57, Δ57-59, Δ60-62, Δ52-61, Δ53-61, Δ52 and Δ51-53. In mutants with Asp-52 deleted, the peptide windows were centred on the subsequent residue occupying position 52.

3.3.7 Comparison with other available methods

As the CasPredictor method is unavailable from the published website, it was not tested. The performance of GrabCas was compared with the SVM method using the current datasets. As the GraBCas scoring matrices are specific for the tripeptide, P_4-P_3-P_2, and assume that P_1 is an Asp (D) residue, the GraBCas matrices were used to score only the positive sequences (cleavage sites) from the P_4P_1 training dataset. As GraBCas scores for different caspases were available, only the highest scores were recorded. The percentage of correctly predicted cleavage sites *(Sensitivity, SE)* were calculated as mentioned earlier. GraBCas was further tested on the P_4P_1 test dataset in a similar manner and the *SE* score was obtained at a GraBCas cut-off of 0.1.

3.4 CASVM: Server for SVM prediction of caspase cleavage sites

3.4.1 Server description

A web server, CASVM, was developed for the SVM-based prediction of caspase substrates cleavage sites as discussed earlier. It is written in Perl and is hosted on a Linux platform accessible at http://www.casbase.org/casvm/server/index.html.

The server homepage presents an intuitive interface for user input and processing (Figure 3-3). Users can submit (through copy and paste) a raw or FASTA-formatted protein sequence and select a number of options for server prediction. Upon form submission, the input sequence will be scanned over the entire length of the sequence with the scanning window selected by the user. Three scanning window sizes are available: P_4P_1, P_4P_2' and $P_{14}P_{10}'$, each dictating the type of SVM classifier to be used for prediction. For example, if the scanning window size of $P_{14}P_{10}'$ is selected, the $P_{14}P_{10}'$-trained SVM classifier will be used for prediction. The $P_{14}P_{10}'$-trained classifier, having reported the highest accuracy during our experimentation, is selected as default. There is also the option for the selection of the type of P_1 residue to be screened so as to account for the possibility of non-canonical cleavage sites on substrates. Users are able to select for aspartic acid (default) or both aspartic acid and glutamic acid as the required P_1 residue. As the input sequence is being scanned, sequence segments containing the specified P_1 amino acid will be extracted and predicted for the presence of the cleavage site with the selected SVM classifier.

The output of the CASVM prediction displays the name of input sequence (optional), sequence length, an abbreviated version of the sequence, a list of potential cleavage sites (all tetrapeptide sequences with the specified P_1 residue in the input sequence) and the CASVM-predicted cleavage sites. All cleavage sites are labelled with the P_1 residue position (Figure 3-4).

3.4.2 Discussion

An analysis of the cleavage sites datasets (available in Appendix B: Table B-1 and Table B-2) revealed that a large number of caspase cleavage sites differ markedly from the consensus tetrapeptide specificities. More significantly, although caspases are thought to be selective for aspartic acid at the P_1 position, a notable number of substrates were cleaved at tetrapeptide sites bearing glutamic acid at the P_1 position. Interestingly, existing methods for caspase cleavage sites prediction (excepting the method herein) are largely limited to the discovery of cleavage sites with aspartic acid at P_1 as they assume the consensus XXXD motif as the basis for their algorithms. The strict adherence to the inclusion of aspartic acid may be limiting the sensitivity of these tools since it is intuitively likely that many more substrates will have P_1 residues as glutamic acid. Therefore, the option on the server to screen and predict for tetrapeptide sequences containing either aspartic acid or glutamic acid at the P_1 position would be helpful. In addition, as the substrates used in the method are derived from a variety of organisms (human, mouse, rat, fruit fly, cow, chicken, frog, worm and viruses) and are cleaved by various caspases (caspase -1, -3, -6, -7, -8, -9, -12, -13 and -14), the server is applicable to the detection of cleavage sites in substrates from various organisms and is not caspase specific.

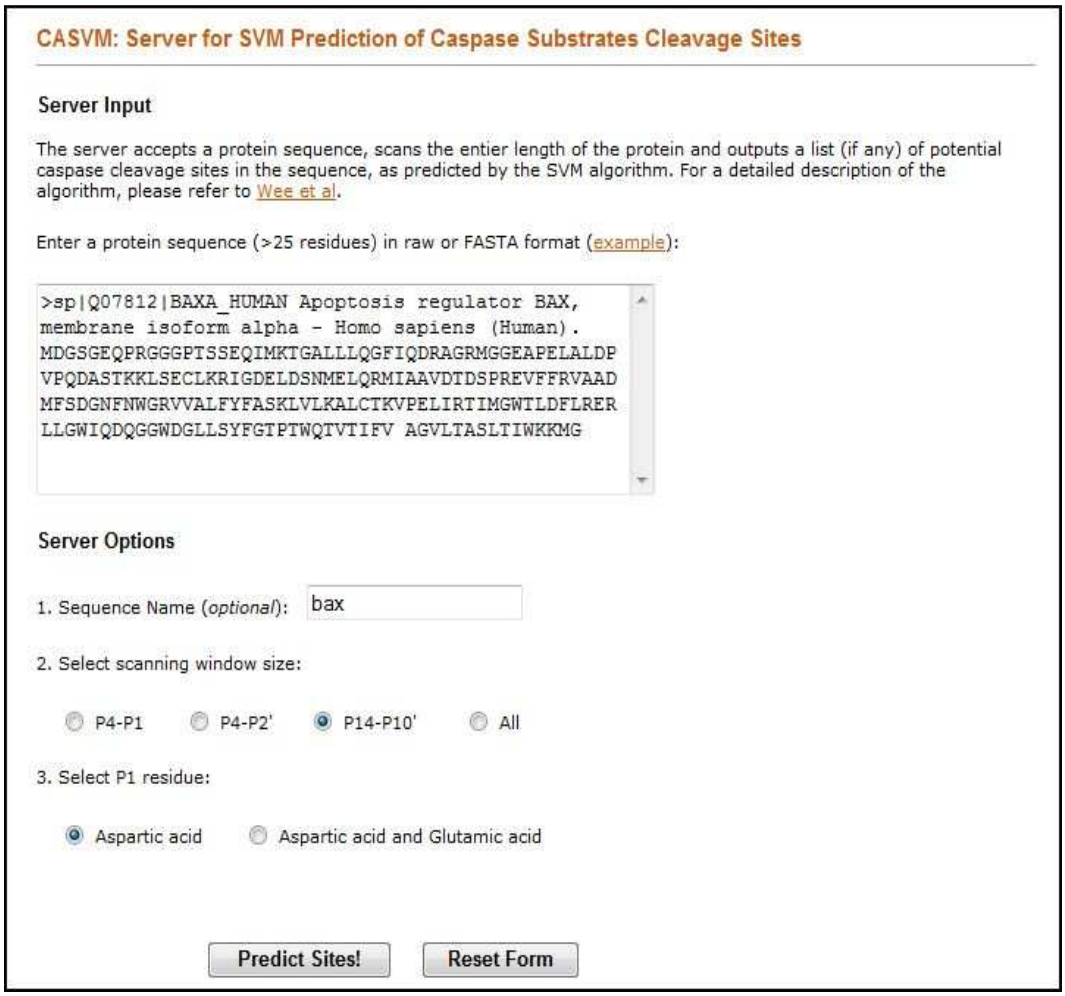
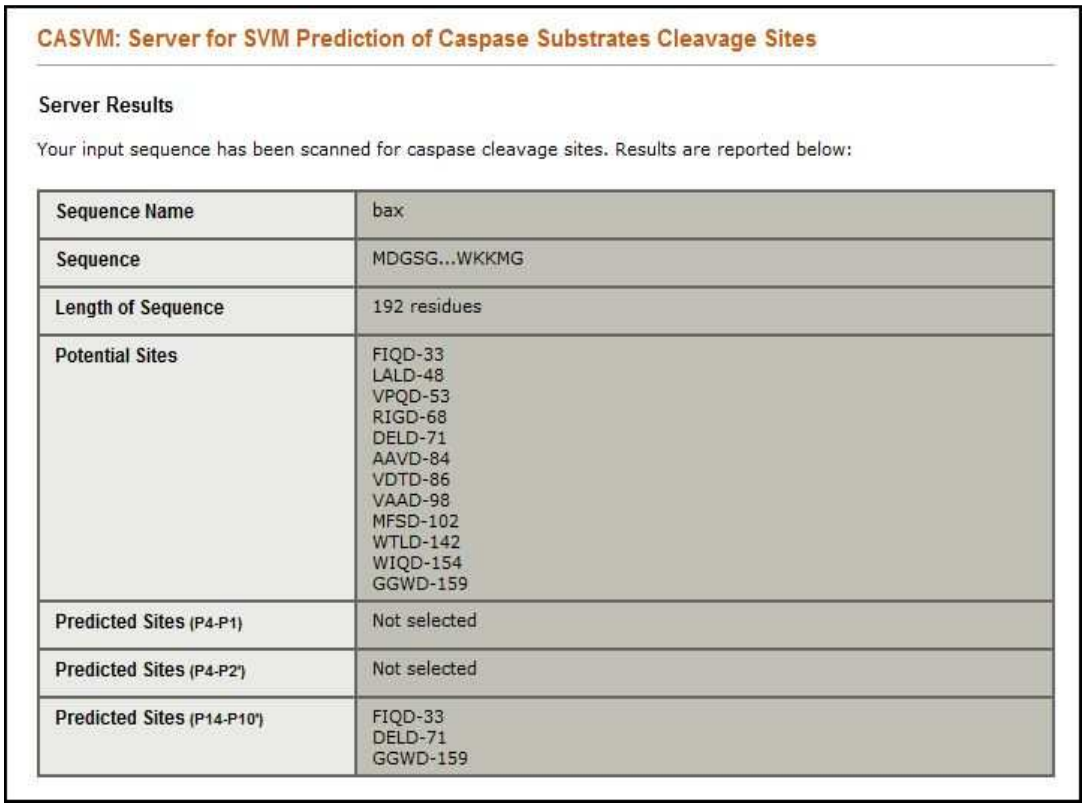


Figure 3-3 CASVM server page. A web form allows users to submit a raw or FASTA-formatted protein sequence to predict for caspase cleavage sites using various options.



CASVM: Server for SVM Prediction of Caspase Substrates Cleavage Sites

Server Results

Your input sequence has been scanned for caspase cleavage sites. Results are reported below:

Sequence Name	bax
Sequence	MDGSG...WKKMG
Length of Sequence	192 residues
Potential Sites	FIQD-33 LALD-48 VPQD-53 RIGD-68 DELD-71 AAVD-84 VDTD-86 VAAD-98 MFSD-102 WTLD-142 WIQD-154 GGWD-159
Predicted Sites (P4-P1)	Not selected
Predicted Sites (P4-P2')	Not selected
Predicted Sites (P14-P10')	FIQD-33 DELD-71 GGWD-159

Figure 3-4 The results of prediction on CASVM server. A list of all predicted cleavage sites on the sequence is listed with the position of the P_1 residue annotated.

3.5 Conclusion

In conclusion, a support vector machines (SVM) approach for predicting caspase cleavage sites was developed and was shown to be complementary to existing methods, if not more effective. The prediction accuracy can also be improved by accounting for sequences at the P_1' and P_2' positions and further upstream and downstream of the cleavage site. In addition, the SVM method may be useful for predicting the non-canonical cleavage sites lacking aspartic acid the P_1 position, such as those found in DIAP1 and other proteins as reported in literature. CASVM, a web server implementing the SVM method was developed. Also, an extensive dataset of unique, experimentally-verified caspase substrates cleavage sites was constructed and is made available on the CASVM server website. The content of this chapter has been published in **Wee LJ, Tan TW, Ranganathan S. SVM-based prediction of caspase substrate cleavage sites. BMC Bioinformatics. 2006, Suppl. 5:S14** and **Wee LJ, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. Bioinformatics. 2007, 23:3241-3243.**

Chapter 4: Towards the prediction of caspase substrates

4.1 Introduction

As discussed in Chapter 3, current approaches for caspase substrates prediction are based on the detection of potential cleavage sites on proteins using information encoded within the tetrapeptide motifs. While the identification of the specific cleavage site on the primary sequence of a protein is necessary for substrate prediction, it is intuitive that the final proteolytic cleavage of a protein is likely the outcome of a multitude of factors in addition to the presence of cleavage sites (Timmer and Salvesen, 2007). There is evidence to suggest that conformation of the local structure of the cleavage site and not just the primary sequence alone is required for protease cleavage. An analysis on a dataset of 176 caspase substrates (derived from the Fischer dataset), 80% of substrates contain at least one other caspase cleavage site which is not reported as the true cleavage site in literature. The cleavage sites of Tpr (DDED) [Ferrando-May *et al*., 2001), p28BAP31 (AAVD) [Ng *et al*., 1997; Granville *et al*., 1998; Maatta *et al*., 2000], golgin 160 (SEVD) [Mancini *et al*., 2000], Topo I (PEDD) [Casiano *et al*., 1998; Samejima *et al*., 1999] and heterogeneous nuclear ribonucleoparticle C1/C2 (GEDD) [Waterhouse *et al*., 1996], are located at two distinct positions on the respective protein but only one was reported to be cleaved. Recent studies have suggested that unstructured regions of substrates appear to be more susceptible to cleavage than regions of secondary structure (such as helices and β-sheets) (Timmer and Salvesen, 2007). Indeed, a number of *in vivo* caspase substrates with reported structures suggest that cleavage sites have a preference for disordered or unstructured extended loops, in line with observations on protease substrates in general. It is further suggested that the location

of cleavage sites may be important as well –a potential cleavage site needs to be located at the surface of the substrate, rather than within the hydrophobic core of the protein, in order to be accessible to the protease active site (Boyd *et al*., 2005).

In this context, it is opportune to quantify the relationship between factors for caspase substrate cleavage and to improve on existing methods for substrate prediction. In this chapter, studies conducted indicate that caspase cleavage sites have a higher propensity to be located in disordered or unstructured extended loops and in solvent exposed regions compared to non-cleavage sites. A quantitative model combining these factors with the previously developed SVM-based method for predicting caspase cleavage sites was developed to refine the prediction of caspase substrates. The multi-factor model was shown to improve accuracy by reducing the false positives from predicted cleavage sites based on different methods.

4.2 Materials and Methods

4.2.1 Dataset

74 unique, experimentally verified cleavage sites were obtained from the dataset of caspase cleavage sites derived from Fischer *et al*. (2003). 24-residue long subsequences comprising of the tetrapeptide sequence with flanking upstream 10 residues and downstream 10 residues ($P_{14}....P_4P_3P_2P_1.....P_{10}'$) were extracted and assigned as "cleavage sites". For every cleavage site, one other tetrapeptide was randomly selected on the respective substrate and subsequences similar in length to the cleavage site sequences were constructed. A total of 74 additional subsequences were constructed and designated "non-cleavage sites". Together, the sequences constitute the analysis dataset (total 148) and were used for analysis of structural features and for optimization of the substrate prediction model parameters.

4.2.2 Quantitative measures of secondary structures and solvent accessibilities

Each sequence in the analysis dataset was predicted for secondary structures and solvent accessibilities using the SABLE II protein structure prediction server (w/approximator predictor selected, server located at http://sable.cchmc.org) (Wagner *et al*., 2005; Adamczak *et al*., 2004; Adamczak *et al*., 2005). SABLE server output was parsed with Perl scripts and quantitative measures of the propensities for helices (H_p), b-strands (E_p), coils (C_p) and solvent accessibilities (S_p) were computed using the following equations:

$$H_p = \frac{\sum_1^N H_n}{N}$$

$$E_p = \frac{\sum_1^N E_n}{N}$$

$$C_p = \frac{\sum_1^N C_n}{N}$$

$$S_p = \frac{\sum_1^N S_n}{S_{max}}$$

Where H_n, E_p, C_p and S_n are the predicted secondary structure for helix, beta-strand, coil and real-value score (ranging 0-6, 0 for fully buried and 6 for maximum exposure) for solvent accessibility for each residue at position n in the sequence of

length *N* (=24). S_{max} is given a value of 144 (=24 x 6), which is the sum of real-value scores for all residues in the sequence assuming that each residue is maximally exposed to solvent.

The variable, *P-score*, which is a combinatorial measure of the propensity for unstructured regions and solvent exposure, was calculated for all sequences in the analysis and test datasets as:

$$P\text{-}score = \alpha\left(\frac{C_p}{2}\right) + \beta\left(\frac{S_p}{2}\right)$$

The coefficients α and β represent the respective fractional weighting of C_p and S_p in the *P-score* variable. Optimal α and β coefficient values were obtained by stepping through various combinations of values (0.0, 0.1, 0.2…1.0), and measuring the percentage of cleavage sites and non-cleavage sites retained at increasing P-score cut-offs.

