DOI 10.1515/hsz-2012-0143 — Biol. Chem. 2012; 393(9): 853–871

Review

Hiroyuki Sorimachi*, Hiroshi Mamitsuka and Yasuko Ono

Understanding the substrate specificity of conventional calpains

Abstract: Calpains are intracellular Ca^{2+}-dependent Cys proteases that play important roles in a wide range of biological phenomena via the limited proteolysis of their substrates. Genetic defects in calpain genes cause lethality and/or functional deficits in many organisms, including humans. Despite their biological importance, the mechanisms underlying the action of calpains, particularly of their substrate specificities, remain largely unknown. Studies show that certain sequence preferences influence calpain substrate recognition, and some properties of amino acids have been related successfully to substrate specificity and to the calpains' 3D structure. The full spectrum of this substrate specificity, however, has not been clarified using standard sequence analysis algorithms, e.g., the position-specific scoring-matrix method. More advanced bioinformatics techniques were used recently to identify the substrate specificities of calpains and to develop a predictor for calpain cleavage sites, demonstrating the potential of combining empirical data acquisition and machine learning. This review discusses the calpains' substrate specificities, introducing the benefits of bioinformatics applications. In conclusion, machine learning has led to the development of useful predictors for calpain cleavage sites, although the accuracy of the predictions still needs improvement. Machine learning has also elucidated information about the properties of calpains' substrate specificities, including a preference for sequences over secondary structures and the existence of a substrate specificity difference between two similar conventional calpains, which has never been indicated biochemically.

Keywords: calpain; multiple kernel learning; PSSM; structure-function relationship; substrate specificity; support vector machine.

***Corresponding author: Hiroyuki Sorimachi,** Calpain Project, Department of Advanced Science for Biomolecules, Tokyo Metropolitan Institute of Medical Science, Tokyo 156-8506, Japan, e-mail: sorimachi-hr@igakuken.or.jp
Hiroshi Mamitsuka: Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
Yasuko Ono: Calpain Project, Department of Advanced Science for Biomolecules, Tokyo Metropolitan Institute of Medical Science, Tokyo 156-8506, Japan

Introduction

Calpains (Clan CA, family C02; EC 3.4.22.17) are a large superfamily of intracellular Ca^{2+}-dependent Cys proteases (Goll et al., 2003; Liu et al., 2008; Sorimachi et al., 2011a,b; Ono and Sorimachi, 2012) that play pivotal roles in a wide range of biological phenomena by mediating limited proteolysis of their substrates. Thus, calpains function as proteolytic processing enzymes. This is in contrast to the major intracellular degradative proteolytic systems, consisting of eraser proteases such as proteasomes and lysosomal peptidases. The specificity of the ubiquitin/proteasome-mediated proteolysis is defined by the specific recognition and tagging of substrates by ubiquitin ligases, whereas the lysosomal peptidases generally function through autophagy, a largely non-specific degradation machinery (although specific autophagic degradations occur within certain contexts). Another major intracellular protease, caspase, shows strict specificity for Asp in P1 amino acid residues (aars). In contrast to all the above intracellular proteolytic systems, calpains show a more complex/ambiguous substrate specificity. Calpains are specific, because the same substrates are always proteolyzed at the same positions under varying conditions; however, the rules governing this specificity are not understood.

Calpains have been identified in most eukaryotes (an intriguing exception is *Schizosaccharomyces pombe*) and a few eubacteria; these homologues have a variety of domain structures and physiological roles. The most studied calpains are ubiquitous mammalian types known as μ-calpain and m-calpain, i.e., 'conventional' calpains. Each is composed of two distinct subunits: a large (~80 kDa) catalytic subunit, CAPN1 (previously called μCL or calpain-1) in μ-calpain or CAPN2 (mCL or calpain-2) in m-calpain; and a smaller (~30 kDa) regulatory subunit,

CAPNS1 (30K or CAPN4), which is common to both conventional calpains. This review refers to calpain enzymes according to their subunit composition. Thus, μ-calpain and m-calpain are referred to, respectively, as CAPN1/S1 (short for CAPN1/CAPNS1) and CAPN2/S1.

CAPN1 and CAPN2 have an identical domain structure: an N-terminal anchor-helix; protease core-domains 1 and 2 (PC1 and PC2, respectively); a C2-domain-like (C2L) domain; and a penta-EF-hand [PEF(L)] domain (Figure 1). The protease domain structure composed of PC1 and PC2 is defined as 'CysPc' (No. cd00044 in the Conserved Domain Database of the National Center for Biotechnology Information). CAPNS1 is composed of a Gly-rich (GR) domain and a penta-EF-hand [PEF(S)] domain, which is similar to a PEF(L) domain. The Ca^{2+}-binding functional domains, PC1, PC2, C2L, and PEF(L)/(S), respectively bind one, one, several, and four Ca ions.

Mammalian calpains

Using the CysPc as the defining domain for calpain homologues, 15 genes are identified in human genome (Sorimachi et al., 2011b). Other vertebrates have one or more orthologs of each human calpain species, which can be classified according to their domain structure. CAPN3 (previously called p94 or calpain-3), CAPN8 (nCL-2), CAPN9 (nCL-4), CAPN11 (μ/mCL), and CAPN12–14 are similar to CAPN1 and 2 (Figure 1), and are collectively called 'classical' calpains. The remaining large subunits [CAPN5 (hTRA-3), CAPN6, CAPN7 (PalBH), CAPN10, CAPN15 (SOLH), and CAPN16 (C6orf103)] are called 'non-classical' calpains, and are further divided into several subfamilies. CAPN5–7 and 10 are categorized as the PalB subfamily, and contain CysPc, C2L, and C2L/C2 domains [CAPN7 additionally contains a microtubule-interacting and transport (MIT) motif at the N-terminus]. CAPN15 belongs to the SOL subfamily, which contains Zn-finger motifs, CysPc, and a SOL-homology (SOH) domain; CAPN16 contains only part of the CysPc domain, i.e., PC1 but not PC2.

Expression patterns also provide good classification criteria for mammalian calpains. CAPN1, 2, 5, 7, 10, 13–16 are expressed in most tissues, whereas CAPN3 (skeletal muscle), CAPN6 (embryonic muscle and placenta), CAPN8/9 (gastrointestinal tract), CAPN11 (testis), and

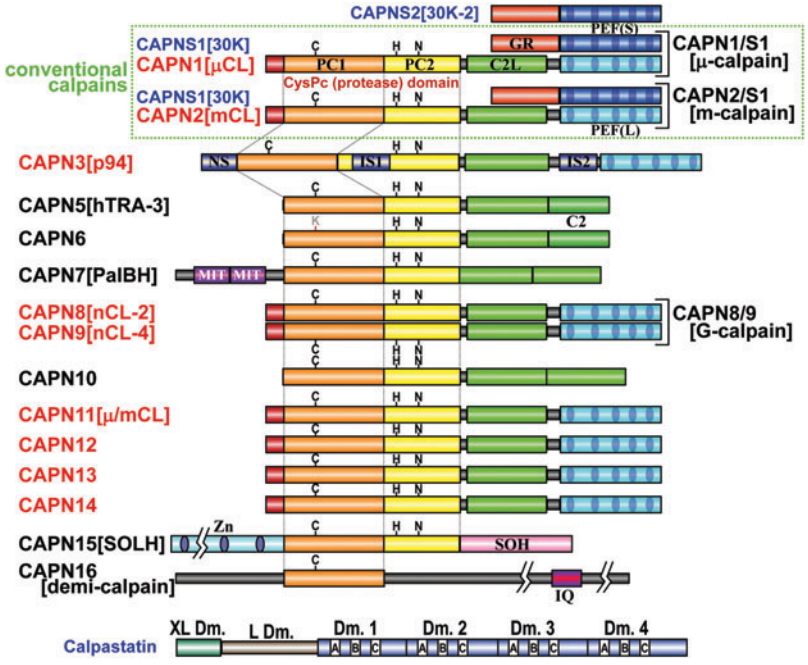


Figure 1 Schematic structures of human calpains and their associated regulatory molecules.
CAPN1–3, 8, 9, and 11–14 (in red) are considered to be classical calpains, which contain the PEF domain; the rest (in black) are nonclassical calpains containing no PEF domain. Their regulatory molecules are shown in blue (CAPNS1 and CAPNS2 are calpain regulatory subunits, and calpastatin is the endogenous specific inhibitor for calpains). The names of the calpain enzyme complexes, whose quaternary structures have been elucidated *in vivo,* are shown at right. Bottom: domain structure of calpastatin. The four repeated inhibitory units are labeled as Dm. 1–4; the A, B, and C regions of each unit are important for inhibitory activity. The consensus aa sequence in the B-region, which directly interacts with the calpain active site, is GxxE/DxTIPPxYR. Symbols: NS/IS1/IS2, CAPN3-characteristic sequences; IQ, a motif that interacts with calmodulin. See text for other symbols.

CAPN12 (hair follicles) are more tissue/organ-specific. Defects in some ubiquitous calpains cause early-stage lethality (Dutt et al., 2006; Takano et al., 2011), suggesting the importance of ubiquitous calpains in early development. By contrast, defects in tissue-specific calpains result in restricted dysfunctions like muscular dystrophy (Richard et al., 1995), indicating specialized functions of these calpain species.