4.2.3 Multi-factor model testing

For model testing, a test dataset of unique caspase cleavage sites (17 sequences) from the Post-Fischer dataset was used. The test dataset was predicted for potential caspase cleavage sites using CASVM server and GraBCas algorithm using default options. Predicted caspase cleavage sites were extracted from substrate sequences and 24-residue long subsequences comprising of the predicted tetrapeptide sequences with flanking upstream 10 residues and downstream 10 residues ($P_{14}....P_4P_3P_2P_1.....P_{10}'$) were constructed and calculated for S_p, C_p and *P-score* values. Percentage of subsequences retained at each P-score cut-off (0.00, 0.05, 0.10...1.00) was calculated.

4.3 Results

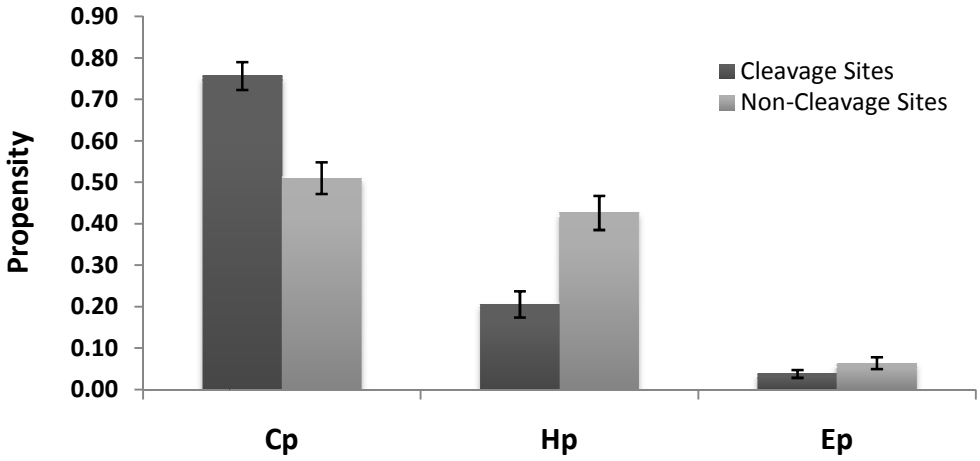
4.3.1 Propensity for unstructured regions

Here, it is hypothesized that the cleavage sites of caspase substrates are likely to locate in unstructured segments spanning across the tetrapeptide sequence on the folded protein. To measure the propensities for secondary structures across the cleavage site region, 24-mer peptide subsequences comprising of the tetrapeptide cleavage site with upstream 10 residues up to P_{14} and downstream 10 residues up to P_{10}' was constructed from respective substrates in the analysis dataset and predicted for secondary structural elements using SABLE II protein structure prediction server. The propensities for secondary structures were quantified as H_p, E_p and C_p scores for helices, beta-strands and coils respectively. As shown in Fig 4-1(A), the propensity for coils were significantly greater for cleavage sites sequences compared to non-cleavage sites (mean C_p=0.76 versus C_p=0.51 respectively, P-value<0.01), while the propensity for helices were greater for non-cleavage sites compared to cleavage sites (mean H_p=0.43 versus H_p=0.21 respectively, P-value<0.01). The distribution of cleavage sites and non-cleavage sites were further analyzed across C_p bins (Fig 4-1B). It was shown that cleavage sites were distributed more frequently to bins of higher C_p scores compared to non-cleavage sites. Most cleavage sites were confined to bins 0.8-1.0 while a greater proportion of non-cleavage sites were distributed to bins less than 0.8. These results suggest that caspase substrate cleavage sites tend to locate in unstructured sequence segments.

4.3.2 Propensity for solvent exposure

It is also hypothesized that cleavage sites will preferentially locate in solvent exposed regions of substrates. Therefore, the solvent accessibilities of the 24-mer peptide subsequences from the analysis dataset were predicted. Using the SABLE II server, real-value scores for each 24-mer subsequence were obtained and computed a score, S_p for measuring the propensity for solvent exposure. As shown in Figure 4-2(A), sequences containing cleavage sites were on the average, more exposed to solvent compared to non cleavage sites (mean $S_p = 0.50$ versus $S_p=0.43$ respectively, P-value < 0.01). The distribution of sequences, with or without cleavage sites across S_p bins (Figure 4-2B) was further analyzed. Both cleavage and non-cleavage sites were found to be increasing distributed into regions of greater solvent exposure as S_p increases from 0 to 0.40. However, distribution of non-cleavage sites peaked at S_p of 0.40, while the distribution for cleavage sites peaked at 0.60, before falling off at higher S_p bins values. These results suggest that caspase cleavage sites tend to locate in solvent exposed regions on substrates.

A



B.

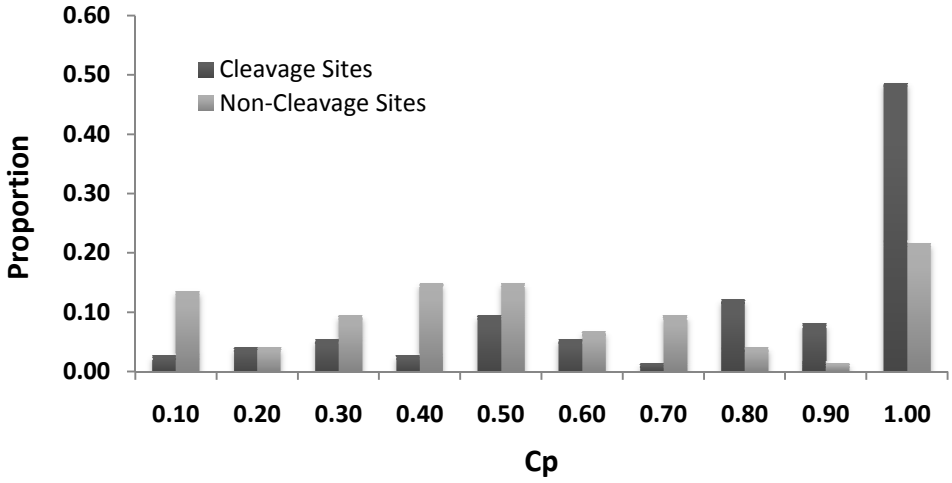
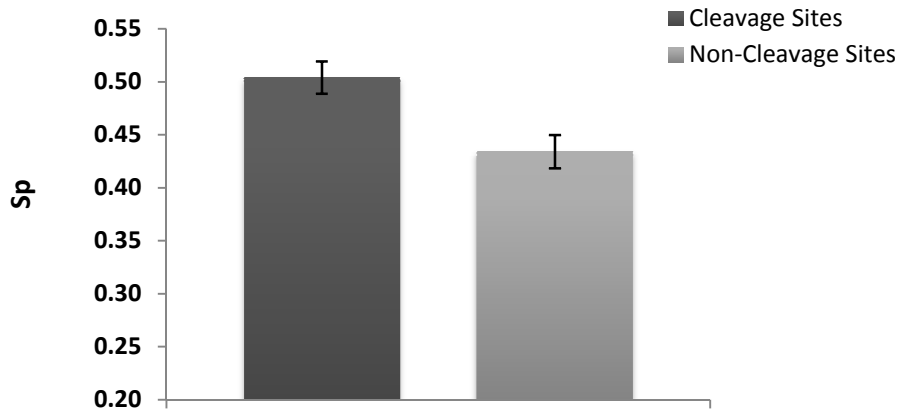


Figure 4-1 Propensity for secondary structures. (A) Propensities for different secondary structure elements (*coils*, C_p; *helices*, H_p; *beta-strands*, E_p) were measured for 24-mer sequences with or without caspase cleavage sites (labeled "*cleavage sites*" and "*non-cleavage sites*" respectively). (B) Distribution of "*cleavage sites*" and "*non-cleavage sites*" to C_p bins. Each C_p bin (0.10, 0.20...1.00) was allocated a proportion of sequences with C_p scores falling within the bin range (0-0.10, 0.11-0.19...0.91-1.00).

A.



B.

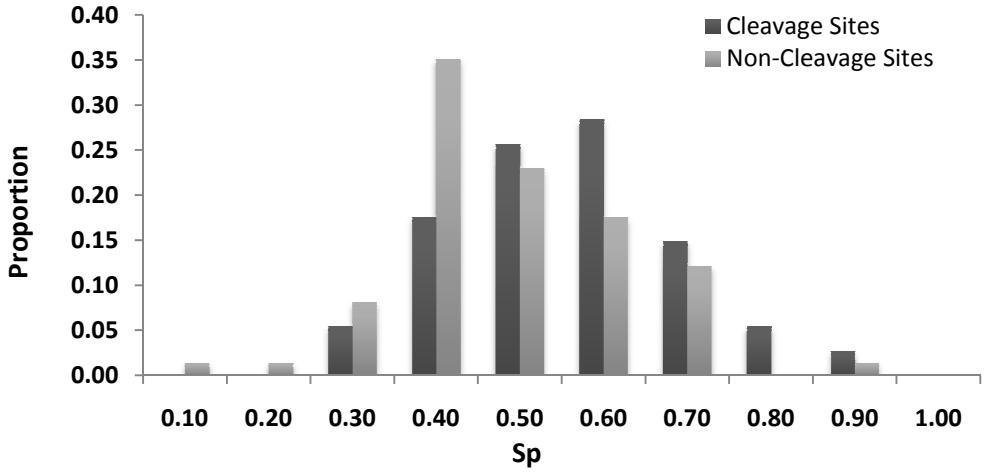
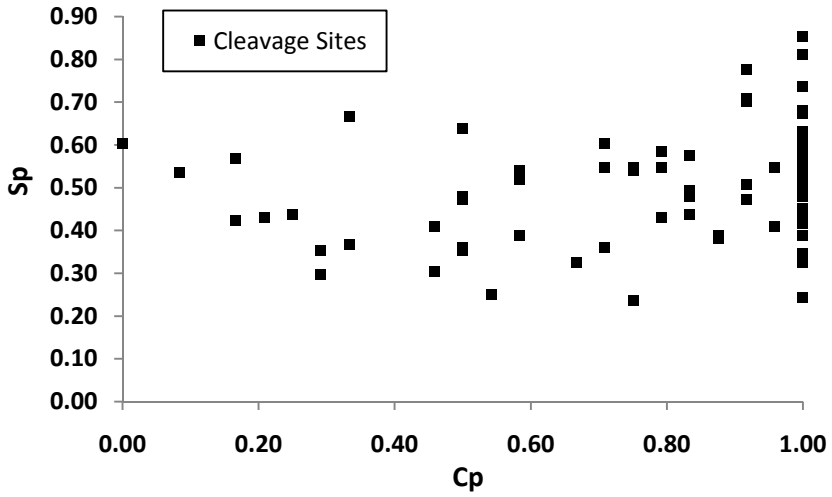


Figure 4-2 Propensity for solvent accessibility. (A) Propensities for solvent accessibilities (S_p) were measured for 24-mer sequences with or without caspase cleavage sites (labeled "*cleavage sites*" and "*non-cleavage sites*" respectively). (B) Distribution of "*cleavage sites*" and "*non-cleavage sites*" to S_p bins. . Each S_p bin (0.10, 0.20...1.00) was allocated a proportion of sequences with S_p scores falling within the bin range (0-0.10, 0.11-0.19...0.91-1.00).

A.



B.

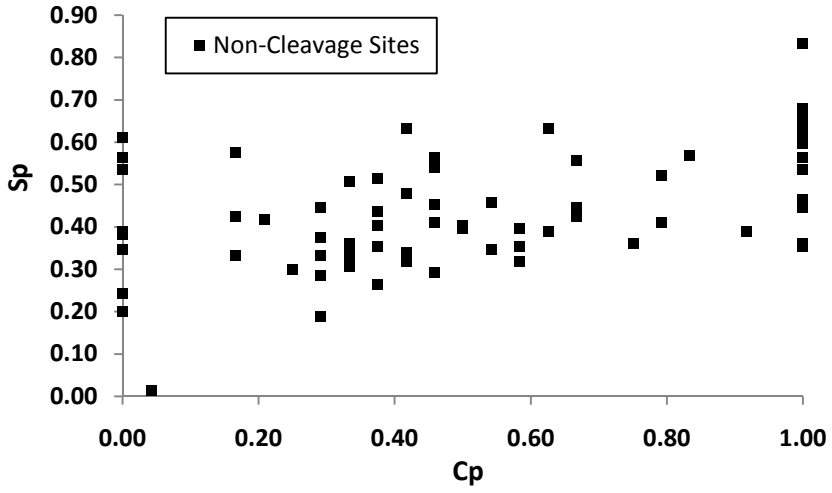


Figure 4-3 Scatter plots of S_p and C_p values for cleavage sites (A) and non-cleavage sites (B). Each data point corresponds to the S_p and C_p values of a single 24-mer sequence in the respective datasets.

4.3.3 Multi-factor model for prediction of caspase substrates

The major factor determining caspase cleavage is the presence of specific tetrapeptide sequences recognized by caspases. With the exception of GraBCas and CASVM, computational tools rely primarily on these sequences in their algorithms for prediction. However, as mentioned earlier, the usage of tetrapeptide specificities solely for detection of substrate cleavage is likely to produce a high percentage of false positives. Given that bona fide cleavage sites tend to locate in unstructured and solvent exposed regions, it is hypothesized if these two factors can be used to improve the prediction of caspase substrates by filtering out predicted sequences with unfavorable structural characteristics. Accordingly, a model is proposed for prediction of caspase substrates using a two-step approach. In the first step, a set of potential caspase cleavage sites is predicted for a given protein sequence using the relevant cleavage site prediction tool. In the second step, predicted sequences locating in highly structured and solvent inaccessible regions will be eliminated, leaving behind sequences with greater likelihood for cleavage. Here, to account for structural factors, 24-mer sequences consisting of the predicted tetrapeptide sequences and upstream and downstream flanking residues are constructed, and values for S_p and C_p are calculated. Further analyses indicate a positive correlation between S_p and C_p values across cleavage sites and non-cleavage sites (Figure 4-3, overall correlation coefficient, r = 0.43). Accordingly, the S_p and C_p values are combined into a single variable; *P-score* with optimized weightings of both factors selected at 0.7 and 0.3 respectively (data not shown). Based on the *P-score*, cut-off levels are assigned where predicted sequences from the first step scoring above the cut-off are retained while the rest are eliminated. The two-step model is further illustrated in Figure 4-4.

To corroborate the model, two distinctive caspase cleavage sites prediction tools, CASVM (described in Chapter 3) and GraBCas (Backes *et al*., 2005), were utilized. CASVM uses the support vector machines algorithm for prediction and has an optimal sensitivity of 89% when tested on an independent dataset of caspase cleavage sites. GraBCas, on the other hand, utilizes position-specific scoring matrices for scoring tetrapeptide sequences and achieved a sensitivity of 70% (cut-off > 0.1) using the same independent dataset. The test dataset consists of 14 caspase substrates, with a total of 17 unique caspase cleavage sites. Stepping through the first stage of the model, CASVM predicted 80 tetrapeptide sequences across all substrates as potential caspase cleavage sites, including all 17 cleavage site sequences from the test dataset. On the hand, GraBCas predicted a total of 223 tetrapeptide sequences, with only 15 out of the 17 true cleavage sites included in the prediction results. Here, all predicted caspase cleavage sites verified as cleaved were assigned to be true positives and all others to be false positives. At the stage, all sequences (true positives and false positives) were calculated for their *P-score* values and filtered through cut-off values of increasing stringency. As shown in Figure 4-5(A), all CASVM predicted sequences were retained at *P-score* of 0 but were steadily eliminated as the cut-off progressed to 1.00. In addition, at all *P-score* cut-offs, proportionately more true positives were retained compared to the false positives. At *P-score* of 0.50, about 88% of true positives were retained but more than half of the false positives were eliminated. Similar results were obtained when CASVM was replaced with GraBCas at the first step of the model.

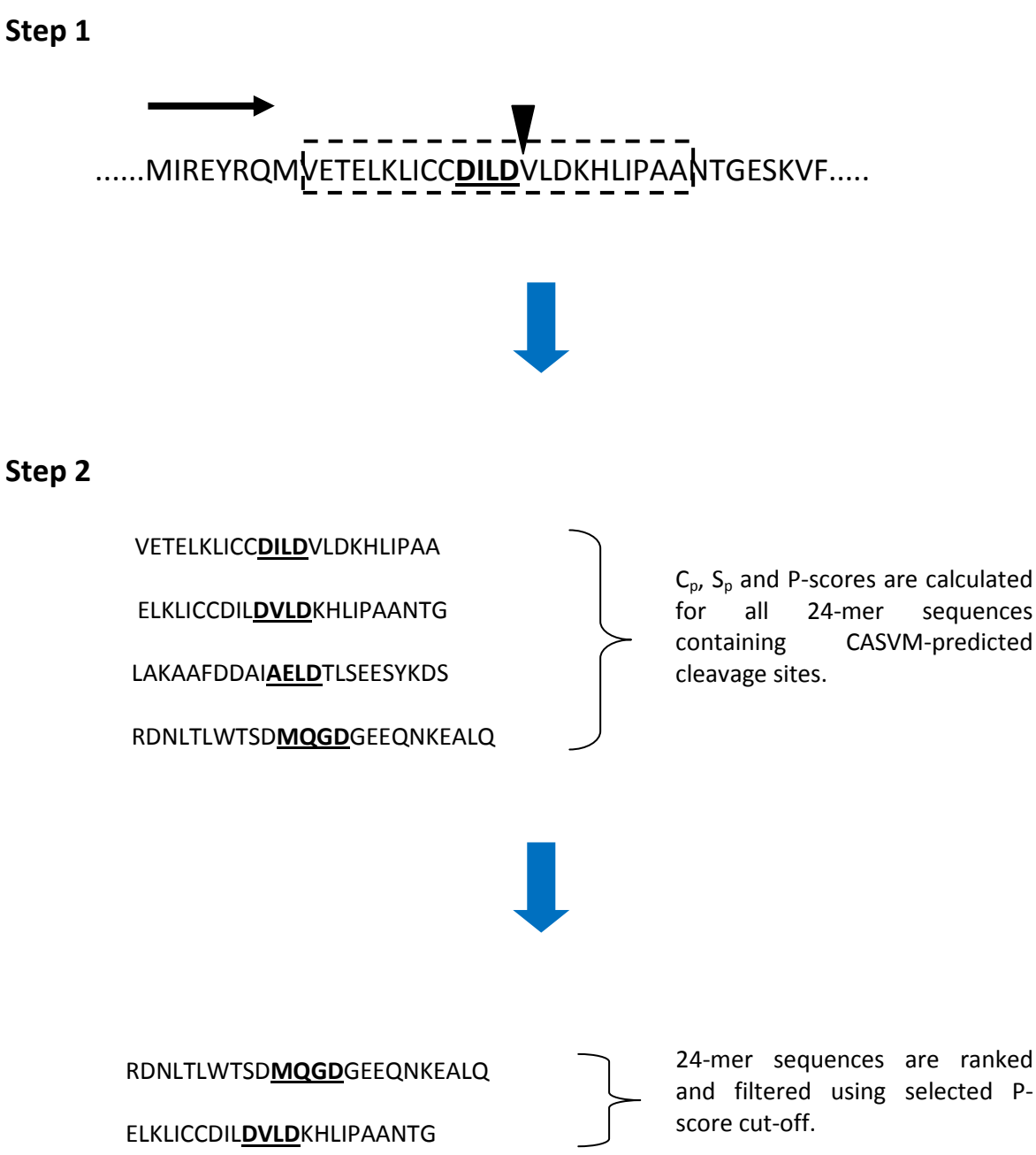


Figure 4-4 Schematic diagram of the two-step model for caspase substrate prediction. *Step1*: A window scans the entire protein (example: 14-3-3, Uniprot ID: P31946) for potential caspase cleavage sites (*bold, underlined*) using a caspase cleavage site prediction tool (a 24 residue window from CASVM shown as example). *Step 2*: Predicted cleavage sites are extracted together with flanking 10 residues downstream and upstream from the tetrapeptide sequence (P_{14} to P_{10}'). C_p, S_p and P-scores are calculated for all subsequences. Subsequences are ranked according to their P-scores and filtered using the desired cut-off score.

As shown in Figure 4-6, all sequences were eliminated in tandem with the increase in the *P-score* cut-off but proportionately more true positives were retained at all P-score cut-offs compared to false positives. Interestingly, results obtained using GraBCas showed that false positives elimination began at *P-score* cut-off of 0.20 and for true positives, at 0.55. When CASVM was used, false positives were eliminated onwards from *P-score* cut-off of 0.15 and true positives from 0.35. The higher *P-score* cut-off for true positives in both cases suggest that the model is likely to improve substrate prediction accuracy since a notable proportion of false positives (up to 13% and 53% in CASVM and GraBCas models respectively) can be eliminated without reduction in the original pool of true positives. Taken together, these results further indicate that structural factors such as secondary structures and solvent accessibilities can be effective as additional factors for identifying bona fide caspase cleavage sites. Specifically, accuracy of substrate prediction can be improved when these factors are used for differentiating the true positives and false positives from results obtained through cleavage sites prediction tools.

4.4 Discussion

In this chapter, a computational model was implemented for the prediction of caspase substrates using a two-step approach. The entire protein sequence is first scanned for potential cleavage sites using a caspase cleavage sites prediction algorithm. The predicted cleavage sites are filtered through a scoring system which quantifies the propensities of predicted cleavage sites to locate in unstructured and solvent exposed regions on the protein. The incorporation of additional factors, such as secondary structures and solvent accessibilities, was found to augment accuracy in HIV protease

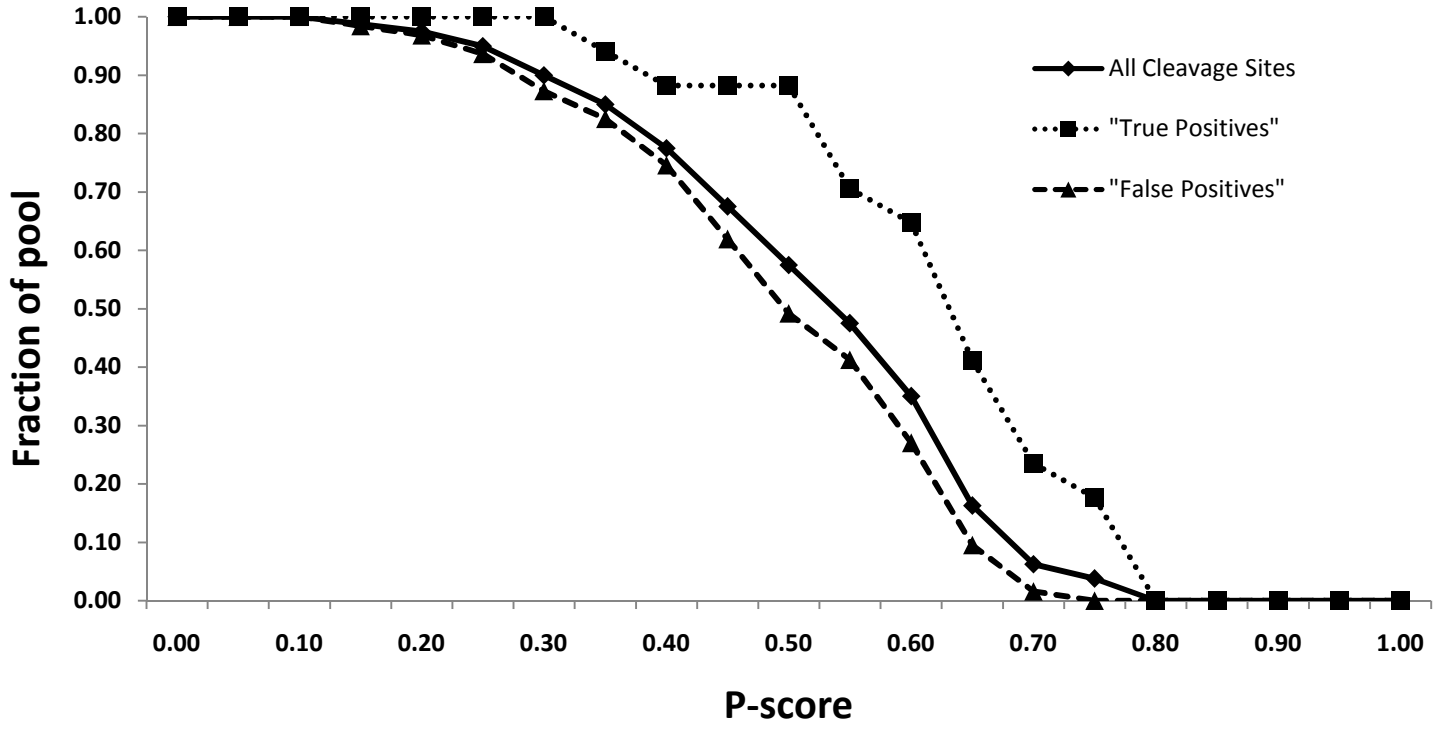


Figure 4-5 Results of caspase substrate prediction model on test dataset. CASVM predicted cleavage sites were assigned to pools containing "true positives" or "false positives". Fractions of cleavage sites in the assigned pools (*vertical axis*) with P-scores above the cut-offs (*horizontal axis*) were measured.

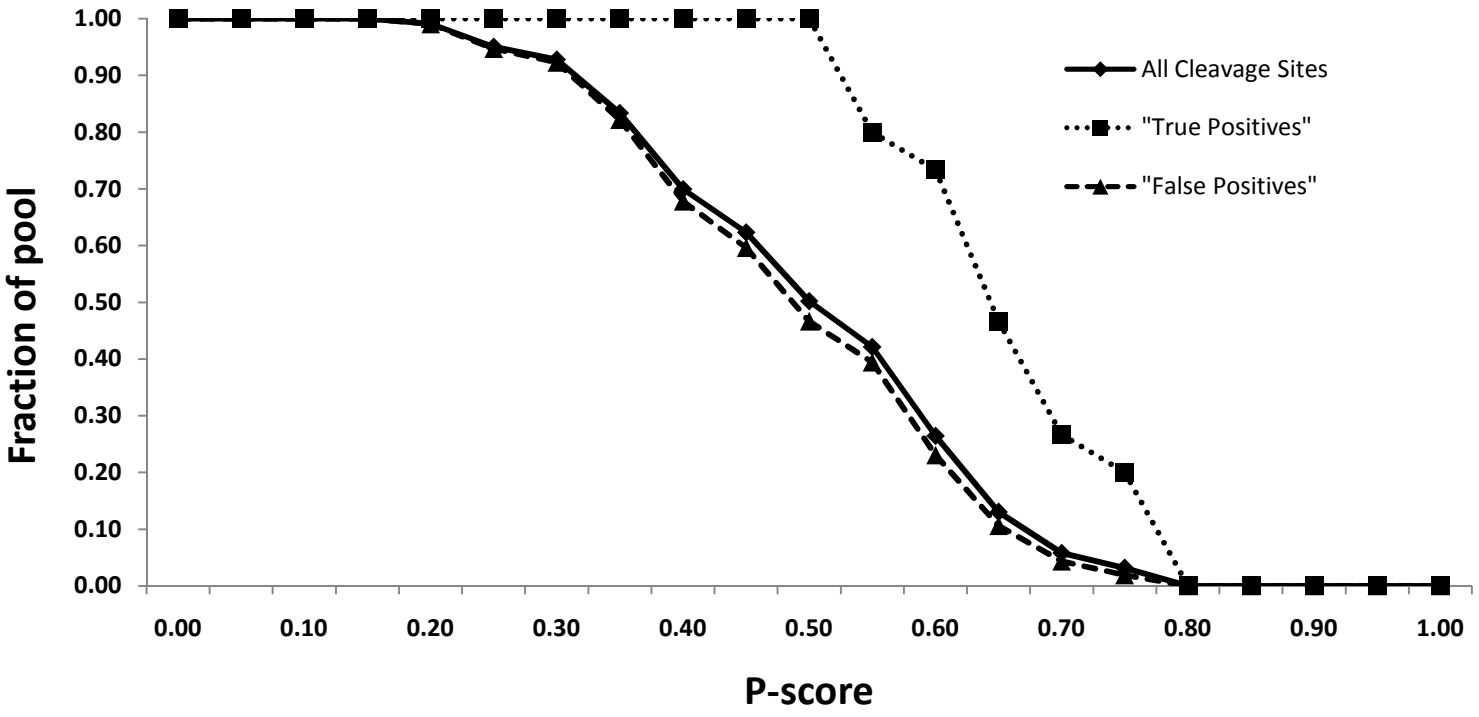


Figure 4-6 Results of caspase substrate prediction model on test dataset. GraBCas predicted cleavage sites were assigned to pools containing "true positives" or "false positives". Fractions of cleavage sites in the assigned pools (*vertical axis*) with P-scores above the cut-offs (*horizontal axis*) were measured.

substrates prediction (You, 2006). There, a three-level hierarchical classifier scans a protein sequence for HIV protease cleavage sites using specificity data and filters the output for sequences located within disorganized secondary structures and solvent exposed regions. These structural factors were similarly integrated into the neural network algorithms for RNA and DNA binding sites prediction and were found to be effective (Ahmad *et al*., 2004; Jeong *et al*., 2004). Here, instead of combining the prediction of cleavage sites specificity, secondary structures and solvent accessibilities into a single predictor, these factors were accounted in two distinct steps for the purpose of addressing a couple of caveats implicit in protease substrate prediction. While cleavage sites were shown to preferentially locate in unstructured and solvent exposed regions, not all predicted cleavage sites with substantial propensity for these factors are cleaved *in vivo* (Timmer and Salvesen, 2007). It is conceivable that regulatory processes such as post-translational modifications and other protein-protein interactions will also influence the final proteolytic event. Conversely, predicted cleavage sites sequences which are hidden in deep hydrophobic cores of proteins – and characterized by low propensities for solvent exposure – cannot be ruled out as it is possible that these sequences may be exposed following a proteolytic cleavage of the protein. Evidently, the caspase-mediated cleavage of ETK (epithelial and endothelial tyrosine kinase) was suggested to proceed in a two-step fashion where the first caspase cleavage site of the protein exposes an internal cleavage site for subsequent cleavage (Wu *et al*., 2001). The retinoblastoma protein, RB, the hepatocyte growth factor receptor, MET and GTP exchange factor for small G-protein Ras, RasGAP were all shown to be cleaved sequentially at multiple sites, suggesting the possibility of structural changes following an

upstream proteolytic cleavage event (reviewed in Fischer *et al*., 2003). To circumvent these constraints, the present model proposes a broad pool of potential cleavage sites which can be narrowed down through a range of *P-score* cut-off levels as guided by experimental data and user requirements.

The two-step model was tested using two cleavage site prediction methods– CASVM and GraBCas. It was shown that in both cases, the discrimination of prospective cleavage sites based on additional structural characteristics would be helpful for reducing false positives. The GraBCas model outperformed the CASVM model by eliminating a greater percentage of false positives before the first reduction in true positives. One possibility for the disparity could be the different sequence windows used for analysis in each case. GraBCas requires only the tetrapeptide cleavage sequence while a 24-mer peptide (tetrapeptide sequence with flanking 10 residues upstream and downstream) is needed for input into CASVM. Presumably, in the latter case, information encoding factors for cleavage sites specificity would have overlapped to a greater extent with that for secondary structures and solvent exposure. In any case, the results suggest that other cleavage sites prediction tools utilizing algorithms with low correlation with secondary structure and solvent accessibility prediction could be integrated into the model. Conversely, the addition of other factors with low correlations with cleavage site prediction would be helpful for improving prediction accuracy. Recent studies have suggested that exosites-interaction sites distal from the enzyme active site-could mediate substrate cleavage and are responsible for non-canonical caspase substrate cleavage. Structural studies by Agniswarmy *et al*. (2007) highlighted a symmetrical pentapeptide binding pocket on caspase-7 situated way from the active site which could function as an

exosite. Exosites was shown also to be involved in proteolytic events mediated by blood coagulation proteases (Bode *et al*., 1997). In addition, it was reported that serine phosphorylation of caspase cleavage sites, particularly on the P_4 and P_1' residues, inhibited substrate cleavage (Tözsér *et al*., 2003). It is expected that the incorporation of these factors will be helpful for further reducing the false positives in prediction of caspase substrates.

4.5 Conclusion

In this chapter, the structural characteristics of reported caspase substrates were analyzed indicating that caspase cleavage sites tend to locate in unstructured and solvent exposed regions. A score-based two-step model was constructed to integrate these factors with existing cleavage sites prediction tools, CASVM and GraBCas. It was shown that the model improved accuracy by reducing false positives from cleavage sites prediction. As it is likely that other factors are involved in determining substrate cleavage, the development of algorithmic tools and availability of larger datasets will complement efforts in computational prediction. In concert with biochemical and biophysical studies, these *in silico* efforts will complement system-level studies on the caspase degradome and further our understanding of the intricacies of caspase-substrate biochemistry. In spite of the differences between protease degradomes, the prediction of specific substrate cleavage is likely to be largely influenced by a set of common factors such as the presence of the required cleavage site and its presentation. Therefore, with relevant and sufficient data, the model presented here can be reasonably extended for the prediction of other protease substrates as well.