Calpain substrates for *in vitro* activity assays and inhibitors

Calpains cause limited proteolysis of their substrates, mainly within inter-domain unstructured regions. Two exceptions are casein and myelin basic protein, which are proteolyzed exhaustively by calpains, and casein is the most common substrate used in *in vitro* calpain assays. Some synthetic oligopeptides, in conjunction with fluorescent probes, are also used as *in vitro* substrates (Table 1). A major problem of using these substrates is that they are not calpain-specific. For example, SLY-MCA is a good substrate for cathepsin-L-like protease (Brady et al., 2000), SLLVY-MCA is also cleaved by chymotrypsin and proteasomes (Ishiura et al., 1985), and BocLM-CMCA is cleaved by fiber cell globulizing aminopeptidase (Chandra et al., 2002). As short oligopeptides are generally poor substrates for calpains, some longer peptide substrates were developed using calpain substrate sequences to improve specificity and efficacy (see Table 1). These substrates, however, are also proteolyzed by other proteases.

Calpastatin is a highly specific endogenous proteinaceous inhibitor of CAPN1/S1 and CAPN2/S1 (both are equally susceptible). Calpastatin contains four inhibitory unit repeats that have varying inhibition efficacies (see Figure 1). Peptides (20–40 mers) corresponding to calpastatin's reactive sites are also used as calpain-specific inhibitors (Table 2). Several low-molecular-weight inhibitors of conventional calpains, such as leupeptin and E-64, have been reported, although they are much less calpain-specific than calpastatin. They also inhibit other Cys proteases, including Cys cathepsins and papain, as well as proteasomes and matrix metalloproteinase-2 (Ali et al., 2012) (see Table 2). PD150606, PD151746, and PD145305 bind PEF domains to inhibit calpains, although they are not specific for calpains (Van den Bosch et al., 2002) and are less effective than calpeptin (Gerencser et al., 2009). Thus, it is necessary to use several different inhibitors to determine whether calpains are involved in specific phenomena.

Ca^{2+} and calpain activation

Mechanistic studies on calpain activation progressed dramatically once their primary (Ohno et al., 1984) and 3D (Hosfield et al., 1999; Strobl et al., 2000) structures were determined. The latter showed that, in inactive calpain, the conformations of the PC1 and PC2 domains separate them from one another, thus maintaining the active site residues of the CysPc in a nonfunctional state.

Identifying the 3D structures of the Ca^{2+}-bound CysPc domains of CAPN1 and CAPN2 facilitated three major

Substrate	Structure and cleavage site	Commercial source	References	Note
SLY-MCA	Suc-LY-/-MCA	MERCK	Sasaki et al. (1984)	
SLLVY-MCA	Suc-LLVY-/-MCA	Peptide Institute	Sasaki et al. (1984)	
BocLM-CMCA	Boc-LM-/-CMCA	Invitrogen	Rosser et al. (1993)	Cell-permeable
KEVYGMMK	K(-ε-N-5(6)-FAM)-EVY-/-GMM-K-ε-N-4,4-Dabcyl	MERCK	Mittoo et al. (2003)	Deduced as most preferred by data mining
TPLKSPPPSPR	Dabcyl-TPLK-/-SPPPSP-R-5-EDANS	MERCK	Tompa et al. (2004)	Cleavage site sequence in α-spectrin
TPLKSPPPSPRE-R_7	Dabcyl-TPLKSPPPSPR-E(-5-EDANS)-RRRRRRR-NH_2	MERCK	Banoczi et al. (2008)	Cell-permeable version of the above
EPLFAERK	EDANS-EPLF-/-AER-K-ε-N-4,4-DABCYL	MERCK	Cuerrier et al. (2005, 2007)	Artificial sequence optimized for calpain

Table 1 Commercially available fluorescent calpain substrates.
Boc, *t*-butoxycarbonyl; CMCA, 7-amino-4-chloromethylcoumarin; Suc, succinyl; Dabcyl, dimethylamino-azobenzene-4′-carboxylic acid; EDANS, [(2-aminoethyl)amino]naphthalene-1-sulfonic acid; FAM, carboxyfluorescein; MCA, 4-methylcoumaryl-7-amide (7-amino-4-methylcoumarin). '/' indicates the cleavage site.

Inhibitor	Structure	Other targets*	Commercial source	Reference, note
Leupeptin	Ac-LL-*L*-argininal	a	Peptide Institute	Aoyagi et al. (1969)
E-64	[(2*S*, 3*S*)-3-carboxyoxirane-2-carbonyl]-L-(4-guanidinobutyl)amide	a	Peptide Institute	Hanada et al. (1978)
E-64-c	[(2*S*, 3*S*)-3-carboxyoxirane-2-carbonyl]-L-(3-methylbutyl)amide	a	Peptide Institute	Hashida et al. (1980), synthetic analog of E-64
E-64-d	[(2*S*, 3*S*)-3-ethoxycarbonyloxirane-2-carbonyl]–L-(3-methylbutyl)amide	a	Peptide Institute	Tamai et al. (1986), cell-permeable analog of E-64; also called EST or loxistatin
Calpain inhibitor I	Ac-LL-*L*-norleucinal	a, b, c	Sigma	Also called MG-101
Calpain inhibitor II	Ac-LL-*L*-methioninal	a, b	Sigma	
Calpain inhibitor III	Z-V-*L*-phenylalaninal	a, c	Sigma	Also called MDL-28170
Calpain inhibitor IV	Z-LLY-CH_2F	b	MERCK	
	Z-LL-*L*-leucinal	b	Bachem AG	Also called MG-132
Calpain inhibitor V	Morpholinoureidyl-V-homophenylalanyl-CH_2F	a	MERCK	
Calpain inhibitor VI	4-fluorophenylsulfonyl-V-*L*-leucinal	a	MERCK	
Calpain inhibitor X	Z-L-Abu-$CONHC_2H_5$	a	MERCK	
Calpain inhibitor XI	Z-L-Abu-$CONH(CH_2)_3$-morpholine	a	MERCK	
Calpain inhibitor XII	Z-L-*L*-norvaline-CONH-CH_2-2-pyridyl	a	MERCK	
Calpeptin	Z-L-*L*-norleucinal	a	MERCK	
SJA6017	N-(4-fluorophenylsulfonyl)-V-*L*-leucinal	a	Senju	Fukiage et al. (1997)
PD150606	3-(4-iodophenyl)-2-mercapto-(Z)-2-propenoic acid	c, d	MERCK	Wang et al. (1996)
PD151746	3-(5-Fluoro-3-indolyl)-2-mercapto-(Z)-2-propenoic acid	d	MERCK	
PD145305	2-mercapto-3-phenylpropanonic acid	d	MERCK	
Calpastatin peptide	(Ac-)DPMSSTYIEELGKREVTIPPKTRELLA(-NH_2)	Not known	Sigma	

Table 2 Small molecule calpain inhibitors.
Ac, acetyl; Z, benzyloxycarbonyl; CH_2F, fluoromethane; CH_2Cl, chloromethane; Abu, α-aminobutyric acid. *Including estimation from molecular structures: a, Cys proteases such as Cys cathepsins and papain; b, proteasome; c, matrix metalloproteinase-2; d, others, including nonproteolytic enzymes.

findings. First, both PC1 and PC2 have a unique Ca^{2+}-binding site (Moldoveanu et al., 2002, 2003). Second, after binding Ca^{2+}, PC1 and PC2 move closer together to form the active site. Third, the active site cleft within the CysPc domain is deeper and narrower than that of other papain-like Cys proteases (Moldoveanu et al., 2004), suggesting that the appropriate substrate conformation must be 'soft' around the cleavage site. This partly explains why calpains preferentially proteolyze interdomain unstructured regions. More recently, this activation mechanism was confirmed by determining the whole 3D structure of active CAPN2/S1 co-crystallized with calpastatin and Ca^{2+} (Hanna et al., 2008; Moldoveanu et al., 2008) (see below and Figure 3B).

A classic calpain research question asks how conventional calpains are activated *in vivo*. This question arises because *in vitro* activation of calpains requires a high [Ca^{2+}] (>10 μM), which is rare *in vivo*. The vicinity of the plasma/endosomal membranes may provide a favorable niche for calpain activation, because phospholipids, a major component of plasma membranes, lower the [Ca^{2+}] required to activate calpain *in vitro* (Saido et al., 1992; Tompa et al., 2001; Shao et al., 2006). Alternatively, a very small number of calpain molecules activated in a small region with a high local [Ca^{2+}] might suffice for physiological calpain functions. In addition, the autolysis of a few N-terminal residues and subunit dissociation during activation may have significance for *in vivo* activation.

Early studies of calpain substrate specificity

As the rules governing calpain specificity are unclear at the aa sequence level, calpains have been thought to recognize the overall 3D, rather than the primary, structures of their substrates (Sakai et al., 1987; Stabach et al., 1997). Nevertheless, some sequence preferences have been extracted by comparing the aa sequences around

the proteolytic sites in calpain substrates. Studies using various small peptide substrates revealed that the P3, P2, P1, and P1′ positions of the calpain proteolytic site were preferentially associated with F/W/L/V, L/V, R/K, and R/K/L, respectively (Ishiura et al., 1979; Hirao and Takahashi, 1984; Sasaki et al., 1984; Takahashi, 1990).