Chapter 5:

Caspase Cleavage of Receptor Tyrosine Kinases

5.1 Introduction

As a step towards elucidating the caspase degradome, the annotated human proteome, based on all protein entries reported in Swiss-Prot database (Release 51.4), was scanned for potential caspase cleavage sites using the CASVM server (with default options). The results showed that 12589 out of 15417 human proteins (~82%) in the Swiss-Prot dataset possess at least one caspase cleavage site. A similarly study on the proteome-wide prevalence of caspase cleavage sites was implemented by Garay-Malpartida *et al*. (2005) using the CasPredictor software. In contrast, the CasPredictor software predicted 16.46% of proteins to possess at least one caspase cleavage site out of a dataset of 9986 proteins from Swiss-Prot database (release version not reported). It is not surprising that both prediction methods produced vastly differing results since they are developed on different algorithms. However, given the higher sensitivity of the SVM method, it is likely that the true percentage is greater than the latter. In any case, these results further support the expectation of many more caspase substrates are to be discovered in the future.

In this chapter, a subset of the annotated human proteome - the family of receptor tyrosine kinases - was predicted for potential caspase substrates. Analyses on the results suggest a general mechanism of regulation of these proteins by caspase cleavage and present useful leads for future experimental studies.

5.2 Biochemistry of receptor tyrosine kinases

Protein kinases are enzymes that play a key regulatory role in nearly every aspect of cell biology (Roskoski, 2004). These enzymes tweak protein function through the catalysis of the transfer of phosphate groups from ATP molecules to specific amino acids on proteins. Based upon the nature of the target amino acid, these enzymes are classified as protein serine/threonine kinases or protein tyrosine kinases. They regulate a plethora of cellular functions such as apoptosis, cell cycle progression, cytoskeletal rearrangement, differentiation and development. Dysregulation of protein kinases occurs in a variety of diseases including cancer, diabetes, autoimmune, cardiovascular, inflammatory, and nervous disorders. Manning *et al*. (2002) identified 478 typical and 40 atypical protein kinase genes in humans (total 518) that correspond to about 2% of all human genes. The family includes 385 serine/threonine kinases, 90 protein tyrosine kinases, and 43 tyrosine kinase-like proteins. Of the protein tyrosine kinases, over 50 are receptor tyrosine kinases (or RTKs for short). The RTK family includes, among others, epidermal growth factor receptor (EGFR), platelet-derived growth factor receptors, fibroblast growth factor receptors (FGFRs), vascular endothelial growth factor receptors, Met (hepatocyte growth factor/scatter factor [HGF/SF] receptor), Ephs (ephrin receptors), and the insulin receptor.

As reviewed in Hubbard and Miller (2007), RTKs are single-pass, type I receptors localized in the plasma membrane. Generally, RTKs are activated through ligand-induced oligomerization, usually dimerization, which juxtaposes the cytoplasmic tyrosine kinase domains. For most RTKs, this juxtaposition facilitates autophosphorylation in *trans* of tyrosine residues in the kinase activation loop or juxtamembrane region, inducing

conformational changes that stabilize the active state of the kinase. The phosphotyrosine residues on the intracellular region of the receptors serve as recruitment sites for the binding of signaling or adaptor proteins. These proteins redirect the trans-membrane signals into several distinctive pathways for cell proliferation and other important cellular responses.

A common cellular signaling circuitry mediated by the RTK is typified in Figure 5-1 using the HGF/SF receptor (or MET) as example (Tulasne and Foveau, 2008). Upon ligand binding, activated MET receptor recruits SH2 domain containing proteins such as GrB2 to its phosphorylated docking sites on the cytoplasmic domain. Sos, a guanine nucleotide exchange factor for Ras, is recruited together with Grb2 to the activated receptor, leading to the activation of membrane bound Ras. Activated Ras activates Raf, which phosphorylates and activates other members of the MAP kinase cascade, culminating in the activation of MAP kinase, which translocates into the nucleus and stimulates transcription of survival genes such as Bcl-2 and Bcl-xL. An alternative, complementary pathway involves the binding of PI-3-kinase, another SH2 domain containing adaptor protein, to the activated MET receptor. Activated PI-3-kinase phosphorylates and activates PKB, which in turn, phosphorylates and inactivates Bad, a pro-apoptotic effector of the family of Bcl-2 proteins. Both downstream pathways, through counteracting external cell death signals and those initiated within the cellular milieu, fine tune the equilibrium of cell survival and death. For most RTKs, several other signaling pathways were found to propagate from the activated receptor through

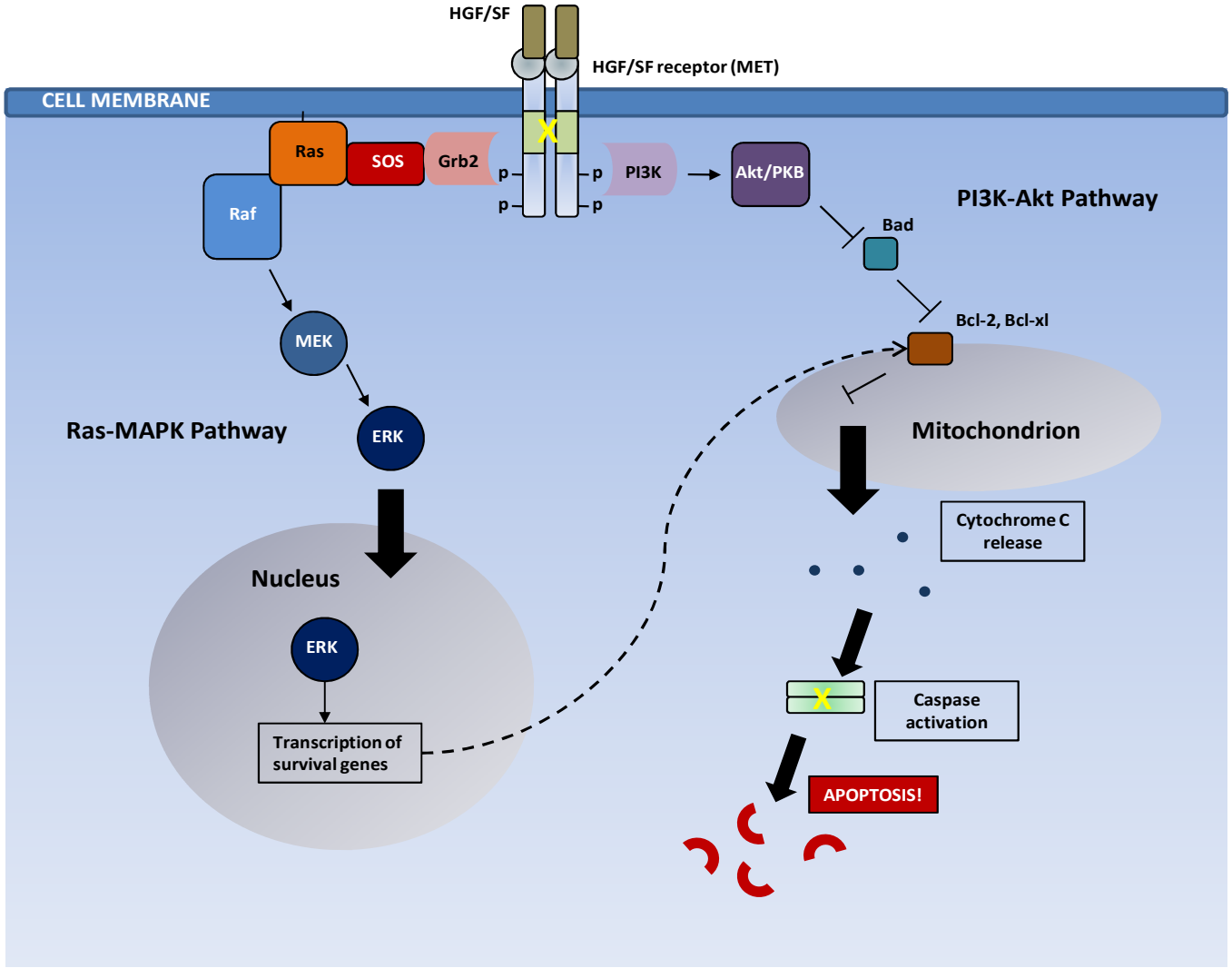


Figure 5-1 Trans-membrane signaling in ligand-activated HGF/SF receptor (MET). Upon ligand binding, MET receptors dimerize and becomes activated through cross-phosphorylation of their cytoplasmic domains. These phosphotyrosine residues serve as binding sites for signaling molecules which propagate the signal into two distinct pathways: the Ras-MAPK and PI3K-Akt pathways.

interactions with diverse classes of molecules, as well as from signaling cross-talks with other membrane receptors such as the G-protein coupled receptors and cytokine receptors.

As essential components of cellular signaling pathways mediating embryonic development, adult homeostasis and other critical processes, RTK activity in resting, normal cells is tightly controlled. It has been shown that mutations or structural aberrations in RTKs convert them to potent oncoproteins, contributing to the development and progression of many cancers. Consequently, RTKs and their cognate ligands have become rational targets for therapeutic intervention using humanized antibodies and small molecule drugs (Gschwind, 2004). In recent years, RTK-based cancer therapies have attained widespread clinical use for a number of cancer phenotypes – examples include trastuzumab (Herceptin) for metastatic breast cancer, imatinib (Glivec) for chronic myelogenous leukemia, gefitinib (Iressa) for non-small cell lung carcinoma, cetuximab (Erbitux) and bevacizumab (Avastin) for colorectal cancer. In addition, a growing pipeline of drug compounds targeting several members of the RTK family, such as EGFR, VEGFR, PDGFR, KIT and FLT3, are currently in clinical development.

5.3 Caspase cleavage of RTKs

Recent studies have implicated several members of the RTK family as proteolytic targets of caspases during apoptosis. Caspase cleavage of RTKs was found to be localized to the cytoplasmic domains of the receptors, frequently leading to the suppression of ligand-mediated trans-membrane signaling through the structural alteration of the kinase domains or the deletion of binding sites for downstream signaling proteins. Since the RTKs mediate downstream growth and survival signals

which are antagonistic to the apoptotic program, the cleavage of RTKs, in concert with other processes, appears to tip the balance of survival and death signals, leading to cellular commitment to apoptosis.

Caspase cleavage of MET receptor at Asp^{1000} results in the inactivation of functional MET by loss of its signaling cytoplasmic part, with the concomitant appearance of membrane bound 100 kDa MET and a soluble intracellular 40 kDa MET (Tulasne *et al*., 2004; Foveau *et al*., 2007; Deheuninck *et al*., 2008). The 100 kDa MET fragment prevents downstream survival activity by trapping the HGF ligand, while the 40 kDa MET fragment becomes ligand-independent and acquires pro-apoptotic activity. Interestingly, the 40 kDa fragment encompasses the entire MET kinase domain and kinase activity was shown to be necessary for pro-apoptotic function. As the 40 kDa fragment does not activate MAP kinase pathway, its pro-apoptotic function was suggested to involve the kinase activity on pro-apoptotic effectors.

ERBB2, a member of the EGFR family of RTKs, was shown paradoxically to function as a cell death effector activated by caspases. Full length ERBB2 protects cells from apoptosis by activating survival pathways while caspase cleavage was shown to unmask a previously unrecognized pro-apoptotic function of the receptor (Strohecker *et al*., 2008). It was demonstrated that the cytoplasmic tail of ERBB2 is cleaved at four sites (Asp^{1016}, Asp^{1019}, Asp^{1087} and Asp^{1087}) by caspases. Cleavage of the receptor at Asp^{1016} or Asp^{1019} releases an intracellular carboxyl terminal 47 kDa fragment that is subsequently proteolyzed by caspases at Asp^{1125} to a predicted 25 kDa product. Interestingly, both the 47 and 25 kDa products were found to induce apoptosis upon ectopic expression in cells and each of them possess a BH3-like domain similar to those found in the BH3-only pro-apoptotic members of the Bcl-2

family. EGFR was also shown to be cleaved Asp^{1012} and less efficiently at Asp^{1172} (Bae *et al*., 2001; He *et al*., 2006). The cleavage at Asp^{1012} terminated downstream signaling through the removal of the binding sites for Cbl and Grb2 in cytoplasmic region of the receptor. However, unlike the ERBB2 receptor, no pro-apoptotic activity downstream of EGFR cleavage was reported.

The RET proto-oncogene was found to induce apoptosis in cells when expressed ectopically (Bordeaux *et al*., 2000). The induction of apoptosis was inhibited in the presence of its ligand, glial cell line derived neurotrophic factor (GDNF), suggesting RET behaves as a dependence receptor. The pro-apoptotic activity of RET was mediated through the cleavage of the receptor by caspases at Asp^{707} and Asp^{1017} - releasing a pro-apoptotic fragment whose activity was localized to the region bound by the two cleavage sites. Another dependence receptor, the anaplastic lymphoma kinase protein (ALK), was similarly shown to induce apoptosis in the absence of its cognate ligand (Mourali *et al*., 2008). The pro-apoptotic activity of unligated ALK is enhanced through the cleavage of the receptor by caspases at the juxtamembrane region (Asp^{1160}). Cleavage of ALK exposes a pro-apoptotic region on the receptor which was mapped to region upstream of the cleavage site in the cytoplasmic domain of the receptor.

As discussed in Chapter 1, the consequences of caspase cleavage of a signaling protein primarily results in the loss or gain of its function, depending on its role in the signaling pathway. However, as mentioned above, there is mounting evidence to suggest that RTK function is intimately regulated by caspase activity and the consequences of RTK cleavage are far more complex than simply the gain or loss of protein function. Also, given the pervasive role of RTKs in cell survival pathways

and their implications in diseases such as cancer, it is plausible that many other RTKs are hitherto undiscovered downstream targets of caspases.

5.4 Prediction of caspase cleavage sites on RTKs

The complete repertoire of RTKs - 52 members across 16 sub-families, as listed in the KEGG database (Kanehisa and Goto, 2000) - was retrieved from Uniprot database and predicted for caspase cleavage sites using the previously developed SVM method. Protein sequences of RTKs were submitted to the CASVM server under default settings and results of prediction are listed in Table 5-1. All RTKs were predicted to possess caspase cleavage sites. Predicted cleavage sites were distributed throughout the length of the extracellular and intracellular regions of RTKs. About 92% of all RTKs (48/52) possess cleavage sites on the intracellular region while about 98% (51/52) contain extracellular cleavage sites. While predicted cleavage sites localize throughout the length of the receptors, certain trends could be observed in the distribution of predicted sites, implying functional significance downstream of caspase cleavage.

A sizeable number of RTKs (~21%) were predicted for caspase cleavage sites at the juxtamembrane region on the cytoplasmic side of the receptors. It is conceivable that caspase cleavage at these sites will truncate the full length receptor into a membrane bound portion and an intracellular fragment. As such, receptor cleavage will likely lead to abrogation of downstream signaling and possibly interfering with normal RTK signaling through the trapping of ligands by the membrane-bound receptors (e.g. ALK cleavage). Significantly, recent studies suggest that an intracellular receptor fragment containing the tyrosine kinase domain may have downstream functional implications. Observations on high-throughput proteomic

screening of caspase substrates in Dix *et al*. (2008), reported that a substantial number of caspase substrates are cleaved into persistently stable, domain-containing fragments, and speculated that caspase-mediated proteolysis yields a class of effector protein fragments with novel functions. Moreover, as discussed earlier, cleavage of MET and ALK at their juxtamembrane region led to the release of pro-apoptotic intracellular receptor fragments. In both cases, the kinase domains were implicated in the pro-apoptotic response, though the exact mechanisms were unclear.

On a related note, close to 80% (41/52) of all RTKs harbor caspase cleavage sites within the tyrosine kinase domain of the receptor. In particular, RTKs from the insulin receptor and FGF receptor sub-families are annotated with multiple cleavage sites within their tyrosine kinase domains. As these domains serve as key mediators of signal transduction for RTKs, structural alterations from caspase cleavage may lead to perturbations of downstream signaling. Interestingly, studies by Tikhomirov *et al*., (2005) indicate that proteolytic fragments bearing the motif "RLLGI" derived from the tyrosine kinase domains of EGFR, Erbb2, Erbb4, TrkA and VEGFR1 were able to induce apoptosis in cells. Indeed, caspase cleavage sites were predicted in the kinase domains of some of these receptors (EGFR; Asp^{770}, Asp^{916}, Erbb4; Asp^{878}, Asp^{922} and VEGFR1; Asp^{958}, Asp^{987} Asp^{1135}), suggesting the possibility of caspase cleavage and release of pro-apoptotic intracellular kinase fragments. As the "RLLGI" motif is suggested to be prevalent among RTKs, it is possible that cleavage of the tyrosine kinase domains of several other RTKs could lead to the similar production of such pro-apoptotic fragments. Studies on Erbb2 cleavage have shown that the caspase cleavage produced pro-apoptotic intracellular fragments downstream of the kinase domain in the C-terminal region of the receptor. Cleavage of EGFR at a comparable location was shown but no pro-apoptotic consequences were reported. Interestingly,

the other members of the EGFR family, Erbb3 and Erbb4, were predicted to possess similarly located caspase cleavage sites, suggesting that these proteins could be caspase targets as well.