Comprehensive analyses of published calpain cleavage sites [106 (Tompa et al., 2004) and 267 (duVerle et al., 2010, duVerle et al., 2011) sites] identified a position-specific scoring matrix (PSSM) for aars around the site [Figure 2A shows a modified Sequence Logo (Crooks et al., 2004) for the most recently extended PSSM version, which was transformed to discriminate favored and disfavored).

PSSM is more informative when considered alongside the AAindex (Nakai et al., 1988), which is a database of numerical indices (Ver. 9.1: 544 criteria) representing various bio/physicochemical properties of aas so far reported. A calpain substrate PSSM was examined to determine whether a specific AAindex correlated with the aa scores (normalized frequency ratios) for each position from P30 to P30′ (544×60=32 640). Surprisingly, only 20 combinations produced a square correlation value (R^2)>0.6 (|R|>0.78), whereas 5 in P3′ and P4′ with biased values were omitted. In general, these 15 correlations show inverse associations with the hydrophobicity at P5′, P7′, and P9′, and with the propensity for a particular kind of secondary structure (SS) formation at P4′ (Table 3 and Figure 2B). These findings suggest that P5′, P7′, and P9′ prefer hydrophilic aars and that P4′ is likely to be unstructured, which indeed makes sense in the 3D structure: the closest aars to P5′, P7′, and P9′ (S5′, S7′, and S9′) in CAPN2 are the hydrophilic residues Q290, E251 (only in 3DF0) and K161, respectively, whereas a substrate bends at P4′ alongside a hydrophilic molecular wall composed of K69, K161, D162, E164, and H169 (see below, Figure 3, and Table 4). However, the lack of correlation between the AAindex and aa score in P30–P3′ may indicate that specific aars, not their attributes, are favored in these positions.

Another approach complementary to PSSM used a mixture of short oligopeptides and found that the optimum sequence (P5–P3′) for calpain substrates was PF[F(>L>P)][L(>V)][L/F]-|-[M(>A>R)]E[R(>K)], where '|' indicates the cleavage site (Cuerrier et al., 2005). This does not necessarily match the consensus sequence derived from the protein substrates shown above. In fact, surprisingly, the optimum sequences, PFFL[L/F]MER, do not exist in the eukaryote protein database. Thus, *in vivo* proteolysis of calpain substrates always occurs at sequences that are calculated to be sub/non-optimal. To allow sub/non-optimal sequences to fit within the protease core, the 3D structure as well as the primary sequence around the cleavage site may define calpain substrate specificity cooperatively. Such apparent complexity might be advantageous for controlling *in vivo* calpain activity, to slow down the hydrolysis reaction.

A limitation of these approaches is that they can detect only preferences or probabilities for any particular aar in each position. The determination of optimum sequences (and a specific aa composition required at each position) for calpains by these methods tends to lack information regarding the context of the sequence. For example, in a hypothetical case in which L-S and T-R in P2-P1 (but not L-R or T-S) are the only cleavable sequences, both of the above-mentioned approaches are likely to assume L-S, T-R, L-R, and T-S are equally favored, even though the last two are non-cleavable. Substrates occupy a specific space in the active-site cleft, so there must be contextual effects in terms of the limitations imposed by molecular size, electrostatic potential, and hydrophobicity, to name a few. This is precisely the case with human immunodeficiency virus proteases (Tozser et al., 1997). The context effect must be incorporated into any approach used to gain a better understanding of substrate specificity.

Subsite specificity of calpain based on its 3D structure

As described above, the 3D structure of active CAPN2/S1 provided important information regarding substrate binding to calpains. Notably, the preference of aa properties in P4′, P5′, P7′ and P9′ can be explained by the 3D structure of calpain. However, there is no clear relationship between the PSSM (see Figure 2A) and calpain subsites (see Figure 3B) at other positions, so it is difficult to deduce a general rule for characterizing the interface between calpain and different substrate sequences. To further explore the substrate specificities of calpains, examining 'the context effect' in the role of calpain domains other than the protease domains offers a reasonable approach. For example, the C2L domain adjacent to the CysPc domain may be crucial for substrate recognition and binding by calpains and for their substrate specificity, because the C2L domain closely contacts with calpastatin in the active CAPN2/S1 structure (see Figure 3B).

The interface between calpain and calpastatin provides a useful example for discussing the contextual effect. The calpastatin reactive site contains the consensus sequence ...**G**xx[**E**/**D**]x**TIPP**x**YR**... (I^{604}KAEHSEKL**G**-ER**D**D**TIPP**E**YR**HLL^{627} in Figure 3B; see also Figure 3C),

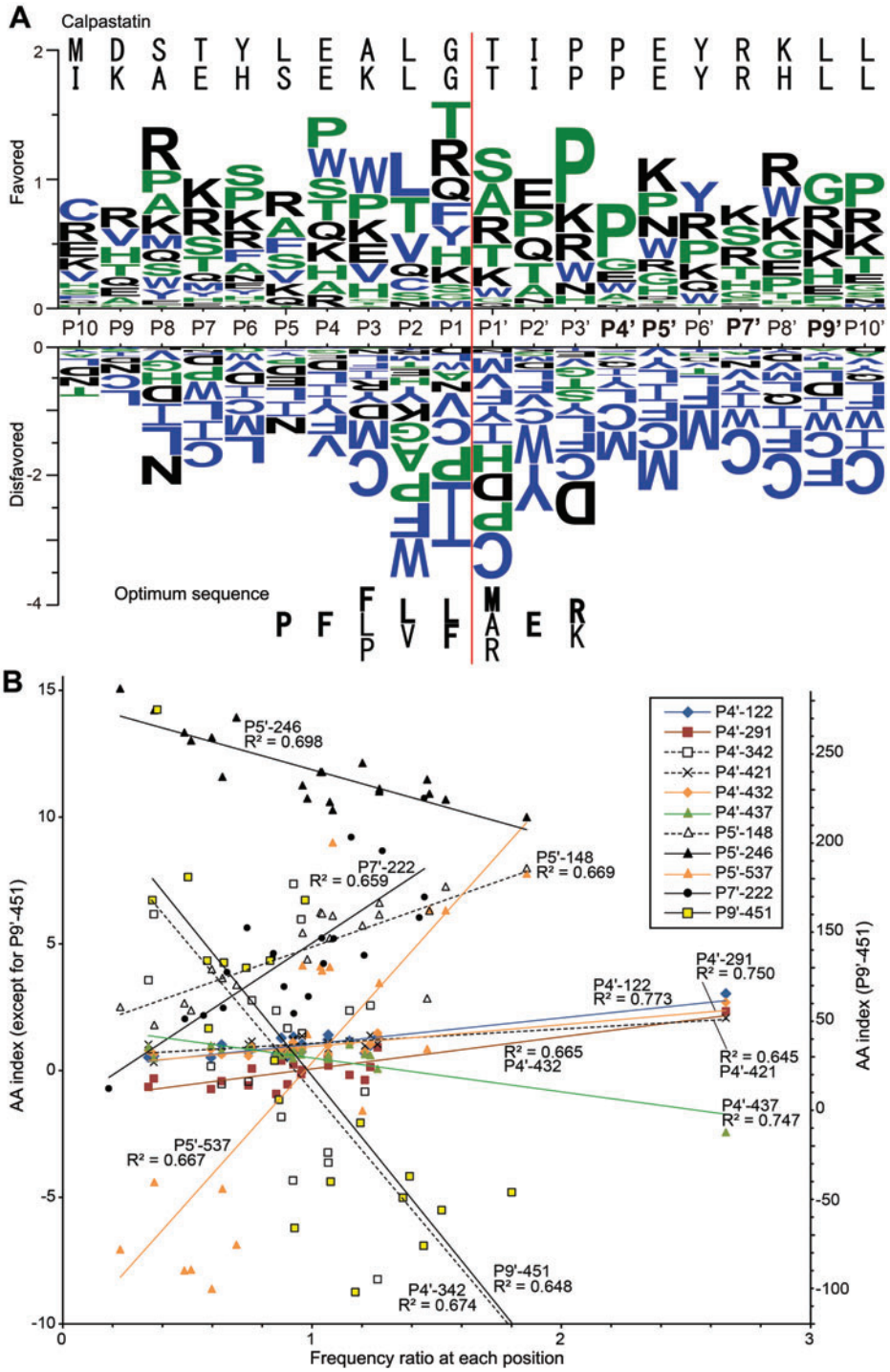


Figure 2 Calpain substrate sequence preferences.

(A) Sequence logo view of aa preferences. After aligning 367 calpain cleavage site sequences of 132 substrates (from P10 to P10′), the scores for the aa in each position were computed by dividing the occurrence ratio for each aa by the composition ratio of each aa (retrieved from UniProtKB/Swiss-Prot protein knowledgebase release 53.3 statistics), and then taking the logarithm. If the value was >0, the aa was preferred; if <0, the aa was disfavored. The values were visualized using the WebLogo program (Crooks et al., 2004). The red line indicates the calpain cleavage site. The color of an aa represents its hydrophobicity (black<green<blue). The two sequences shown at the top are calpastatin sequences M^{167}–G^{176}-/-T^{181}–L^{190} and I^{604}–G^{613}-/-T^{618}–L^{627} (positions are deduced based on the 3D structures, 3DF0 and 3BOW, respectively; see Figure 3B). The sequence at the bottom represents the optimum substrate sequence for calpain, which was determined experimentally by Cuerrier et al. (2005), in which bold aars were preferred to other aars. (B) Position-specific correlation between the preference and the amino acid properties. Each aar in each position (P30 to P30′) was converted to a value using the AAindex (Nakai et al.,

1988) (http://www.genome.jp/aaindex). *R* between the frequency ratios and AAindex values was calculated for 32 640 combinations (544 AAindex×60 positions), and those with R^2>0.6 were selected as significant. Only 15 combinations were significant, and the results for 11 nonredundant AAindex attributes are shown here: five with R^2>0.6 in P3′ and P4′ were omitted because of over-biased values. See Table 3 for the AAindex attributes. In general, residues in P4′ tend to be unstructured, and those in P5′, P7′, and P9′ tend to be hydrophilic.

in which G613 forces the next four aars in the sequence to loop-out from the calpain active site. However, the sequence T[618]IPPEYRHLL[627] binds tightly to the S1′–S10′ subsites within the PC1 and PC2 domains of CAPN2 (see Figure 3B, Table 4). G613 fits into the S1 subsite, and the sequence N-terminal to G613 (I[604]KAEHSEKL[612]) associates with the S2–S10 subsites, which extend into the C2L domain. Aars close to the bound calpastatin are highly conserved in the classical calpains, and 20 out of 24 are conserved in CAPN1 and CAPN2 (see Table 4). This strongly suggests that CAPN1/S1 and CAPN2/S1 have very similar substrate specificities, which is anticipated to be shared among other classical calpains.

As mentioned above, at least 20 aars (I604–G613 and T618–L627) of the bound calpastatin fragment are close to the surface of the calpain molecule (most aars are <3 Å from the calpain aars; see Table 4). In other words, these 20 aars of calpastatin have high affinity for the corresponding calpain subsites and exert strong and specific inhibitory activity by stabilizing the E614–D617 loop, which must have low subsite affinity, outside of the catalytic site. However, it is noteworthy that calpastatin sequences are not well conserved among species or among the four units within the calpastatin molecule (see Figure 3C). For example, calpastatin sequence alignments (data not shown) show that the aar at P10–P7 and P3, respectively, include mostly AKEE and K or [M/I/L][T/S]ST and E, and the aars at P9′ and P10′ are primarily KP, LL, or EE; other combinations rarely occur. These findings indicate that calpastatin sequences have a certain context, not just an aa preference, that influences their affinity for calpain.

Machine learning and artificial calpains

Instead of manually integrating the above observations into a law governing the structure of calpain cleavage sites, if we could generate an 'artificial calpain' *in silico* that recapitulates the proteolytic events elicited by calpains, we would be very close to understanding how

No.	AAindex	R^2	R	Position	Attribute	References
122	ISOY800104	0.773	0.879	P4′	Normalized relative frequency of bend in the first position	Isogai et al. (1980), recalculated
291	QIAN880134	0.665	0.815	P4′	Propensity to random coil structure (weights for coil at the window position of 1)	Qian and Sejnowski (1988)
342	ROBB760104	0.674	-0.821	P4′	Information measure for C-terminal helix	Robson and Suzuki (1976)
421	AURR980119	0.645	0.803	P4′	α-helix break propensity (normalized positional residue frequency at helix termini C‴)	Aurora and Rose (1998)
432	MUNV940104	0.750	0.866	P4′	Free energy required to fix in the β-strand region	Munoz and Serrano (1994)
433*	MUNV940105	0.722	0.850	P4′	Same as above	
437	BLAM930101	0.747	-0.864	P4′	α-helix propensity of position 44 in T4 lysozyme	Blaber et al. (1993)
147**	KRIW710101	0.695	0.834	P5′	Side chain interaction parameter (similar to hydrophilicity)	Krigbaum and Rubin (1971)
148	KRIW790101	0.669	0.818	P5′	Same as above	Krigbaum and Komoriya (1979)
242*	PONP800102	0.654	-0.809	P5′	Surrounding hydrophobicity	Ponnuswamy et al. (1980)
243*	PONP800103	0.682	-0.826	P5′	Same as above	
246	PONP800106	0.698	-0.835	P5′	Same as above	
537	CORJ870108	0.667	0.817	P5′	TOTLS hydrophobicity scale (multiplied by -1)	Cornette et al. (1987)
222	OOBM850105	0.659	0.812	P7′	Optimized side chain interaction parameter	Oobatake et al. (1985)
451	NADH010106	0.648	-0.805	P9′	Hydropathy scale based on self-information values in the two-state model (36% accessibility)	Naderi-Manesh et al. (2001)

Table 3 AAindices significantly correlated with specific positions of calpain substrate sequences.
*Nos. 433 and 242/243 are AAindices highly similar to 432 and 246, respectively, and were omitted in Figure 2B. **No. 148 is an updated version of 147, and, thus, 147 was omitted in Figure 2B.

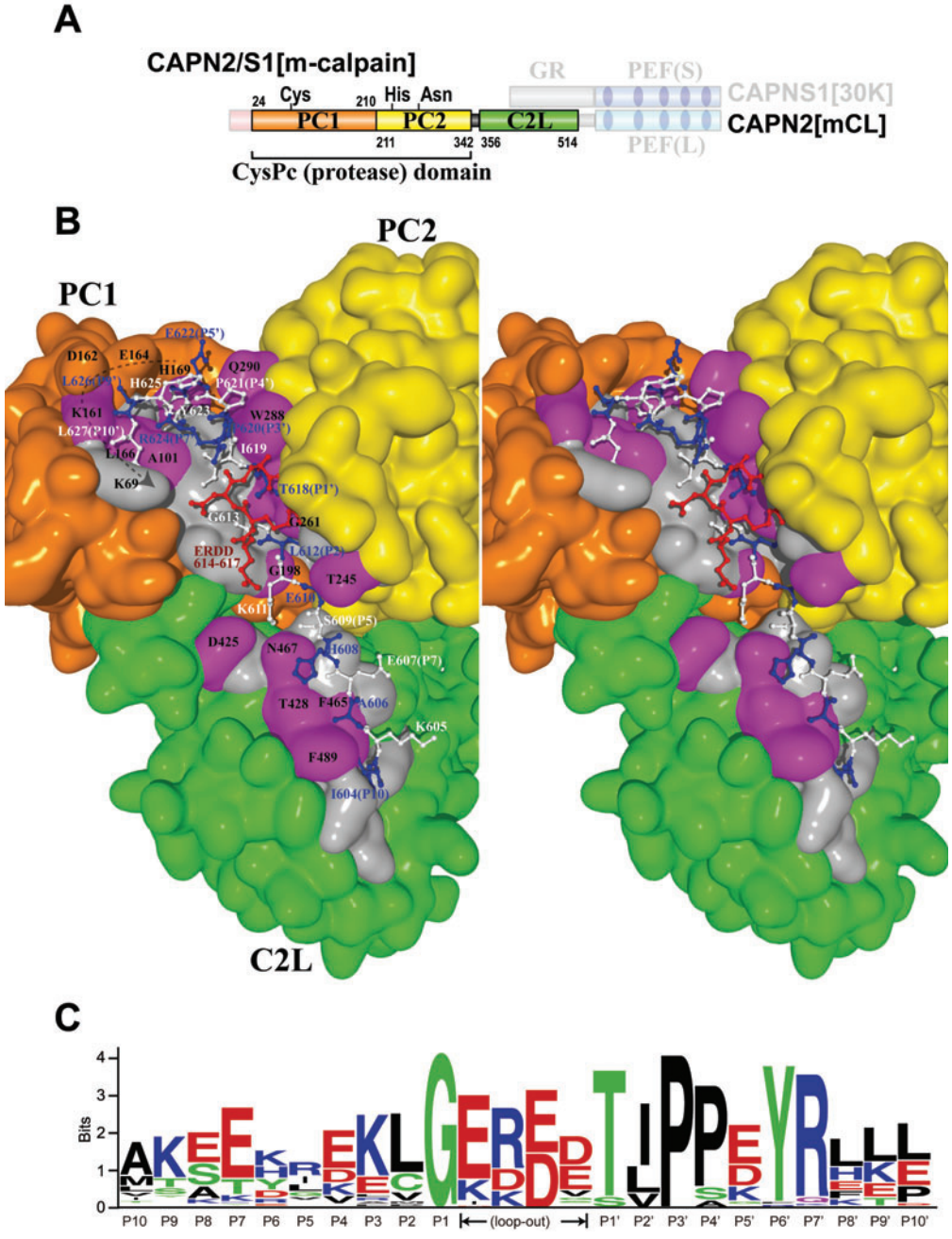


Figure 3 3D structure of calpastatin bound to active CAPN2/S1.
(A) Schematic domain structure of CAPN2/S1 (see Figure 1 for abbreviations). Numbers indicate aar numbers at the borders of the PC1, PC2, and C2L domains. (B) Cross-eyed stereo view of active (Ca^{2+}- and calpastatin-bound) rat CAPN2/S1, based on Protein DataBank account, 3BOW (Hanna et al., 2008). Only the PC1, PC2, and C2L domains are shown, in the same colors as in A. The ball and stick oligopeptide structure illustrates bound calpastatin, in which the aars corresponding to [P10, 8, 6, 4, 2, 1′, 3′, 5′, 7′, 9′], [P9, 7, 5, 3, 1, 2′, 4′, 6′, 8′, 10′], and the looped-out structure (ERDD, see Table 4) are shown in blue, white, and red, respectively, with the residue name and number (604–624) in the same color [like I604(P10) in blue; the N-terminus is at the bottom]. In CAPN2, the aar closest to each calpastatin aar, and those within 4 Å (see Table 4), are shown with pink and gray surfaces, respectively (residue numbers shown in black). (C) Sequence logo view of the conservation of calpastatin sequences. After aligning 101 calpastatin sequences [from 28 species (human ~ fishes)×1 ~ 4 units] corresponding to rat calpastatin I604–L627 (P10 to P10′), the aa conservation at each position was visualized using the WebLogo program (Crooks et al., 2004) as in Figure 2A. Here, the scores were not converted to the logarithm. The sum of the height of the aa logos at each position correlates with the disproportionate impact (bits) on the aa composition at the position, compared with the average, whereas the size of each aa logo indicates which aa is more preferred. The color of an aa represents its hydrophobicity (black<green<blue). Note that Gly and Pro at P1 and P3′ are 100% conserved (bits=4.32).