Taken together, the presence and distribution of these predicted cleavage sites across the RTK family suggest that specific roles of caspase cleavage in regulating RTK function. It is tempting to speculate a general phenomenon whereby caspase cleavage of RTKs leads to a molecular "life-death" switch which converts the pro-survival protein to a pro-apoptotic one through the exposure and/or the release of pro-apoptotic domains. Such reversal of protein function is similarly observed in the caspase cleavage of serine/threonine protein kinases, MEKK1 and MEKK4, which generated pro-apoptotic fragments upon cleavage at their kinase domains. Several other anti-apoptotic proteins such as Bcl-2 and Bcl-xl have been shown to be converted into pro-apoptotic molecules by caspase cleavage. The elegant integration of both anti- and pro- apoptotic functionalities on the same signaling protein is an uncommon but economical feature. Could it be reasoned that in an apoptotic cell, functionally disrupted cell survival and growth machineries would be further used for executing the stepwise destruction of the cell since their original function will no longer be needed?

As most of the RTKs were predicted to harbor caspase cleavage sites on the extracellular domains, it is appealing to speculate if there are specific functional consequences of cleavage at these locations. While caspases have been shown to be involved in the cleavage of a myriad of substrates, most if not all, are localized in the cytoplasm. However, active caspases were found to be released into the extracellular environment during apoptosis (Hentze *et al*., 2001). In addition, work by Cowan and co-workers (2005) provided evidence for the localization of active caspase-2, caspase-

3 and caspase-7 to the membrane surfaces of apoptotic smooth muscle cells. The caspases were thought to be bound to the cell surface receptors following release from the apoptotic cells and are involved in degrading the extracellular matrix as part of mediating the regression of advanced vascular disease. Clearly, future investigations on caspase activity in the extracellular environment will shed light on the possibility of RTK cleavage and downstream consequences.

5.5 Conclusion

In the chapter, a global scan of caspase cleavage sites on the family of receptor tyrosine kinases was conducted. The results suggest that RTK activity could be generally regulated by caspase cleavage. The presence of cleavage sites at the juxtamembrane and kinase domains of the RTKs indicate a likelihood of caspase-mediated abrogation of RTK signaling and the production of pro-apoptotic intracellular fragments. Future biochemical and structural studies on caspase-mediated RTK cleavage will be necessary to validate these issues. The results and discussion in this chapter is presented in the manuscript: **Wee et al., (2008) Multi-factor model for caspase degradome prediction**

Table 5-1 Global mapping of predicted caspase cleavage sites on receptor tyrosine kinases

RTK Family	RTKs	UNIPROT ID	Predicted Caspase Cleavage Sites[1]															
EGF receptor	EGFR	P00533	321	458	587	770	916	1006	1009	1012	1083	1127	1152	1171				
	ERBB2	P04626	277	326	382	639	1016	1019	1087	1125								
	ERBB3	P21860	162	165	242	581	1010	1020	1327									
	ERBB4	Q15303	218	245	300	335	510	564	585	595	878	922	1012	1015	1018	1068	1241	
Insulin receptor	INSR	P06213	75	483	526	546	549	672	704	716	949	985	1145	1210	1259	1330	1344	
	INSRR	P14616	154	585	676	816	916	1101	1166	1207	1280							
	IGF1R	P08069	156	300	342	519	539	542	675	1121	1186	1235	1294	1306				
	ROS1	P08922	100	358	483	513	684	711	842	1202	1391	1853	2058	2062	2135	2247		
PDGF receptor	PDGFRA	P16234	215	244	287	422	568	733	763	846	902	919	1015	1024	1033	1074		
	PDGFRB	P09619	78	200	285	575	691	737	1091									
	CSF1R	P07333	51	63	269	741	746											
	KIT	P10721	439	479	768													
	FLT3	P36888	200	455	600	959												
FGF receptor	FGFR1	P11362	69	90	110	130	131	132	133	142	218	527	768	782				
	FGFR2	P21802	75	126	135	136	138	506	521	530	785	794	795					
	FGFR3	P22607	77	136	139	143	147	497	516	521	776	792						
	FGFR4	P22455	119	129	187	240	507	516	575	770	779							

VEGF receptor	VEGFR1	P17948	372	495	630	958	987	1135	1165	1168	1262								
	VEGFR3	P35916	19	45	77	304	371	556	725	728	1130	1216	1274						
	VEGFR2	P35968	173	180	295	392	639	717	852	1141	1171	1174	1189	1259	1315				
HGF receptor	MET	P08581	174	231	352	824	1002	1231	1376	1380									
	MST1R	Q04912	126	176	204	299	355	375	671	805	927	936	1030	1045	1235				
TRK receptor	TRKA	P04629	53	209	306														
	TRKB	Q16620	173	277	349	406	409	424	476	579									
	TRKC	Q16288	61	193	476	641													
EPH receptor	EPHA1	P21709	32	45	158	252	592	778	841										
	EPHA2	P29317	33	232	250	708													
	EPHA3	P29320	17	34	159	282	299	531	708										
	EPHA4	P54764	35	161	241	319	402	542											
	EPHA5	P54756	65	187	190	348	430	995											
	EPHA6	Q9UF33	57	163	243	440	970												
	EPHA7	Q15375	37	314	404														
	EPHA8	P29322	55	61	222	241	729	790	940										
	EPHB1	P54762	24	118	138	528	771	840											
	EPHB2	P29323	25	139	318	774	777												
	EPHB3	P54753	138	333	785	786	918	935											

	EPHB4	P54760	226	242	836									
	EPHB6	O15197	63	139	142	271								
AXL receptor	AXL	P30530	87	260	270	389	407	551	648	769	843			
	TYRO3	Q06418	73	79	121	250	335	576	638	763				
	MERTK	Q12866	309	402	610	706	827	843	900	983				
LTK receptor	LTK	P29376	87	193	335	340	348	557	705					
	ALK	Q9UM73	305	516	885	951	954	993	1017	1163	1311	1606		
TIE receptor	TIE1	P35590	287	391	474	560	578	586	861	883				
	TIE2	Q02763	137	357	389	473	663	846	868	923				
ROR receptor	ROR1	Q01973	167	387	395	580	591							
	ROR2	Q01974	42	51	262	390	903							
DDR receptor	DDR1	Q08345	44	46	68	70	189	216	598	604	630	729		
	DDR2	Q16832	69	125	189	234	240							
RET receptor	RET	P07949	43	264	267	547	567	707	797	1017	1031			
RYK receptor	RYK	P34925	258	351	554									
MuSK receptor	MUSK	O15146	94	474	622	743	817							

[1] Positions of CASVM predicted cleavage sites on protein sequence of each RTK member are listed. Numbers indicate the positions of P_1 (Asp) residues on protein sequences. All cleavage site positions are color coded; grey indicates location of cleavage site within the extracellular domain, blue indicates location within intracellular domain and darker blue indicates location within kinase domain (all kinase domains of RTKs are located in the intracellular domain of the receptor). Predicted cleavage sites corresponding to true experimentally verified cleavage sites on EGFR, ERBB2, MET, ALK and RET are underlined.

Chapter 6: Conclusion

6.1 Summary of thesis

Caspases belong to a unique class of proteases which distinctly recognize and cleave after specific tetrapeptide sequences on proteins. By cleaving a diverse set of substrates, caspases serve as critical effectors in important cellular processes such as inflammation, apoptosis, cell survival and differentiation. As protease function is invariably linked to the nature of its downstream targets, the characterization of the caspase degradome is an essential step for furthering the current knowledge of caspase biology in health and disease.

In this thesis, the primary objective is the elucidation of the caspase degradome through the identification of known and hitherto undiscovered caspase substrates. Instead of utilizing experimental methods, an unconventional method using computational prediction was adopted. The framework for computational prediction of caspase substrates began with the construction of clean and reliable datasets on caspase substrates from published literature and biomedical databases. The outcome of this process was the establishment of a web-accessible database for caspase substrates (available at www.casbase.org/casvm/squery/index.html) and downloadable datasets containing experimentally-verified caspase substrates and associated annotations. Next, a computational prediction model based on the support vector machines (SVM) algorithm was developed for predicting caspase cleavage sites. By accounting for the preferential tetrapeptide cleavage site sequences and upstream and downstream flanking sequences, the model was shown to be comparable, if not superior to existing computational tools. Additionally, CASVM - a

web server integrating the SVM method for predicting caspase cleavage sites - was implemented and is available at www.casbase.org/casvm/index.html.

While the presence of cleavage sites is required for substrate cleavage, it is not sufficient for robust prediction of substrates on its own. Other factors such as the exposure of cleavage sites to solvent, local secondary structures, substrate localization and other post-translational modifications may be important as well. Consequently, the computational prediction model was extended to account for other structural features such as secondary structures and solvent exposure. Analyses of caspase substrates concluded that cleavage sites preferentially locate in regions on the substrates which are unstructured and solvent exposed. Accordingly, a two-step model for predicting caspase substrates was developed. In the first step, a set of potential caspase cleavage sites is predicted by the SVM model (using CASVM). In the second step, a score-based system quantifies predicted cleavage sites for their propensities to locate in solvent exposed and unstructured regions, and selects for those which are above a user-defined cut-off score. The two-step model was shown to improve the accuracy of substrate identification over standalone prediction of caspase cleavage sites by reducing proportion of false positives.

To explore the application of computational prediction for elucidating the caspase degradome, the receptor tyrosine kinase (RTK) family was predicted for potential caspase substrates. The RTKs belong to a class of membrane receptors which play critical roles in cell survival, proliferation and differentiation, and have been implicated in several human cancers. A total of 52 RTKs across 16 sub-families were predicted for the presence of caspase cleavage sites using the SVM method developed here. The predicted results showed a wide spread presence of caspase

cleavage sites among the RTKs. The observed patterns of cleavage site locations at the juxtamembrane and tyrosine kinase domains of the RTKs indicate a likelihood of caspase-mediated abrogation of RTK signalling and the production of pro-apoptotic intracellular fragments. The caspase-mediated cleavage of RTKs could be a novel mode of regulation for these receptors and possibly implicate them as molecular "life-death" switches.

6.2 Future directions

The prediction of caspase substrates serves as a complementary tool to experimental methods for the elucidation of the caspase degradome. While experimental methods are necessary for validation of true substrates, it is often cumbersome and expensive. On the other hand, computational prediction, much like weather forecasting, provides a peek into the likelihood of substrate cleavage without having to get the hands dirty - literally. The price of such convenience is accuracy. It is evident, then, that future work on this domain will be driven by the question: *what more can be done to improve prediction accuracy?*

As discussed in Chapter 4, a robust model for predicting caspase substrates – or any other protease substrates – will be dependent on integrating meaningful, non-correlated factors affecting substrate cleavage. The primary determinant of caspase cleavage remains the presence of distinctive tetrapeptide cleavage sites. Besides accounting for these specificities and the neighbouring flanking sequence tendencies, other structural factors such as the local secondary structures and degree of exposure to solvent are important. These factors have been carefully considered and are integrated into the model for two-step prediction of caspase substrates (as mentioned

in Chapter 4). However, as the inputs for these factors rely on results from secondary structures and solvent accessibilities prediction, there are still degrees of uncertainty. Clearly, future development of better secondary structure and solvent accessibilities prediction tools and algorithms, as well as the elucidation of three-dimensional structural data on caspase substrates, will be helpful for efforts here.

Beyond the cleavage sites, it has also been suggested other structural features on the substrate are important for cleavage. As discussed in Chapter 4, exosite-interaction sites distal from the enzyme active site could mediate substrate cleavage and may be responsible for non-canonical caspase substrate cleavage. As exosites were shown to be involved in other protease-substrate systems as well, it may be helpful to account for such features in the prediction model. In addition, recent studies have proposed that certain post-translational modifications on the substrate could mediate caspase cleavage. Serine phosphorylation of residues at or close to the substrate cleavage site was found to be inhibitory to substrate cleavage (Tözsér *et al*., 2003). Indeed, the phosphorylation of tau - a protein implicated in the formation of neuronal tangles in Alzheimer's disease - at Ser^{422}, was shown to inhibit cleavage by caspases *in vivo* (Guillozet-Bongaarts *et al*., 2006). More recently, SATB1, a protein associated with the matrix attachment regions, was shown to exhibit enhanced cleavage by caspases when the small-ubiquitin modifier (SUMO) protein was over-expressed in cells – suggesting that sumolyation may yet be another mode of regulation for caspase cleavage (Tan *et al*., 2008).

The involvement of caspases in seemingly unrelated and perhaps antagonistic processes such as cell proliferation and apoptosis argues for an even more complex blend of factors regulating substrate cleavage. If caspases are activated during mitosis,

how could caspase cleavage be restricted to those cell cycle regulators, while leaving other critical proteins intact? As discussed in Fischer *et al*. (2003), when caspases were activated and the cell cycle regulator Wee1 was cleaved after mitogenic T-cell stimulation, neither DNA replication factor RFC140 nor ICAD were cleaved in proliferating T cells. Cleavage of RFC140 and ICAD would lead to inhibition of DNA replication and fragmentation of genomic DNA, events that are not compatible with cell proliferation. Perhaps the answer could lie in a specific sub-cellular compartmentalization of caspases, the existence of scaffold proteins or a different accessibility of cleavable substrates. Other factors such as the coordinated expression of anti-apoptotic molecules or overall level of caspase activity may play a role in the selectivity of caspase cleavage. For instance, it was reported that the partial cleavage of Ras-GAP, a GTPase in the Ras signaling pathway, owing to low caspase activity first generates an N-terminal fragment that is anti-apoptotic by activating the PI3K pathway. Increased caspase levels, in contrast, result in the further cleavage of Ras-GAP into two proapoptotic fragments.

Future progress in computational prediction of caspase substrates, and possibly for any other protease-substrate system, will clearly hinge on the careful aggregation and integration of factors for substrate cleavage. While it is likely that certain fundamental elements of proteolysis are universally applicable across all protease-substrate systems, the intricacies and idiosyncrasies of each system – as clearly shown for the caspase substrates – necessitate deeper analysis and thought for building robust prediction models. It is also certain that such efforts will be greatly assisted as more data on caspase substrates are made available through experimental efforts. At the time of this writing, there are possibly close to 800 caspase substrates

experimentally elucidated - a clear jump in numbers when compared to those reported by Fischer *et al*. in 2003. A large part of this advance is contributed from improved high-throughput screening of caspase substrates by Mahrus *et al*. (2008) and Dix *et al*. (2008). Continual advancements in high-throughput screening of protease substrates will play an increasing role in the future for identifying novel substrates and protease-substrate relationships.

6.3 Key contributions

The original work in this thesis makes several important contributions to the fields of bioinformatics and caspase biology:

- Creation of expertly-curated datasets of caspase substrates. The datasets comprise of cleavage sites sequences and location, and other data related to the substrates. These datasets can be readily manipulated for use in information systems or for the development and testing of prediction algorithms.
- Creation of a relational database for caspase substrates (available at www.casbase.org/squery/index.html). The database allows users to retrieve data on caspase substrates using multiple query modes and provides viewing of useful annotations related to substrate cleavage.
- Development of a support vector machines based method for predicting caspase cleavage sites on protein sequences. The method was shown to perform comparably, if not better than existing computational methods. Unlike other methods, the method utilizes additional information from sequences upstream and downstream of the tetrapeptide cleavage sites. The method was integrated onto a web server which allows users to scan their sequences for potential caspase cleavage sites (available at www.casbase.org/casvm/index.html).
- Analyses on the structural characteristics of cleavage sites on caspase substrates were conducted. Caspase cleavage sites were found to preferentially locate in solvent exposed and unstructured regions on substrates.

- A two-step approach for predicting caspase substrates was developed. The algorithm comprises of first predicting the protein sequence for caspase cleavage sites followed by a filtering step where less structurally favourable cleavage sites were eliminated according to a user-defined cut-off score. The method proved to be helpful for improving the overall accuracy of predicting caspase substrates by reducing the proportion of false positives.
- A global prediction of caspase cleavage sites across all proteins in the human proteome was made. The receptor tyrosine kinase (RTK) family was analysed in detail for the occurrences of cleavage sites and consequences of caspase cleavage. Based on the analyses, the RTKs are suggested to be commonly regulated by caspase cleavage, with implications for their roles in apoptosis.

6.4 Publications

- Wee LJ, Tan TW, Ranganathan S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. Bioinformatics. 2007, 23:3241-3.
- Wee LJ, Tan TW, Ranganathan S. SVM-based prediction of caspase substrate cleavage sites. BMC Bioinformatics. 2006, Suppl 5:S14.
- Wee LJ, Tan TW, Ranganathan S. Multi-factor model for caspase degradome prediction. *Manuscript in preparation*.