Position	3DF0[1]				3BOW[1]				CAPN2 domain
	CAST	CAPN2			CAST	CAPN2			
	aar	Proximate aar	Distance (Å)[2]	Other aars ≤4 Å	aar	Proximate aar	Distance (Å)[2]	Other aars ≤4 Å	
P10	M167	N376+	1.9	C374, R375, ***F431***, ***G432***, ***A458***, F489, R461, F465	I604	F489+	1.6	C374, R375, N376, R461, F465	C2L
P9	D168	F465++	2.3	N376, T464, F489	K605	F465++	2.8	N376, T464	C2L
P8	S169	F465+	2.1	T428, T464, F489	A606	T428+	2.0	F465, F489	C2L
P7	T170	T464+	1.7	T428, F465, I466, N467	E607	F465+	1.7	T428, T464, I466, N467	C2L
P6	Y171	N467++	3.0	I466	H608	N467++	2.6	T428, I466	C2L
P5	L172	N467++	1.9	D243, R337, I466, L468	S609	N467++	1.7	***W214***, I466, L468	C2L, PC2
P4	E173	I244+	1.8	D243, R337	E610	***T245***	1.2	D243, I244, G261, R337	PC2
P3	A174	G198+	2.6	G197	K611	***D425***	1.9	G197, G198, A199, ***E202***, ***M426***, N467, L468, ***R469***	C2L, PC1, PC2
P2	L175	G197+	1.8	G103, S105, W106, G198, A199, T200, S241, D243, G261, H262, A263, R337	L612	G198+	2.0	G103, S105, W106, G197, A199, T200, S241, G261, H262, A263, ***E339***	PC1, PC2
P1[3]	(G176)	G261++	1.9	G103, S105, G197, H262	(G613)	G261++	1.8	G103, S105, W106, ***S196***, G197, H262	PC2, PC1
P1′[3]	T181	W288++	2.0	Q99, G103, S105, V259, K260, G261, H262	T618	W288++	1.8	Q99, ***A252***, V259, K260, G261, H262	PC2, PC1
P2′	I182	A101++	2.4	Q99, G100, L102, G103, W288	I619	A101++	2.3	Q99, G100, L102, G103, W288	PC1, PC2
P3′	P183	Q99+	2.5	G100, A101, L102, W288	P620	Q290	2.3	Q99, G100, A101, W288	PC2, PC1
P4′	P184	W288++	2.6	Q290	P621	W288++	3.0	***V291***	PC2
P5′	E185	Q290++	2.9	–	E622	Q290++	3.1	–	PC2
P6′	Y186	L165+	2.1	Q99, G100, A101, E164, L166, H169	Y623	H169+	2.1	Q99, G100, A101, E164, L165, L166, Q290	PC1, PC2
P7′	R187	***E251***	3.8	–	R624	A101	4.3	–	PC1, PC2
P8′	K188	K161++	6.0	–	H625	K161++	5.4	–	PC1
P9′	L189	K161++	2.1	D162, L166	L626	K161++	2.5	A101, D162, E164, L165, L166	PC1
P10′	L190	L63+	2.3	K69, A101, K161, L166	L627	L166+	2.2	L63, K69, I73, A101, K161, ***F167***	PC1

Table 4 Calpain residues in contact with calpastatin residues.

[1]'3DF0' (Moldoveanu et al., 2008) and '3BOW' (Hanna et al., 2008) are Protein DataBank account names for 3D structures of active rat CAPN2/S1 co-crystallized with calpastatin (CAST).

[2]'Distance' indicates the distance measured between the closest atoms in the concerned residues of calpastatin and CAPN2/S1 in each of the 3D structures. [3]A looped-out structure (3DF0: $I^{177}KEG^{180}$, 3BOW: $E^{614}RDD^{617}$) is present between P1 and P1′. ++ and + at the right of CAPN2 aar indicate the proximate aar and aars ≤4 Å, respectively, which are conserved in both the 3DF0 and 3BOW 3D structures. ***Bold-italics*** indicate that an aar is unique for either structure. All proximate aars are conserved between rat CAPN1 and 2, except for T464 (Q in CAPN1), T245 (S), Q290 (E), and L166 (V) (underlined).

calpains 'assess' the 3D structure and local sequences of substrate proteins and select the appropriate sites for proteolysis. Because bioinformatics has proved fruitful for such applications, we launched construction of a prediction tool for calpain cleavage sites using the machine learning (ML) technique, support vector machine (SVM), and its recently extended version, multiple kernel learning (MKL) (duVerle et al., 2011).

ML is one of the most active research fields in computer science (Hastie et al., 2009). It began around 1980 and matured during the early 2000s. ML techniques are used for a wide range of applications in engineering and science. The procedures involved in ML are shown schematically in Figure 4 (the logic is discussed in more detail in the following sections). The essential advantage of this strategy is that the learning process is mathematically, rather than arbitrarily, refined to fit the existing empirical knowledge; hence, a good learning process could reveal novel aspects of calpain cleavage preferences.

SVM, a key concept in ML

Problem definition is the first step in applying ML, and the second is the conversion of data samples into numerical vectors usable for training the machine. The question we ask here is whether the machine can properly discriminate sequences cleavable by calpains from uncleavable ones. This type of problem is a 'classification' problem, where all given samples (e.g., sequences) have classified attributes, such as cleavable or non-cleavable; exons or introns; or Ca^{2+}-, Mg^{2+}-, or Zn^{2+}-binding. By ML, 'classification' generates a hypothesis that can categorize unknown samples into given classes (e.g., cleavable or uncleavable), and SVM is one of the most powerful techniques used for two-class classification (i.e., only two kinds exist; for example, positives and negatives) (Cristianini and Shawe-Taylor, 2000). In other words, SVM is one of the best-suited methods for predicting calpain cleavage sites.

As exemplified in Figure 4A, a set of numerical vectors that can be written as $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3, \ldots, \boldsymbol{a}_N$ (for N samples) is used for ML. For our cleavage site problem, the samples are 20 aar sequences, of which some are cleaved by calpain in the middle (= positive samples) and others are not (= negative samples). To convert these data into numerical vectors, a common bioinformatics approach would be to transform each aa into a unique integer. In this case, the 20 aars, A, C, D, ..., Y, are converted to 1, 2, 3, 4, ..., 20, respectively, and the *i*th peptide sequence can be expressed as $\boldsymbol{a}_i=(a_{i1}, a_{i2}, a_{i3}, \ldots, a_{i(n-1)}, a_{in})$ (for an *n*-mer), where a_{ij} corresponds to the *j*th aar. This transformation method also allows biochemical attributes, such as hydrophobicity, SS, and solvent accessibility (SA), to be used as numerical inputs. In Figure 4A, each sample sequence is converted to a vector consisting of 40 integers representing the aar and SS for 20 positions.

The SVM procedure can be summarized in the following two steps: (1) samples are distributed/mapped over a high-dimensional space, and (2) an optimum line [more precisely, a 'hyperplane' in a high- (>2) dimensional space] that is most distant from the positives and negatives is sought. Figuratively, such a hyperplane, described by a certain discriminant function, would correspond to a classification mechanism governing the substrate specificities of calpains. For a linear SVM (the simplest procedure), step 1 is omitted, i.e., the high-dimensional space is considered the same as the original input space; and for step 2, a linear function (or a first-order polynomial) is used as the hyperplane. For instance, when the samples are composed of *P* positives (i.e., cleavable in our case; written as ○; $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3, \ldots, \boldsymbol{b}_P$) and *Q* negatives (i.e., non-cleavable; written as ×; $\boldsymbol{b}_{P+1}, \boldsymbol{b}_{P+2}, \ldots, \boldsymbol{b}_{P+Q}$), the function can be written as $f(\boldsymbol{x})=\boldsymbol{k}\cdot\boldsymbol{x}+C=0$ (where $\boldsymbol{x}$ is a vector variable having the same dimension as $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{P+Q}$), and the actual task is to estimate $\boldsymbol{k}$ (a vector constant also having the same dimension) and *C* (a constant) using the given samples. Mathematically, this process is an iterative estimation of $\boldsymbol{k}$ and *c* that maximizes the distance ('margin') between the temporarily closest sample (called the 'support vector') and the hyperplane $f(\boldsymbol{x})=0$. Hence, this method is called SVM.

Figure 4B(i) shows an intuitive image for these processes: the optimized hyperplane is represented by the solid line (not dotted lines). As the distribution of samples becomes more complicated, as shown in Figure 4B(ii), a linear function is often not sufficient for classification. In this case, the discriminant line/hyperplane ($f(\boldsymbol{x})=0$) can be more complex, such as a second-order polynomial [called the polynomial SVM; the solid curve in Figure 4B(ii)] or a Gaussian function (the Gaussian kernel SVM).

SVM performance evaluation

To obtain the most efficient discriminant function, its validation is very important, and this is where bioinformatics also has a systematic advantage. The most popular criterion used in ML is the 'area under the ROC curve' (AUC) (ROC once stood for 'receiver operating characteristic,' but

the original meaning is no longer relevant) (Mamitsuka, 2006). The AUC is an index showing the relative efficacy of a function in solving a problem by defining perfect performance as AUC=1 and the worst performance as AUC=0.5 (see Figure 4C). Accordingly, the AUC of a given $f(\boldsymbol{x})$ is between 1 and 0.5, and the higher the AUC is, the better the discriminant function is. Details of actual evaluation procedures for discriminant functions, i.e., computing the AUC by cross-validation, are described in the legend for Figure 4C.

Using multiple vectors and kernel functions – MKL

The discriminant function, $f(\boldsymbol{x})$, is easily rewritten mathematically with a 'kernel' function, which intuitively corresponds to a function that evaluates the similarity of two vectors as variants. SVM can be regarded as a specific, simplest version of MKL, where all the information – such as aa sequences, hydrophobicity, SS, and SA – is put into a single vector and calculated by one kernel function (as described above; see Figure 4A). Each piece of information, however, may contribute differently to the prediction of substrate specificity. For example, SA information may be more important and complex (thus requiring a more complex kernel function) for the prediction of calpain cleavage sites than is SS. MKL focuses on the advantage of differentiating each information source; it puts different information into different vectors and performs classification using different kernel functions.

In MKL, each distinct kernel function, when selected appropriately for each different kind of information, can be weighted automatically according to its importance in classification performance for the given samples (see Figure 4D) (Sonnenburg et al., 2006; Gönen and Alpaydin, 2011). Therefore, one distinguishing feature of MKL is its ability to suggest the relative contribution of each information source for calculating the prediction. For example, our recent calpain cleavage prediction study (described below) using the MKL method generated a two-kernel prediction function in which 'sequence string' and 'SS' information weighed 1.0 and 0.09, respectively (duVerle et al., 2011). This result indicates that sequence information is probably more important than SS for determining calpains' cleavage site preferences. The details of MKL are omitted here due to space limitations, but MKL typically outperforms SVM (duVerle and Mamitsuka, 2011).

MKL prediction of calpain substrate cleavage

The MKL-based calpain cleavage site prediction tool is available at http://www.calpain.org, in which an AUC of 0.837 was produced when strings, SS, and SA were taken into account as independent kernel functions using 267 published calpain cleavage sites (duVerle et al., 2011). Table 5 lists some newly reported calpain cleavage sites, i.e., novel samples, which were predicted successfully using our predictor. Although the success rate was not 100%, more training samples (i.e., sequences known to be cleavable or non-cleavable by calpains) will equip the predictor with more precision and power. Thus, the current predictor provides a very good starting point, and it is expected that as the predictor is improved, it will reveal novel aspects of calpain substrate specificity. In addition, our predictor has been used in recent reports, and has provided significant information on cleavage sites (Huang et al., 2011; Arnandis et al., 2012; Kaczmarek et al., 2012). An ideal predictor, which does not yet exist, would help identify the functional consequences of substrate proteolysis by calpains.

Another intriguing feature of our MKL prediction method is that it appears to discriminate between the substrate specificity of CAPN1/S1 and CAPN2/S1. These two calpains have long been considered to have the same substrate specificity (Goll et al., 2003). However, the MKL approach generated a predictor for CAPN1/S1 with good AUCs in the range of *ca*. 14 and 10 aars in the N-terminal (left) and C-terminal (right) sides of the cleavage site, respectively (i.e., P14–P10′). However, a predictor for CAPN2/S1 showed good AUCs over a longer range, *ca*. 20 aars on both sides (P20–P20′). This result deserves our attention: The two calpains may use different areas of the molecular surface for substrate recognition, and CAPN2/S1 may recognize a wider range of substrates than CAPN1/S1. The PSSMs for each calpain were also slightly different (Figure 5), supporting the difference between the two calpains. The sizes of the data sets used for each calpain species were relatively small (~100 sites), so this result will be improved by using more data in the future.

Concluding remarks

The rationale for predicting the substrate specificities of calpains mainly on the basis of the primary structure of cleaved sequences is that the reaction conditions seldom

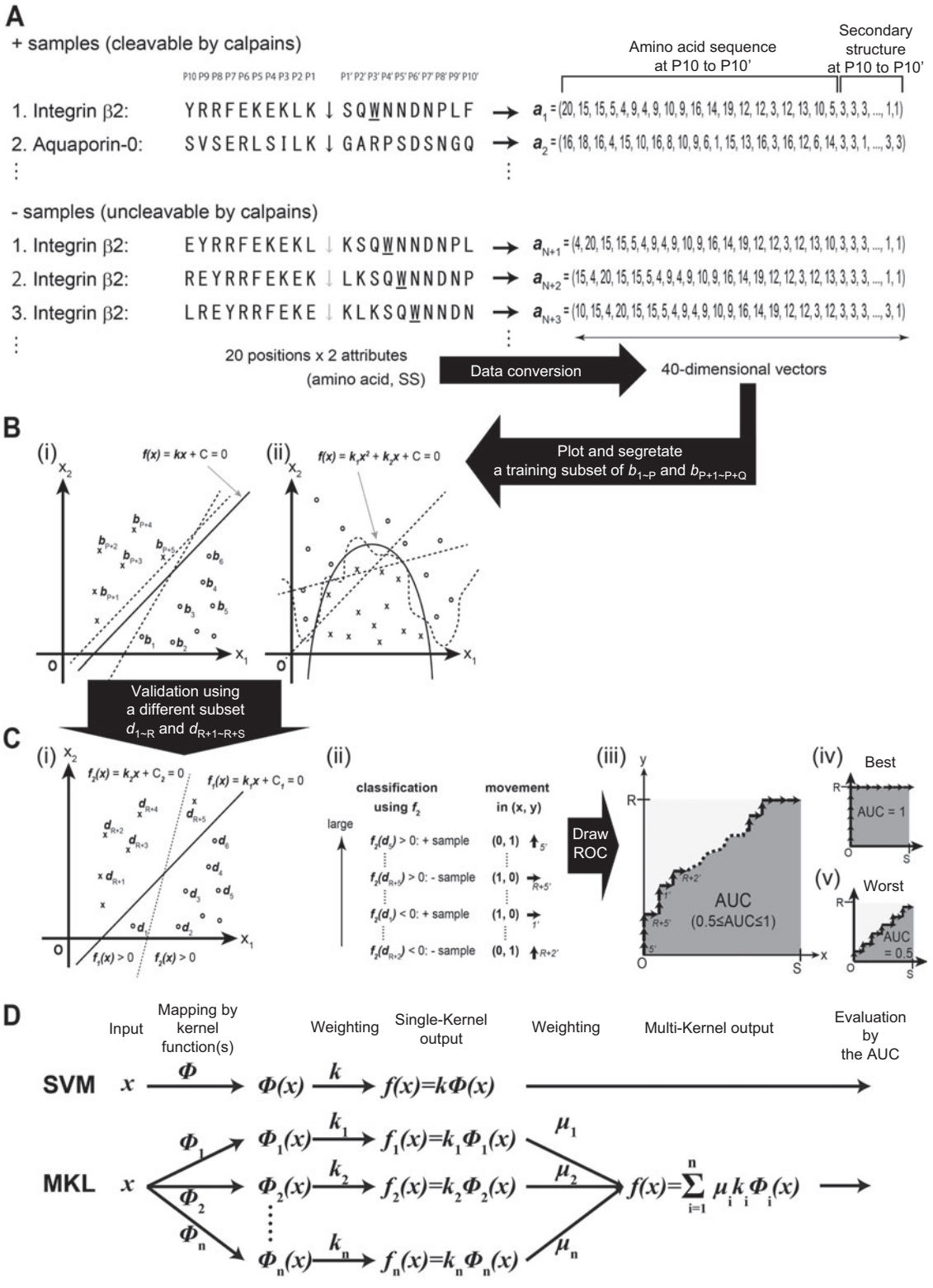


Figure 4 ML using SVM with AUC validation.