Bibliography

Adamczak R, Porollo A, Meller J. **Accurate prediction of solvent accessibility using neural networks-based regression**. *Proteins* 2004, **4**:753-67.

Agniswamy J, Fang B, Weber IT. **Plasticity of S2-S4 specificity pockets of executioner caspase-7 revealed by structural and kinetic analysis**. *FEBS J* 2007, 18:4752-65.

Ahmad S, Gromiha MM, Sarai A. **Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information**. *Bioinformatics* 2004, **4**:477-86.

Alnemri ES, Livingston DJ, Nicholson DW, Salvesen G, Thornberry NA, Wong WW, Yuan J. *Cell*. 1996, **87:**171.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. **Basic local alignment search tool.** *J Mol Bio*l. 1990, **3**:403-10.

Backes C, Kuentzer J, Lenhof HP, Comtesse N, Meese E. **GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences.** *Nucleic Acids Res*. 2005, **33**:208-13.

Bae SS, Choi JH, Oh YS, Perry DK, Ryu SH, Suh PG. **Proteolytic cleavage of epidermal growth factor receptor by caspases.** *FEBS Lett*. 2001, **491**:16-20.

Barrett, A. J. *et al*. (eds) **Handbook of Proteolytic Enzymes** (Academic London, 1998).

Bellows DS, Chau BN, Lee P, Lazebnik Y, Bums WH and Hardwick JM **Antiapoptotic herpesvirus Bcl-2 homologs escape caspase-mediated conversion to proapoptotic proteins.** *J Virol*. 2000,**74**:5024-31.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol*. 2004, **4:**783-95.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL.**GenBank.** *Nucleic Acids Res*. 2008, **Database issue**:D25-30.

Bhasin M, Singh H, Raghava GP. **MHCBN: a comprehensive database of MHC binding and non-binding peptides.** *Bioinformatics*. 2003, **5**:665-6.

Bordeaux MC, Forcet C, Granger L, Corset V, Bidaud C, Billaud M et al. **The RET proto-oncogene induces apoptosis: a novel mechanism for Hirschsprung disease.** *EMBO J*. 2000, **19**:4056-63.

Bode W, Brandstetter H, Mather T, Stubbs MT. **Comparative analysis of haemostatic proteinases: structural aspects of thrombin, factor Xa, factor IXa and protein C.** *Thromb Haemost* 1997, **78**:501–511.

Boyd SE, Pike RN, Rudy GB, Whisstock JC, Garcia de la Banda M. **PoPS: a computational tool for modeling and predicting protease specificity**. *J Bioinform Comput Biol* 2005, **3**:551-85.

Bradford JR, Westhead DR: **Improved prediction of protein-protein binding sites using a support vector machines approach.** *Bioinformatics* 2005, 21: 1487-1494.

Brazas MD, Fox JA, Brown T, McMillan S, Ouellette BF. **Keeping pace with the data: 2008 update on the Bioinformatics Links Directory.** *Nucleic Acids Res*. 2008, **Web Server issue**:W2-4.

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc. Natl Acad. Sci. USA* 2000, 97: 262-267.

Brusic V, Rudy G, Kyne AP, Harrison LC. **MHCPEP, a database of MHC-binding peptides: update.** *Nucleic Acids Res*. 1996, **25**:269-71.

Burges CJC: **A tutorial on support vector machines for pattern recognition.** *Data Mining and Knowledge Discovery* 1998, 2: 121-167.

Busuttil S, Abela J, Pace GJ: **Support vector machines with profile-based kernels for remote protein homology detection.** *Genome Inform* 2004, 15: 191-200.

Byvatov E, Schneider G: **Support vector machine applications in bioinformatics.** *Appl Bioinformatics* 2003, 2: 67-77.

Cai YD, Liu XJ, Xu XB, Chou KC: **Support vector machines for predicting HIV protease cleavage sites in protein.** *J Comput Chem* 2002, 23: 267–274.

Casiano CA, Ochs RL and Tan EM. **Distinct cleavage products of nuclear proteins in apoptosis and necrosis revealed by autoantibody probes**. *Cell Death Differ* 1998, **5**: 183–190.

Cerretti DP, Kozlosky CJ, Mosley B, Nelson N, Van Ness K, Greenstreet TA, March CJ, Kronheim SR, Druck T, Cannizzaro LA et al. **Molecular cloning of the interleukin-1 beta converting enzyme**. *Science*. 1992, **256**:97-100.

Chang CC, Lin CJ: **LIBSVM: a library for support vector machines.** 2001. [http://www.csie.ntu.edu.tw/~cjlin/libsvm]

Cheng EH, Kirsch DG, Clem RJ, Ravi R, Kastan MB, Bedi A, Ueno K and Hardwick JM **Conversion of Bcl-2 to a Bax-like death effector by caspases.** *Science*. 1997, **278**:1966-68.

Chou JJ, Li H, Salvesen GS, Yuan J, Wagner G. **Solution structure of BID, an intracellular amplifier of apoptotic signaling**. *Cell* 1999, **96**: 615–624.

Clem RJ, Cheng EH, Karp CL, Kirsch DG, Ueno K, Takahashi A, Kastan MB, Griffin DE, Earnshaw WC, Veliuona MA and Hardwick JM **Modulation of cell death by Bcl-XL through caspase interaction.** *Proc. Natl. Acad. Sci. USA*. 1998, **95**:554-59.

Codd, E. F. **A Relational Model of Data for Large Shared Data Banks**. *Comm. ACM*. 1970, **13**:6.

Cortes C, Vapnik V: **Support vector networks.** *Machine Learning* 1995, 20: 273–293.

Cowan KN, Leung WC, Mar C, Bhattacharjee R, Zhu Y, Rabinovitch M. **Caspases from apoptotic myocytes degrade extracellular matrix: a novel remodeling paradigm.** *FASEB J*. 2005, **13**:1848-50.

Cui J, Han LY, Lin HH, Tang ZQ, Jiang L, Cao ZW, Chen YZ. **MHC-BPS: MHC-binder prediction server for identifying peptides of flexible lengths from sequence-derived physicochemical properties.** *Immunogenetics*. 2006, **8**:607-13.

Deak JC, Cross JV, Lewis M, Qian Y, Parrott LA, Distelhorst CW and Templeton DJ. **Fas-induced proteolytic activation and intracellular redistribution of the stress-signaling kinase MEKK1.** *Proc Natl Acad Sci USA*. 1998, **95**:5595-5600.

Degterev A, Boyce M, Yuan J. **A decade of caspases.** *Oncogene*. 2003, **53**:8543-67.

Deheuninck J, Foveau B, Goormachtigh G, Leroy C, Ji Z, Tulasne D, Fafeur V. **Caspase cleavage of the MET receptor generates an HGF interfering fragment.** *Biochem Biophys Res Commun*. 2008, **3**:573-7.

Ding CHQ, Dubchak I: **Multi-class protein fold recognition using support vector machines and neural networks.** *Bioinformatics* 2001, 17: 349–358.

Dix MM, Simon GM, Cravatt BF. **Global mapping of the topography and magnitude of proteolytic events in apoptosis.** *Cell*. 2008, **4**:679-91.

Galperin MY. **The Molecular Biology Database Collection: 2008 update.** *Nucleic Acids Res*. 2008, **Database issue**:D2-4.

Garay-Malpartida HM, Occhiucci JM, Alves J, Belizario JE. **CaSPredictor: a new computer-based tool for caspase substrate prediction.** *Bioinformatics*. 2005, Suppl **1**:i169-176.

Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. **Protein Identification and Analysis Tools on the ExPASy Server.** *In The Proteomics Protocols Handbook*. Edited by: Walker JM. Humana Press. 2005, 571-607.

Granville DJ, Carthy CM, Jiang H, Shore GC, McManus BM and Hunt DW. **Rapid cytochrome c release, activation of caspases 3, 6, 7 and 8 followed by Bap31 cleavage in HeLa cells treated with photodynamic therapy**. *FEBS Lett*. 1998, **437**: 5–10

Gschwind A, Fischer OM, Ullrich A. **The discovery of receptor tyrosine kinases: targets for cancer therapy.** *Nat Rev Cancer*. 2004, **5**:361-70.

Guillozet-Bongaarts AL, Cahill ME, Cryns VL, Reynolds MR, Berry RW, Binder LI. **Pseudophosphorylation of tau at serine 422 inhibits caspase cleavage: in vitro evidence and implications for tangle formation in vivo.** *J Neurochem*. 2006, **4**:1005-14.

Ferrando-May E, Cordes V, Biller-Ckovric I, Mirkovic J, Gorlich D and Nicotera P **Caspases mediate nucleoporin cleavage, but not early redistribution of nuclear transport factors and modulation of nuclear permeability in apoptosis**. *Cell Death Differ* 2001, **8**: 495–505.

Fischer U, Janicke RU, Schulze-Osthoff K: **Many cuts to ruin: a comprehensive update of caspase substrates.** *Cell Death Differ* 2003, **10:**76-100.

Foveau B, Leroy C, Ancot F, Deheuninck J, Ji Z, Fafeur V et al. **Amplification of apoptosis through sequential caspase cleavage of the MET tyrosine kinase receptor.** *Cell Death Differ*. 2007, **14**:752–64.

Furey TS, Cristianini N, Duffy N, Bednarski DW Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data**. *Bioinformatics* 2000, 16: 906–914.

He YY, Huang JL, Chignell CF. **Cleavage of epidermal growth factor receptor by caspase during apoptosis is independent of its internalization.** *Oncogene*. 2006, **10**:1521-31.

Hengartner MO. **The biochemistry of apoptosis**. *Nature*. 2000, **407**:770-6.

Hentze H, Schwoebel F, Lund S, Keel M, Ertel W, Wendel A, Jäättelä M, Leist M. **In vivo and in vitro evidence for extracellular caspase activity released from apoptotic cells.** *Biochem Biophys Res Commun*. 2001, **5:**1111-7.

Hua S, Sun Z: **A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach.** *J Mol Biol* 2001, 308: 397–407.

Hubbard SR, Miller WT. **Receptor tyrosine kinases: mechanisms of activation and signaling.** *Curr Opin Cell Biol*. 2007, **2**:117-23.

Jeong E, Chung IF, Miyano S. **A neural network method for identification of RNA-interacting residues in protein**. *Genome Inform* 2004, **1**:105-16.

Kanehisa M, Goto S. **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res*. 2000, **1**:27-30.

Kang SJ, Wang S, Hara H, Peterson EP, Namura S, Amin-Hanjani S, Huang Z, Srinivasan A, Tomaselli KJ, Thornberry NA, Moskowitz MA, Yuan J. **Dual role of caspase-11 in mediating activation of caspase-1 and caspase-3 under pathological conditions**. *J Cell Biol*. 2000, **149**:613-22.

Koenig U, Eckhart L, Tschachler E. **Evidence that caspase-13 is not a human but a bovine gene**. *Biochem Biophys Res Commun*. 2001, **285**:1150-4.

Kulikova et al. **EMBL Nucleotide Sequence Database in 2006.** *Nucleic Acids Res*. 2007, **Database issue**:D16-20.

Launay S, Hermine O, Fontenay M, Kroemer G, Solary E, Garrido C. **Vital functions for lethal caspases.** *Oncogene*. 2005, **33**:5137-48.

Lohmuller T, Wenzler D, Hagemann S, Kiess W, Peters C, Dandekar T, Reinheckel T: **Toward computer-based cleavage site prediction of cysteine endopeptidases**. *Biol Chem*. 2003, **384**:899–909.

López-Otín C, Overall CM. **Protease degradomics: a new challenge for proteomics**. *Nat Rev Mol Cell Biol*. 2002, **3**:509-19.

Lord SJ, Rajotte RV, Korbutt GS, Bleackley RC. **Granzyme B: a natural born killer.** *Immunol Rev*. 2003, **193**:31-8.

Los M, Stroh C, Janicke RU, Engels IH, Schulze-Osthoff K: **Caspases: more than just killers?** *Trends Immunol* 2001, **22:**31-34.

Maatta J, Hallikas O, Welti S, Hilden P, Schroder J and Kuismanen E. **Limited caspase cleavage of human BAP31**. *FEBS Lett* 2000, **484**: 202–206.

Mahrus S, Trinidad JC, Barkan DT, Sali A, Burlingame AL, Wells JA. **Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N termini.** *Cell*. 2008, 5:866-76.

Mancini M, Machamer CE, Roy S, Nicholson DW, Thomberry NA, Casciola-Rosen LA and Rosen A. **Caspase-2 is localized at the Golgi complex and cleaves golgin-160 during apoptosis**. *J. Cell Biol* 2000, **149**: 603–612.

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. **The protein kinase complement of the human genome.** *Science*. 2002, **298**:1912–34.

Mittl PR, Di Marco S, Krebs JF, Bai X, Karanewsky DS, Priestle JP, Tomaselli KJ, Grütter MG. **Structure of recombinant human CPP32 in complex with the tetrapeptide acetyl-Asp-Val-Ala-Asp fluoromethyl ketone.** *J Biol Chem*. 1997, **10**:6539-47.

Mourali J, Benard A, Lourenco FC, Monnet C, Greenland C, Moog-Lutz C et al. **Anaplastic lymphoma kinase is a dependence receptor whose proapoptotic functions are activated by caspase cleavage.** *Mol Cell Biol*. 2006, **26**:6209-22.

Ng FW, Nguyen M, Kwan T, Branton PE, Nicholson DW, Cromlish JA and Shore GC. **p28 Bap31, a Bcl-2/Bcl-XL- and procaspase-8-associated protein in the endoplasmic reticulum**. *J. Cell Biol* 1997, **139**: 327–338.

Nguyen MN, Rajapakse JC: **Two-stage multi-class support vector machines to protein secondary structure prediction.** *Pac Symp Biocomput* 2005, 346-357.

Nicholson DW. **Caspase structure, proteolytic substrates and function during apoptotic cell death**. *Cell Death Differ*. 1999, **6**:1028-42.

Overall CM, Blobel CP. **In search of partners: linking extracellular proteases to substrates.** *Nat Rev Mol Cell Biol*. 2007, **3**:245-57.

Pistritto G, Jost M, Srinivasula SM, Baffa R, Poyet JL, Kari C, Lazebnik Y, Rodeck U, Alnemri ES. **Expression and transcriptional regulation of caspase-14 in simple and complex epithelia**. *Cell Death Differ*. 2002, **9**:995-1006.

Puente, XS, Sanchez, LM, Overall, CM, Lopez-Otin, C. **Human and mouse proteases: a comparative genomic approach**. *Nat Rev Genet.* 2003, **4**:544-58.

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics*. 1999, **3-4**:213-9.

Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ. **MEROPS: the peptidase database**. *Nucleic Acids Res*. 2008, **Database issue**:D320-5.

Rechsteiner M, Rogers S: **PEST sequences and regulation by proteolysis.** *Trends Biochem Sci* 1996, 21: 267–271.

Rogers S, Wells R, Rechsteiner M: **Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis.** *Science* 1986, 234: 364–368.

Roskoski R Jr. **The ErbB/HER receptor protein-tyrosine kinases and cancer.** *Biochem Biophys Res Commun*. 2004, **319**:1-11.

Samejima K, Svingen PA, Basi GS, Kottke T, Mesner PW, Jr., Stewart L, Durrieu F, Poirier GG, Alnemri ES, Champoux JJ, Kaufmann SH and Earnshaw WC. **Caspase-mediated cleavage of DNA topoisomerase I at unconventional sites during apoptosis**. *J. Biol. Chem* 1999, **274**: 4335–4340.

Schechter I, Berger A. **On the size of the active site in proteases. I. Papain.** *Biochem Biophys Res Comm*. 1967, **18**:77–82.

Schönbach C, Kowalski-Saunders P, Brusic V. **Data warehousing in molecular biology.** *Brief Bioinform*. 2000, **2**:190-8.

Shen HB, Chou KC. Signal-3L: **A 3-layer approach for predicting signal peptides. Biochem Biophys Res Commun.** 2007, **2**:297-303.

Siegel RM. **Caspases at the crossroads of immune-cell life and death.** *Nat Rev Immunol*. 2006, **4**:308-17.

Stennicke HR, Salvesen GS. **Catalytic properties of the caspases**. *Cell Death Differ*. 1999, **6**:1054-9.

Strohecker AM, Yehiely F, Chen F, Cryns VL. **Caspase cleavage of HER-2 releases a Bad-like cell death effector**. *J Biol Chem*. 2008, **26**:18269-82.

Tan JA, Sun Y, Song J, Chen Y, Krontiris TG, Durrin LK. **SUMO conjugation to the matrix attachment region-binding protein, special AT-rich sequence-binding protein-1 (SATB1), targets SATB1 to promyelocytic nuclear bodies where it undergoes caspase cleavage.** *J Biol Chem*. 2008, **26**:18124-34.

Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T. **DNA Data Bank of Japan (DDBJ) for genome scale research in life science.** *Nucleic Acids Res*. 2002, **1**:27-30.

The UniProt Consortium. **The Universal Protein Resource (Uniprot)** *Nucleic Acids Res*. 2008, **36**:D190-D195.

Thornberry NA, Bull HG, Calaycay JR, Chapman KT, Howard AD, Kostura MJ, Miller DK, Molineaux SM, Weidner JR, Aunins J, et al. **A novel heterodimeric cysteine protease is required for interleukin-1 beta processing in monocytes**. *Nature*. 1992, **356**:768-74.

Thornberry NA, Rano TA, Peterson EP, Rasper DM, Timkey T, Garcia-Calvo M, Houtzager VM, Nordstrom PA, Roy S, Vaillancourt JP, Chapman KT, Nicholson DW. **A combinatorial approach defines specificities of members of the caspase**

family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem*. 1997, **272:**17907-11.

Tikhomirov O, Dikov M, Carpenter G. **Identification of proteolytic fragments from ErbB-2 that induce apoptosis.** *Oncogene*. 2005, **24**:3906–13.

Tikhomirov O, Carpenter G. **Caspase-dependent cleavage of ErbB-2 by geldanamycin and staurosporin.** *J Biol Chem*. 2001, **36**:33675-80.

Tikhomirov O, Carpenter G. **Identification of ErbB-2 kinase domain motifs required for geldanamycin-induced degradation.** *Cancer Res*. 2003, **1**:39-43.

Timmer JC, Salvesen GS: **Caspase substrates**. *Cell Death Differ* 2007, **1**:66-72.

Tözsér J, Bagossi P, Zahuczky G, Specht SI, Majerova E, Copeland TD. **Effect of caspase cleavage-site phosphorylation on proteolysis.** *Biochem J*. 2003, **372**:137-43.

Tulasne D, Deheuninck J, Lourenco FC, Lamballe F, Ji Z, Leroy C et al. **Proapoptotic function of the MET tyrosine kinase receptor through caspase cleavage.** *Mol Cell Biol*. 2004, **24**: 10328-39.

Tulasne D, Foveau B. **The shadow of death on the MET tyrosine kinase receptor.** *Cell Death Differ*. 2008, **3**:427-34.

Tung CW, Ho SY. **POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties.** *Bioinformatics*. 2007, **8:**942-9.

Ussat S, Werner U, Adam-Klages S. **Species-specific differences in the usage of several caspase substrates.** *Biochem Biophys Res Commun*. 2002, **5**:1186-90.

Vlahovicek K, Kajan L, Agoston V, Pongor S: **The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines**. *Nucleic Acids Res* 2005, 33, Database Issue: D223-225.

Wang YP, Biernat J, Pickhardt M, Mandelkow E, Mandelkow EM. **Stepwise proteolysis liberates tau fragments that nucleate the Alzheimer-like aggregation of full-length tau in a neuronal cell model**. *Proc Natl Acad Sci U S A*. 2007, **24**:10252-7.

Wagner M, Adamczak R, Porollo A, Meller J. **Linear regression models for solvent accessibility prediction in proteins**. *J Comput Biol* 2005, **3**:355-69.
Ward JJ, McGuffin LJ, Buxton BF, Jones DT: **Secondary structure prediction with support vector machines.** *Bioinformatics* 2003, 19: 1650–1655.

Waterhouse N, Kumar S, Song Q, Strike P, Sparrow L, Dreyfuss G, Alnemri ES, Litwack G, Lavin M and Watters D. **Heteronuclear ribonucleoproteins C1 and C2, components of the spliceosome, are specific targets of interleukin 1beta-converting enzyme-like proteases in apoptosis**. *J. Biol. Chem* 1996, **271**: 29335–29341.

Wee LJ, Tan TW, Ranganathan S. **CASVM: web server for SVM-based prediction of caspase substrates cleavage sites**. *Bioinformatics* 2007, **23**:3241-3.

Wee LJ, Tan TW, Ranganathan S. **SVM-based prediction of caspase substrate cleavage sites**. *BMC Bioinformatics*. 2006, **7** Suppl 5:S14.

Wen LP, Madani K, Martin GA and Rosen GD. **Proteolytic cleavage of ras GTPase activating protein during apoptosis.** *Cell Death Differ*. 1998, **5**:729-34.

Widmann C, Gibson S and Johnson GL. **Caspase-dependent cleavage of signaling proteins during apoptosis. A turn-off mechanism for anti-apoptotic signals.** *J Biol Chem*. 1998, **273**:7141-47.

Wu YM, Huang CL, Kung HJ, Huang CY. **Proteolytic activation of ETK/Bmx tyrosine kinase by caspases.** *J Biol Chem*. 2001, **21**:17672-8.

Yan H, Brouha B, Liu T, Raj D, Biddle D, Lee R, Grossman D: **Proteolytic cleavage of Livin (ML-IAP) in apoptotic melanoma cells potentially mediated by a non-canonical caspase.** *J Dermatol Sci* Epub 2006 Jun 27.

Yang JY and Widmann C. **Antiapoptotic signaling generated by caspase-induced cleavage of RasGAP.** *Mol. Cell. Biol*. 2001, **21**:5346–58.

Yang JY and Widmann C. **The RasGAP N-terminal fragment generated by caspase cleavage protects cells in a Ras/PI3K/Akt-dependent manner that does not rely on NFkappa B activation.** *J Biol Chem*. 2002, **277**:14641-46.

Yuan, J, Horvitz HR. **Genetic mosaic analyses of *ced-3* and *ced-4*, two genes that control programmed cell death in the nematode.** *C.elegans Dev Bio*. 1990, **138**:33-41.

Yuan J, Shaham S, Ledoux S, Ellis HM, Horvitz HR. **The C. elegans cell death gene ced-3 encodes a protein similar to mammalian interleukin-1 beta-converting enzyme**. *Cell*. 1993, **75**:641-52.

Yang ZR: **Biological applications of support vector machines.** *Brief Bioinform* 2004, 5: 328-338.
Yang ZR: **Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks.** *Bioinformatics* 2005, **21:**1831-1837.

You L. **Detection of cleavage sites for HIV-1 protease in native proteins**. *Comput Syst Bioinformatics Conf* 2006, 249-56.

Zhang SW, Pan Q, Zhang HC, Zhang YL, Wang HY: **Classification of protein quaternary structure with support vector machine.** *Bioinformatics* 2003, 19: 2390–2396.

Zhao Y, Pinilla C, Valmori D, Martin R, Simon R: **Application of support vector machines for T-cell epitopes prediction.** *Bioinformatics* 2003, 19: 1978–1984.

Appendix A

Table A-1 Fischer Dataset

Caspase Substrate	Uniprot ID	Cleavage_sites[1]
Apaf-1	O14727	SVTD (271) and a second unknown site
Bad	Q92934	Human: EQED (14) Mouse: SATD (61)
Bax	Q07812	FIQD (33)
Bcl-2	P10415	DAGD (34)
Bcl-xL	Q07817	HLAD (61), SSLD (76)
Bid	P55957	LQTD (59)
c-FLIP	O15519	LEVD (376)
c-IAP1	Q13490	ENAD (372)
XIAP	P98170	SESD (242)
APC	P25054	DNID (777)
CALM	O60641	Not reported
Cas	Q63767	DVPD (416), DSPD (748)
Beta-Catenin	P35222	SYLD (32), ADID (83), TQFD (115), YPVD (751), DLMD(764)
Gamma-Catenin	Q86W21	Not reported
Desmoglein-3	P32926	DYAD (781) and additional unknown sites
Desmocollin 3	Q14574	Not reported
Desmoplakin	P15924	Not reported
E-cadherin	P12830	DTRD (750)
N-cadherin	P19022	Not reported
P-cadherin	P22223	Putative site: ETAD (695)
FAK	Q05397	DQTD (772)
HEF1	Q14511	DLVD (363), DDYD (630)
Connexin 45.6	P36381	DEVE (367)
Paxillin	P49023	Early: NPQD (102), SQ LD (301) Late: DDLD (5), SELD (146), FPAD (165), SLLD (222)
Plakophilin-1	Q13835	Not reported
alpha-actin	P04270	Not reported
beta-actin	P60709	ELPD (244)
alpha-actinin	P12814	Not reported
alpha-adducin	P35611	DDSD (633)
CD-IC	O88485	DSGD (99) and an additional unknown site

Cortactin	Q14247	Not reported
Filamin	P21333	Not reported
alpha-II-fodrin	Q13813	DETD (1185)
beta-II-fodrin	Q01082	DEVD (1457)
Gas2	O43903	SRVD (37)
Gelsolin	P06396	DQTD (403)
HIP-55	Q6IAI8	EHID (361)
HS1	P14317	Not reported
Cytokeratin 14	P02533	Not reported
Cytokeratin 17	Q04695	VEMD (241), EVQD (416)
Cytokeratin 18	P05783	VEVD(238) -caspase-3, caspase-6 and caspase-7; DALD(397)- caspase-3 and caspase-7
Cytokeratin 19	P08727	Not reported
MHC	P35579	Not reported
vMLC	P08590	DFVE (134)
p150-Glued (dynactin)	Q14203	Not reported
Plectin	Q15149	ILRD (2395)
beta-II spectrin	Q01082	DSLD (1478), DEVD (1457), ETVD (2146)
Tau	P10636	DMVD (421)
Troponin T	P45379	VDFD (96)
alpha-tubulin	P05209	LEKD (431)
Vimentin	P08670	DSVD (85)-caspase-3, caspase-7; IDVD (259)-caspase-6, caspase-9; TNLD (429)
Emerin	O08579	Not reported
LBR	Q14739	Not reported
Lamin A	P02545	VEID (230)
Lamin B1	P20700	VEVD (231)
Lamin C	P02545-2	VEID (230)
LAP2-alpha	P42166	Putative sites: KRID (413), EERD (441), SQHD (483)
Nup153	P49790	DITD (343)
Nup214	P35658	Not reported
RanBP2/Nup358	P49792	Not reported
SAF-A	Q00839	SALD (100)
SATB1	Q01826	VEMD (254)

Tpr	P12270	Putative sites: DSQD (1892), DGTD (1999), DDED (2117), DDGD (2250), DESD (2285)
p28BAP31	P51572	AAVD (164)
Golgin-160	Q08378	ESPD (59), CSTD (139), SEVT(311)
GRASP65	Q91X51	SLLD (320), SFPD (375), TLPD (393)
Kinectin	Q86UP2	Not reported
c-Abl	P00519	Putative sites: DTTD (546), DTAD (655)
Bcr-Abl	NA	Putative sites: DTTD (546), DTAD (655)
Cdc6	Q99741	LVRD (99), SEVD (442)
CDC27	P30260	Not reported
Cyclin A	P47827	DEPD (90)
Cyclin E	P24864	Not reported
MDM2/HDM2	Q00987	DVPD (361)
MDMX	O15151	DVPD (361)
NuMA	Q14980	DSLD (1712)
p21Waf1	P38936	DHVD (112)
p27Kip1	P46527	DPSD (139), ESQD (108)
PITSLRE	P21127-2	YVPD (391)
Prothymosin-alpha	P06454	Three overlapping sites at the C-terminus: DDEDDDVD(101)
Rb	P06400	DEAD (886)
Wee1	P30291	Not reported
Acinus	Q9UKV3	DELD (1093)
ATM	Q13315	DYPD (863)
BLM	P54132	TEVD (415)
BRCA-1	P38398	DLLD (1154)
DNA-PKCs	P78527	DEVD (2713)
ICAD	O00273	DETD (117), DAVD (224)
Helicard	Q8R5F7	DNTD (208), SCTD (251)
MCM3	P25205	Not reported
PARG	Q86W56	DEID (256), MDVD (307)
PARP-1	P09874	DEVD (214)
PARP-2	O88554	LQMD (186)
Pol e catalytic subunit A	Q07864	DQLD (189), DMED (1185)

RAD21	O60216	DSPD (279)
RAD51	Q06609	DVLD (187)
RFC140	P35251	DEVD (722)
Topo I	P11387	PEDD (123)-caspase-6; DDVD (146), EEED (170) -caspase-3
Topo II-alpha	P11388	Not reported
XRCC4	Q13426	Not reported
AP-2 alpha	P05549	DRHD (19)
CREB	P16220	Putative site: ILND (140) or LSSD (144)
c-Rel	Q04864	Not reported
GAL4	P04386	Unknown C-terminal cleavage sites
GATA-1	P15976	EGLD (42), EDLD (125), LSPD(144)
HSF	Q00613	Not reported
hTAF(II)80 d	P49848	Not reported
IkBa	P25963	DRHD (32)
LEDGF	O75475	EVPD (30), WEID (85), DAQD(486)
Max	P61244	IEVE (10), SAFD (135)
MEF2A	Q02078	SSYD (466)
MEF2C	Q06413	SSYD (422)
MEF2D	Q14814	LTED (288), DHLD(291)
NF-kB p50	P19838	Not reported
NF-kB p65	Q04206	VFTD (465)
NRF2	Q16236	TEVD (208), EELD (366)
PML-RAR alpha	Q15156	PHLD (523)
RAR alpha	P10276	Not reported
Relish	Q94527	Not reported
Sp1	P08047	NSPD (590)
SREBP-1	P36956	Not reported
SREBP-2	Q12772	DEPD (468)
SRF	P11831	Not reported
STAT1	P42224	MELD (694)
BTF3	P20290	Putative site: QSVD (175)
hnRNPs A0	Q31351	Not reported

hnRNP A2/B1	P22626	KLTD (49), VMRD (55), AEVD (76), putative sites.
hnRNP A3	P51991	Not reported
hnRNP C1	P07910-2	NKTD (10), EGED (295), DDRD (298), GEDD (305), putative sites.
hnRNP C2	P07910	NKTD (10), EGED (295), DDRD (298), GEDD (305), putative sites.
hnRNP I	P26599	IVPD (7), LKTD (139), AAVD (172).
hnRNP K	P61978	Not reported
hnRNP R	O43390	RAID (66) and DYYD (472) or KESD (87) and DYHD (481), putative sites
KHSRP	Q92945	Putative sites: IRKD (72), AFAD (76), IGGD (91), STPD (102), QLED (114), EDGD (116), SQGD (128)
NONO/ p54nrb	Q15233	Putative site: MMPD (421)
NS1-associated protein1	O60506	Not reported
Nucleolin	P19338	Putative sites: TEID (455), and AMED (629) or GEID (633)
RHA	Q08211	EEVD (167)
SFRS1	Q07955	Putative sites: DLKD (139), CYAD (151), VYRD (155), RKLD (176)
SFRS9	Q13242	Putative site: GWAD (6)
SRPK1	Q96SB4	Not reported
SRPK2	P78362	Not reported
SS-B/La-autoantigen	P05455	DEHD (371) or DEHD (374)
U1-70-kDa snRNP	P08621	DGPD (341)
60S acidic ribosomal protein P0	P05388	Putative sites: PRED (5), EESD (308), SDED (310)
DAP5	P78344	DETD (792)
eIF2a	Update in progress	AEVD (301) or DGDD (304)
eIF3	Update in progress	DLAD (242), DYED (256)
eIF4B	Update in progress	DETD (45)
eIF4E-BP1	Update in progress	VLGD (25)
eIF4GI	Update in progress	DLLD (492), DRLD (1136)
eIF4GII	Update in progress	Not reported
NAC-alpha	Q13765	Not reported
PABP4	Q13310	Not reported
SRP72	O76094	SELD (614)
pro-IL-1b	P01584	YVHD (116)
pro-IL-16	Q14005	SSTD (510)
pro-IL-18	Q14116	LESD (36)

pro-EMAP-II	P31230	ASTD (144)
DCC	P43146	LSVD (1290)
EGF-R	P00533	Putative sites: DEED (1006), DMDD (1009)
ErbB-2	P04626	SETD (1125)
GluR1	Update in progress	Asp 865
GluR2	Update in progress	update in progress
GluR3	Update in progress	update in progress
GluR4	Update in progress	update in progress
RET	P07949	VSVD (707), DYLD (1017)
TCR zeta	P20963	GLLD (28) or YLLD (36), and DTYD (153)
TNF-R1	P19438	GELE (260)
GrpL/Gads	O75791	DIND (241)
TRAF1	Q13077	LEVD (163)
TRAF3	Q13114	EEAD (348), ESVD (368)
TXBP151	Q13311	Not reported
ETK/BMX	P51813	DFPD (242) and a second unknown site
Fyn	P06241	EERD (19)
Lyn	P07948	DGVD (18)
Src	P12931	Not reported
AKT	P31749	TVAD (108), EEMD (119), ECVD (462)
CaMK IIa	Q9UQM7	Not reported
CaMK IV	Q16566	YWID (35), PAPD (178)
CaMKK	Q9BQH3	Not reported
CaMKLK	Update in progress	DEND (62),
CaMKLK	Update in progress	putative DEND site at 369
CaMKLK	Update in progress	putative DEND site at 369
HPK-1	Q92918	DDVD (385)
MASK	Q9P289	DESD (305)
MEK	Q02750	Not reported
MEKK1	P53349	DTVD (874)
Mst1	Q13043	DEMD (326)
Mst2	Q13188	DELD (322)