(A–C) Flow chart of ML. One starts with a learning data set, i.e., a set of known events, and uses it to define the kind of question to be asked when a novel data set is considered (A); knowledge from a learning data set is integrated into a mathematical function (B); and the problem-solving power of the function is evaluated quantitatively using the AUC where the best and the worst scores are set to 1 and 0.5, respectively (C). An effective ML procedure is one that creates a good function in B by examining various modes of data integration. In more detail, (A) illustrates the conversion of sequence data into numerical vectors. SVM uses a learning data set composed of positive (+)

samples (in this case, sequences cleaved by calpains) and several-fold negative (-) samples (noncleaved, e.g., neighboring sequences; those corresponding to the first + samples are shown; note that the underlined W residues change their relative position in each sequence extracted from Integrin β2). These data are converted into numbers [see text; 1–20 is used for an intuitive explanation; in practice, binary codes (0 and 1) are usually used instead (duVerle and Mamitsuka, 2011)]. Here, the SS information for each aar is also added as 1 (α-helix), 2 (β-sheet), or 3 (unstructured). Thus, a sample composed of 'n' aars can be expressed as a 2*n*-dimensional vector. B. A subset of + (○) and - (×) samples, i.e., $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, ..., $\boldsymbol{b}_P$, and $\boldsymbol{b}_{P+1}$, $\boldsymbol{b}_{P+2}$, ..., $\boldsymbol{b}_{P+Q}$ ($\subset\{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{N+M}\}$; containing P positives and Q negatives, P+Q<N+M), was plotted to 2*n*-dimensional coordinates (shown as a two-dimensional plane here). The linear SVM detected the $f(\boldsymbol{x})$ (i; the solid line) that maximally discriminated between ○s and ×s. In this case, the dotted lines represent nonpreferred functions, because smaller margins are available for the $\boldsymbol{b}_1$ and/or $\boldsymbol{b}_{P+5}$ points. In more complex cases, no straight line can be drawn that completely discriminates ○ from × (ii; straight dotted lines). Instead, this can be done with a parabolic curve (continuous curve), which corresponds to a second-order polynomial SVM. In addition, if a sufficiently high-dimensional curve, such as the dotted curve, is used, any finite number of samples can be completely discriminated. This is known as 'over-fitting,' which is meaningless for the prediction of unknown samples, and SVM is equipped with algorithms to avoid this. C. The validity of a discriminant function is evaluated using the AUC. When a function $f(\boldsymbol{x})$ is generated as above, its performance is examined using a novel (i.e., not used for the ML training procedure) subset of samples containing R positives and S negatives (represented by vectors $\boldsymbol{d}_1$, $\boldsymbol{d}_2$,...,$\boldsymbol{d}_R$ and $\boldsymbol{d}_{R+1}$, $\boldsymbol{d}_{R+2}$,...,$\boldsymbol{d}_{R+S}$ ($\subset\{\boldsymbol{a}_1, \boldsymbol{a}_2,...,\boldsymbol{a}_{N+M}\}$; R+S<N+M), which are subjected to classification by $f(\boldsymbol{x})$. In the ideal case, $f(\boldsymbol{x})=0$ perfectly discriminates all the R+S samples by dividing them on one side and the other of the line [i; solid line, $f_1(\boldsymbol{x})$]; all the outputs $f(\boldsymbol{d}_1)$, $f(\boldsymbol{d}_2)$,...,$f(\boldsymbol{d}_R)$ can be arbitrarily assigned a positive value, and all $f(\boldsymbol{d}_{R+1})$, $f(\boldsymbol{d}_{R+2})$,...,$f(\boldsymbol{d}_{R+S})$ a negative value. In the actual case, however, a given function sometimes fails in discriminating + and –, as shown for $f_2(\boldsymbol{x})$ ($f_2(\boldsymbol{d}_1)<0$ while $f_2(\boldsymbol{d}_{R+5})>0$, both of which are wrong by definition). To validate this, the ROC is drawn as follows: first, outputs are sorted in descending order (ii; assuming larger values are more likely to be +); second, each output is converted to a two-dimensional vector (x, y), defining movement as beginning from (0, 0), starting from the highest value. Thus, if the class label for $\boldsymbol{d}_i$ is +, i.e., $1 \le i \le R$, an upward movement (0, 1) is given, and if it is -($R+1 \le i \le R+S$), a rightward movement (1, 0) will be given (iii). The AUC is expressed as a ratio relative to the maximum (R · S). It is expected that using a perfect function, the first top R outputs all consist of + labeled samples, i.e., from $\boldsymbol{d}_1$ to $\boldsymbol{d}_R$ in random order, meaning the movement will be straight up from (0, 0) to (0, R), followed by S samples labeled -, proceeding to the final position (S, R) (iv). In the worst case, the function abandons discrimination and the + and-samples appear randomly and alternately, resulting in a ROC like the one shown in (v). Thus, the AUC of any given $f(\boldsymbol{x})$ is between 1 and 0.5. The actual evaluation is performed by a '10-fold cross-validation': that is, all given samples $\{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{N+M}\}$ are divided randomly into 10 folds named, e.g., group 1, 2, ..., 10. First, groups 1–9 $\{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_{P+Q}\}$ $\left(\mathrm{P+Q}=\frac{9(\mathrm{N+M})}{10}\right)$ are used for training and group 10 $\{\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_{R+S}\}$ $\left(\mathrm{R+S}=\frac{\mathrm{N+M}}{10}\right)$ for evaluation; second, groups 1–8 and 10 are for training and group 9 for evaluation, and so forth. This process (random division into 10 folds and 10 different calculations) is repeated 10 times with different random divisions, resulting in 10×10-fold cross-validation (total of 100 calculations). The AUC in this cross-validation is defined as the average of the values obtained from the 100 runs. The cross-validation is a standard protocol to avoid the problem of over-fitting described above. (D) Comparison of SVM and MKL. Whereas SVM uses one kernel function [$\Phi(\boldsymbol{x})$] to map an input ($\boldsymbol{x}$) to construct a discriminant function [$f(\boldsymbol{x})$], MKL uses multiple kernels and automatically determines weighting constants ($\mu_1, \mu_2, ..., \mu_n$) for corresponding kernel functions [$\Phi_1(\boldsymbol{x})$, $\Phi_2(\boldsymbol{x})$, ..., $\Phi_n(\boldsymbol{x})$].

affect the cleavage sites in calpain substrates. However, in fact, the tertiary/quaternary structures of substrates are critical for determining the accessibility of the sequence to calpain activity. That is, even if a calpain-preferred sequence is present, it cannot be cut if it is buried deep within a protein fold. Therefore, a complete understanding of calpain substrate specificities and their precise prediction requires an evaluation of the relationship between tertiary/quaternary structures and the sequence of the proteolyzed site in the substrate protein. The power of bioinformatics or ML, when used alongside conventional methodologies, has been exemplified in various fields in biology, and calpain research should benefit greatly from this trend.

In developing bioinformatics techniques, however, it is important to take into account the biological/biochemical data obtained in earlier studies. Furthermore, despite the general enthusiasm for these approaches, researchers in bioinformatics have the responsibility not to confuse theoretically possible scenarios that have little relevance to biological/biochemical properties (e.g., calculations dependent on superficial parameters) with results that illuminate real and important biological questions. With these caveats in mind, collaborations between experts in bioinformatics and biology/biochemistry hold great promise for revealing new insights into biological functions and will become increasingly important. MKL in particular has been successful in linking equations with biologically driven hypotheses, proving to be an appropriate and powerful method for elucidating sequence-related biological phenomena, including protease substrate specificity. ML has provided us with a practical predictor for calpain cleavage sites, although its accuracy is still being improved, and has shed light on the properties of calpain substrate specificities, such as their preference for sequences over secondary structures, and the discovery

Protein	Accession No.	Predicted site (MKL)		SVM (Gaussian)	SVM (linear)	PSSM	Note	References
Filaggrin	NP_001014364	1713 and 1788 (2 out of 4 sites)	$p<1.3\times10^{-4}$, and $<5.7\times10^{-3}$	–	–	–	Sites 1741 and 1771 could not be detected by any of these methods.	Hsu et al. (2011)
Rad21	NP_006256	192 (1/1)	$p<5.7\times10^{-3}$	–	–	–		Panigrahi et al. (2011)
Calcineurin	NP_058737	422 (1/4)	$p<5.7\times10^{-3}$	–	–	–	Sites 421, 423, and 425 could not be detected by any of these methods.	Wu et al. (2004)
Tau	EDM06300	120 and 380 (2/3)	$p<5.7\times10^{-3}$, and $<6.2\times10^{-4}$	–	–	–	Site 209 could not be detected by any of these methods.	Liu et al. (2011)
Vesicular GABA transporter	NP_113970	59 (1/2)	$p<5.7\times10^{-3}$	–	–	–	Site 51 could not be detected by any of these methods.	Gomes et al. (2011)
Caspase-9	NP_001220	143 (1/2)	$p<1.3\times10^{-4}$	+	+	–	Site 120 could not be detected by any of these methods.	Wolf et al. (1999)
Caspase-7	NP_001218	36 (1/1)	$p<1.3\times10^{-4}$	+	+	+		Gafni et al. (2009)
Type 1 inositol 1,4,5-triphosphate receptor	NP_001007236	1917 (1/1)	$p<6.2\times10^{-4}$	+	+	+		Kopil et al. (2011)
Transient receptor potential canonical 6	NP_038866	16 (1/1)	$p<6.2\times10^{-4}$	+	+	+		Du et al. (2010)
Paxillin*	NP_990315	– (0/1)	–	+	+	+		Cortesio et al. (2011)
Ezrin*	NP_062230	– (0/1)	–	+	+	+		Wang et al. (2008)

Table 5 Prediction of calpain cleavage sites using MKL.
Some of novel calpain cleavage sites that were not used for the original MKL predictor construction in 2011 (duVerle et al.) were analyzed. Only successful (or partially successful) results are shown. –, not predicted, +, predicted. *Sites in these two out of newly examined 28 substrates in this study were not predicted by MKL, although other methods could predict.