Mst3	Q9Y6E0	AETD (325)
PAK2	Q13177	SHVD (212)
PKC delta	Q05655	DMQD (329)
PKC epsilon	Q02156	SSPD (383),
PKC epsilon	P16054	Mouse: SATD (383)
PKC eta	P24723	Unknown site in or upstream of the V3 region
PKC mu	Q15139	CQND (378)
PKC theta	Q04759	DEVD (354)
PKC zeta	Q05513	EETD (210), DGVD (239)
PKR	P19525	DLPD (251)
PRK1	Q16512	Not reported
PRK2	Q16513	DITD (117)
Raf-1	P04049	Not reported
RIP-1	Q13546	LQLD (324)
ROCK-1	Q13464	DETD (1113)
SPAK	Q9UEW8	DEMD (392)
SPAK	O88506	rat: DEMD (398)
SPAK	Q9Z1W9	Mouse: DEMD (402)
p70S6K	P23443	Not reported
FTase	P49354	VSLD (59)
GGTase I	P49354	VSLD (59)
tTG	P21980	Not reported
Calpastatin	P20810	ALDD (137), LSSD (203), ALAD (404)
Cbl	P22681	Not reported
Cbl-b	Q13191	Not reported
Nedd4	P46934	DQPD (206)
PA28-gamma	P61289	DGLD (80)
PAI-2	P05120	Not reported
UFD2	O95155	MDID (109), VDVD (123)
Cdc42	P60953	DLRD (121)
D4-GDI	P52566	DELD (19)
Rabaptin-5	Q15276	DESD (438)

Rac	P15154	DLRD (121)
Ran-GAP1	P46060	Not reported
Ras-GAP	P20936	DEGD (157), DTVD (459)
TIAM1	Q13009	DETD (993)
Vav-1	P15498	DQID (150), DLYD (161)
CCT-alpha	P49585	TEED (28)
IP3 receptor-1	P11881	DEVD (1892)
IP3 receptor-2	Q9Z329	Not reported
PIP5K-I alpha	Q99756	DIPD (279)
PDE4A5	O89084	DAVD (72)
PDE5A1	O76074	Not reported
PDE6	P16499	Putative site: DFVD (167)
PDE10A2	Q9ULW9	DLFD (333)
PDE10A3	Q9QYJ6	DLFD (315),
PMCA-2	Q01814	Putative site: EEID (1072)
PMCA-4	P23634	DEID (1080)
iPLA2	O60733	DVTD (183)
cPLA2a	P47712	DELD (336)
PLC-g1	P19174	AEPD (770)
Androgen receptor	P10275	DEDD (155)
APLP1	P51693	VEVD (620)
APP	P05067	VEVD (739)
Ataxin-3	P54252	Putative sites: LISD (145), DLPD (171), LDED (225), DEED (228)
Atrophin-1	P54259	DSLD (109)
Calsenilin	Q9Y2W7	DSSD (64)
Huntingtin	P42858	DSVD (513), DEED (530), IVLD (586)
Parkin	O60260	LHTD (126)
Presenilin-1	P49768	AQRD (345)
Presenilin-2	P49810	DSYD (329)
CrmA	P07385	LVAD (303)
M2(influenza A)	P21430	DVDD (88)
NP(influenza A and B)	P18277	Influenza A: METD (16), Influenza B: MDID (7), SEAD(61)

FEM-1	P17221	ELLD (320)
FKBP46	Q26486	Not reported
GCL	P48506	AVVD (499)
Hsp90 alpha	P07900	DEED (259)
Hsp90 beta	P08238	DEED (259)
PDC-E2	P10515	Not reported

[1] Cleavage sites are reported as tetrapeptides in the order: P_4-P_3-P_2-P_1. Numbers in parentheses following cleavage sites are location of P_1 residue on substrate.

Table A-2 Post-Fischer Dataset

Caspase Substrate	Uniprot ID	Cleavage Site(s)[1]
SATB1	Q01826	VEMD (254)
p73	O15350	TSPD (10) and at least one other cleavage site
BAG3	O95817	Not reported
HAX-1	O00165	TLRD (127)
HuR	Q15717	MGVD (226)
Nogo-B	Q9JK11	SSTD (15)
CD74	P04441	DQRD(6)
PDI	P07237	VAFD (383)
BNIP-2	Q12982	Putative: IDLD (84), DGLD (86)
BNIP-xl	Q58A63	Putative: VETD(2131), DNSD(2134)
IKK1	O15111	Not reported
NEMO	Q9Y6k9	RIED (355)
TDP-43	Q13148	Putative: DEND (13), DETD (89), DVMD (219)
EAAT2	P43006	DTID (504)
Matrin 3	P43243	DETD (680)
Mcl-1	Q07820	EELD (127), TSTD (157),
Neurocresin	Q789F1	DESD (358)
GRASP65	O35254	SLLD (319), SFPD (374), TLPD (392)
Human homolog of Ufd2p	O95155	VDVD (123) MDID (109)
NHE1 Na+/K+ exchanger	P19634	DEDD (758)
CEACAM1-L	P31809	DQRD (460)
Tensin	Q04205	DYPD (1237)
14-3-3	P62258	MQGD (238)
Sm-F	P62306	EEED (81)
PTEN	P60484	QEID (301), DVSD (371),NEPD (375), DTTD (384)
α2-spectrin	Q13813	DETD (1185)
JNK 1 β2	P45983-4	SDTD (413)
JNK 2 β2	P45984-4	SDTD (410)
SCL/Tal-1	P17542	EITD (180), SSLD (296)
SRF-N	P11831	EETD (245), SESD (254)

Rad9	Q6FI29	Putative: EEAD (187), SDTD (269) and DDID (304)
Cdc6	Q99741	SEVD (442)
Livin	Q96CA5	DHVD (52)
AP-1 complex (β-adaptin)	O35643	DLFD (701), DQPD (620)
AP-1 complex (γ-adaptin)	P22892	DMTD (746), DLLD (629)
Bim_{EL}	O43521	SECD(13)
BAT3	P46379	DEQD (1001)
Syntaxin 5	Q08851	DEQD (209)
Her-2	P04626	SETD (1125) DVFD (1087)
ERK2/MAPK	P63086	ELDD (334)
NDUSF1	P28331	DVMD (255)
Histone deacetylase 4	P56524	DVTD (289)
p23 co-chaperone	Q15185	PEVD (142), DGAD(145)
MET	P16056	ESVD (1000)
Notch1	P46531	DQTD (1840), DCMD (1874), EEED (1906), DHMD (2095), CLLD (2193)
DIAP1	Q24306	DQVD (20), VQPE (205)
CTEN	Q8IZW8	DSTD (570)
LIM-Kinase 1	P53667	DEID (240)
p65	Q04206	DCRD (97)
Ca2+ ATPase (isoform 4b)	P23634	DEID (1080)
Twist	P26687	DELD (173)
Claspin	Q9HAW4	DEYD (1072)
MITF	O75030-9	DLTD (345)

[1] Cleavage sites are reported as tetrapeptides in the order: P_4-P_3-P_2-P_1. Numbers in parentheses following cleavage sites are location of P_1 residue on substrate.

Appendix B

Table B-1 Dataset of caspase substrate cleavage sites (for cross-validation and SVM training).

Caspase Substrate	Uniprot Accession ID	Cleavage Site[1]	P_1 Position[2]
Acinus	Q9UKV3	DELD	1093
Akt	P31749	TVAD	108
		EEMD	119
		ECVD	462
α-Adducin	P35611	DDSD	633
α-II-Fodrin	Q13813	DETD	1185
Androgen Receptor	P10275	DEDD	155
AP-2 α	P05549	DRHD	19
Apaf-1	O14727	SVTD	271
APC	P25054	DNID	777
Ataxin-3	P54252	LISD	145
		LDED	225
ATM	Q13315	DYPD	863
Bad	Q92934	EQED	14
Bax	Q07812	FIQD	33
Bcl-2	P10415	DAGD	34
Bcl-xL	Q07817	HLAD	61
		SSLD	76
β-Actin	P60709	ELPD	244
β-Catenin	P35222	TQFD	115
		ADID	83
		SYLD	32
		YPVD	751
		DLMD	764
β-II Spectrin	Q01082	ETVD	2146
		DEVD	1457
Bid	P55957	LQTD	60
BLM	P54132	TEVD	415
BRCA-1	P38398	DLLD	1155
BTF3	P20290	QSVD	175
c-Abl	P00519	DTTD	546
		DTAD	655
Calcineurin	P48452	DGFD	385
Calsenilin	Q9Y2W7	DSSD	64
Cas	Q63767	DSPD	748
		DVPD	416
CCT-α	P49585	TEED	28
Cdc42	P60953	DLRD	121
Cdc6	Q99741	LVFD	99
CD-IC	O14576	DSGD	99
c-FLIP	O15519	LEVD	376
c-IAP1	Q13490	ENAD	372
Connexin 45.6	P36383	DEVE	367
CREB	P16220	ILND	140
		LSSD	144
CrmA	P07385	LVAD	303
Cyclin A2	P18606	DEPD	90
Cytokeratin 18	P05783	DALD	396
		VEVD	237
DCC	P43146	LSVD	1290
Desmoglein-3	P32926	DYAD	781
E-cadherin	P12830	DTRD	750
EGF-R	P00533	DMDD	1009
		DEED	1006
eIF2α	P05198	DGDD	303

Caspase Substrate	Uniprot Accession ID	Cleavage Site[1]	P_1 Position[2]
eIF3	O75822	DLAD	242
eIF4E-BP1	Q13541	VLGD	24
eIF4GI	Q04637	DRLD	1176
Erb-2	P04626	SETD	1125
ETK/BMX	P51813	DFPD	242
FAK	Q05397	DQTD	772
FEM-1	P17221	ELLD	320
Ftase	P49354	VSLD	59
γHSV68 Bcl-2 homolog	P89884	DCVD	31
Gas2	O43903	SRVD	278
GATA-1	P15976	LSPD	144
		EDLD	125
		EGLD	42
GCL	P48506	AVVD	498
Golgin 160	Q08378	ESPD	59
		SEVD	311
		CSTD	139
GRASP65	Q91X51	TLPD	392
		SFPD	374
GrpL/Gads	O75791	DIND	241
HEF1	Q14511	DLVD	363
		DDYD	630
Helicad	Q8R5F7	DNTD	208
		SCTD	251
HIP-55	Q6IAI8	EHID	361
hnRNP A2/B1	P22626	VMRD	55
		AEVD	76
		KLTD	49
hnRNP C1/C2	P07910	EGED	295
hnRNP I	P26599	LKTD	139
hnRNP R	O43390	DYHD	481
		KESD	87
		DYYD	472
		RAID	66
HPK-1	Q92918	DDVD	385
Huntingtin	P42858	IVLD	586
ICAD	O00273	DAVD	224
iPLA2	O60733	DVTD	183
KHSRP	Q92945	QLED	114
		EDGD	116
		IGGD	91
		AFAD	76
		SQGD	128
		IRKD	72
		STPD	102
Lamin A	P02545	VEID	230
LAP2-α	P42166	SQHD	482
		EERD	440
		KRID	412
LEDGF	O75475	DAQD	486
		WEID	85
		EVPD	30
Lyn	P07948	DGVD	17
Max	P61244	SAFD	135
Mcl-1	Q07820	EELD	127
		TSTD	157
MEF2A	Q02078	SSYD	466

Caspase Substrate	Uniprot Accession ID	Cleavage Site[1]	P_1 Position[2]
MEF2D	Q14814	LTED	288
MEKK1	P53349	DTVD	874
Mst1	Q13043	DEMD	326
Mst3	Q9Y6E0-2	AETD	313
Nedd4	P46935	DQPD	237
NF-kappa-B p65	Q04206	VFTD	465
NONO/p54nrb	Q15233	MMPD	421
NP	Q701N7	METD	16
Nucleolin	P19338	AMED	628
		TEID	454
		GEID	632
NuMA	Q14980	DSLD	1726
Nup153	P49790	DITD	349
p21Waf	P38936	DHVD	112
p27Kip1	P46527	DPSD	139
		ESQD	108
p28BAP31	P51572	AAVD	163
PA28γ	P61289	DGLD	80
PAK2	Q13177	SHVD	212
PARG	Q86W56	MDVD	307
		DEID	256
Parkin	O60260	LHTD	126
PARP-2	O88554	LQMD	187
Paxillin	Q8VI37	SELD	146
		SQLD	301
		SLLD	222
		FPAD	165
		NTQD	102
PDE10A2	Q9QYJ6	DLFD	315
PDE6	P16499	DFVD	166
PIP5K-1α	Q99756	DIPD	279
PKCδ	Q05655	DMQD	329
PKCε	Q02156	SSPD	383
PKCμ	Q15139	CQND	378
PKCζ	Q05513	DGMD	239
		EETD	210
PKR	P19525	DLPD	251
PLCγ1	P19174	AEPD	770
Plectin	Q15149	ILRD	2395
Presenilin-1	P49768	AQRD	345
Presenilin-2	P49810	DSYD	329
pro-EMAP-II	P31230	ASTD	144
pro-IL-1β	P01584	YVHD	116
pro-IL-16	Q14005	SSTD	510
pro-IL-18	Q14116	LESD	36
Rad51	Q06609	DVLD	187
Ras-GAP	P20936	DEGD	157
Rb	P06400	DEAD	886
RET	P07949	VSVD	707
		DYLD	1017
RHA	Q08211	EEVD	167
SATB1	Q01826	VEMD	254
SFRS1	Q07955	VYRD	154
		DLKD	138
		RKLD	175
		CYAD	150
SLK	O54988	DTQD	436

Caspase Substrate	Uniprot Accession ID	Cleavage Site[1]	P_1 Position[2]
Sp1	P08047	NSPD	590
SS-B/La autoantigen	P05455	DEHD	371
STAT1	P42224	MELD	694
TCRζ	P20963	YLLD	36
		GLLD	28
		DTYD	154
TNF-R1	P19438	GELE	260
Topo I	P11387	PEDD	123
		EEED	170
		DDAD	146
Tpr	P12270	DESD	2285
		DSQD	1892
		DDGD	2250
		DDED	2117
TRAF3	Q13114	ESVD	368
		EEAD	348
U1-70-kDa snRNP	P08621	DGPD	341
UFD2	O95155	VDVD	123
		MDID	109
Vav-1	P15498	DQID	150
		DLYD	161
Vimentin	P08670	TNLD	428
		IDVD	258
		DSVD	84
vMLC	P08590	DFVE	134
XIAP	P98170	SESD	242

[1] Cleavage sites are reported as tetrapeptides in the order: P4-P3-P2-P1. Except for DEVE (from connexin 45.6), GELE (fromTNF-R1) and DFVE (from vMLC), all cleavage sites have an Asp (D) in the P_1 position.
[2] Indicate the position of the P_1 amino acid in the protein sequence as reported in Uniprot.

Table B-2 Dataset of caspase substrate cleavage sites (for independent out-of-sample testing).

Caspase Substrate	Uniprot Accession ID	Cleavage Site[1]	P_1 Position[2]
14-3-3	P62258	MQGD	238
AP-1 complex (γ-adaptin)	P22892	DMTD	746
BAT3	P46379	DEQD	1001
CEACAM1-L	P31809	DQRD	460
Claspin	Q9HAW4	DEYD	1072
CTEN	Q8IZW8	DSTD	570
DIAP1	Q24306	DQVD	20
		VQPE	205
ERK2/MAPK	P63086	ELDD	334
Her-2	P04626	DVFD	1087
JNK 1 β2	P45983-4	SDTD	413
MITF	O75030-9	DLTD	345
NDUSF1 (p75 subunit of complex1)	P28331	DVMD	255
Notch1	P46531	CLLD	2193
		DCMD	1874
		DHMD	2095
p23 co-chaperone	Q15185	DGAD	145
		PEVD	142
p65	Q04206	DCRD	97
PTEN	P60484	DVSD	371
		NEPD	375
		QEID	301
Rad9	Q6FI29	DDID	304
SCL/Tal-1	P17542	EITD	180

[1] Cleavage sites are reported as tetrapeptides in the order: P_4-P_3-P_2-P_1. Except for VQPE (from DIAP1), all cleavage sites have an Asp (D) in the P_1 position.
[2] Indicates the position of the P_1 amino acid in the protein sequence as reported in Uniprot.