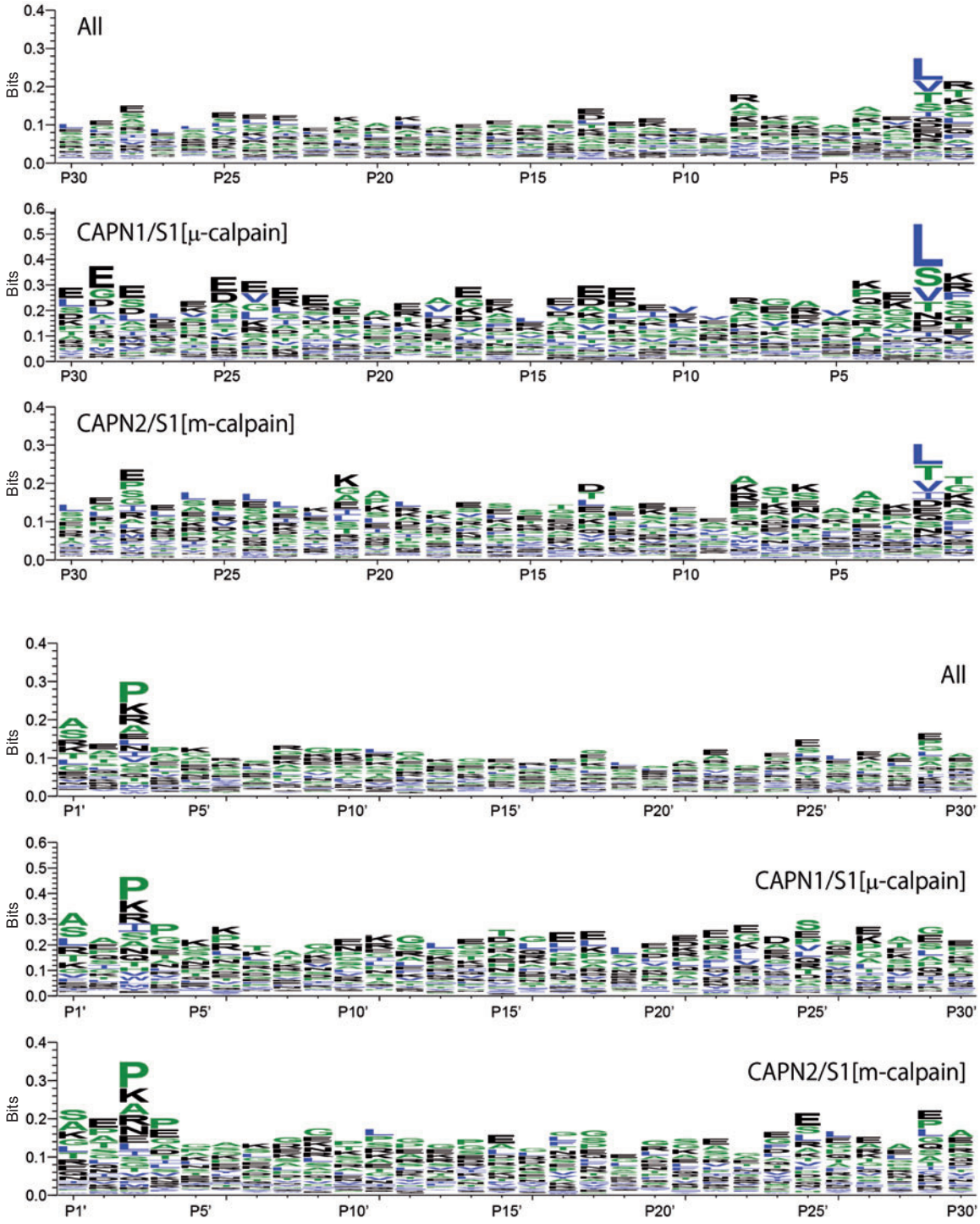


Figure 5 Comparison of the substrate specificities of CAPN1/S1 and CAPN2/S1 by PSSM.
Of the reported 367 calpain cleavage site sequences (from 132 substrate proteins) used in Figure 2, 104 and 209 sites (from 54 and 57 substrates, respectively) were results from experiments using CAPN1/S1 and CAPN2/S1, respectively. These sequences [from P30 to P1 (upper) and P1′ to P30′ (lower)] were aligned for both calpains and for each calpain separately, and their sequence logos were drawn using the WebLogo program, as in Figure 3C.

of a possible substrate specificity difference between two similar conventional calpains, which, most importantly, have never been indicated biochemically.

Acknowledgements: We thank Dr. David A. duVerle for program running, Drs. Leslie Miglietta and Grace Gray for excellent English editing, and all the Calpain Project

laboratory members for their invaluable support. This work was supported in part by JSPS.KAKENHI 20370055 and 23247021 (H.S.), and 22770139 (Y.O.), a Takeda Science Foundation research grant (H.S.), and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University [grant 2010-15 (H.S.), grants 2011-18 and 2012-30 (Y.O.)].

Received February 29, 2012; accepted July 23, 2012

References

Ali, M.A., Stepanko, A., Fan, X., Holt, A., and Schulz, R. (2012). Calpain inhibitors exhibit matrix metalloproteinase-2 inhibitory activity. Biochem. Biophys. Res. Commun. *423*, 1–5.

Aoyagi, T., Takeuchi, T., Matsuzaki, A., Kawamura, K., and Kondo, S. (1969). Leupeptins, new protease inhibitors from Actinomycetes. J. Antibiot. (Tokyo). *22*, 283–286.

Arnandis, T., Ferrer-Vicens, I., Garcia-Trevijano, E.R., Miralles, V.J., Garcia, C., Torres, L., Vina, J.R., and Zaragoza, R. (2012). Calpains mediate epithelial-cell death during mammary gland involution: mitochondria and lysosomal destabilization. Cell Death Differ.

Aurora, R. and Rose, G.D. (1998). Helix capping. Protein Sci. *7*, 21–38.

Banoczi, Z., Alexa, A., Farkas, A., Friedrich, P., and Hudecz, F. (2008). Novel cell-penetrating calpain substrate. Bioconjug. Chem. *19*, 1375–1381.

Blaber, M., Zhang, X.J., and Matthews, B.W. (1993). Structural basis of amino acid α-helix propensity. Science *260*, 1637–1640.

Brady, C.P., Brinkworth, R.I., Dalton, J.P., Dowd, A.J., Verity, C.K., and Brindley, P.J. (2000). Molecular modeling and substrate specificity of discrete cruzipain-like and cathepsin L-like cysteine proteinases of the human blood fluke *Schistosoma mansoni*. Arch. Biochem. Biophys. *380*, 46–55.

Chandra, D., Ramana, K.V., Wang, L., Christensen, B.N., Bhatnagar, A., and Srivastava, S.K. (2002). Inhibition of fiber cell globulization and hyperglycemia-induced lens opacification by aminopeptidase inhibitor bestatin. Invest. Ophthalmol. Vis. Sci. *43*, 2285–2292.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J. Mol. Biol. *195*, 659–685.

Cortesio, C.L., Boateng, L.R., Piazza, T.M., Bennin, D.A., and Huttenlocher, A. (2011). Calpain-mediated proteolysis of paxillin negatively regulates focal adhesion dynamics and cell migration. J. Biol. Chem. *286*, 9998–10006.

Cristianini, N. and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods (Cambridge, UK: Cambridge University Press).

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

Cuerrier, D., Moldoveanu, T., and Davies, P.L. (2005). Determination of peptide substrate specificity for mu-calpain by a peptide library-based approach: the importance of primed side interactions. J. Biol. Chem. *280*, 40632–40641.

Cuerrier, D., Moldoveanu, T., Campbell, R.L., Kelly, J., Yoruk, B., Verhelst, S.H., Greenbaum, D., Bogyo, M., and Davies, P.L. (2007). Development of calpain-specific inactivators by screening of positional scanning epoxide libraries. J. Biol. Chem. *282*, 9600–9611.

Du, W., Huang, J., Yao, H., Zhou, K., Duan, B., and Wang, Y. (2010). Inhibition of TRPC6 degradation suppresses ischemic brain damage in rats. J. Clin. Invest. *120*, 3480–3492.

Dutt, P., Croall, D.E., Arthur, S.C., De Veyra, T., Williams, K., Elce, J.S., and Greer, P.A. (2006). m-Calpain is required for preimplantation embryonic development in mice. BMC Dev. Biol. *6*, 3.

duVerle, D.A. and Mamitsuka, H. (2011). A review of statistical methods for prediction of proteolytic cleavage. Brief. Bioinform.

duVerle, D.A., Takigawa, I., Ono, Y., Sorimachi, H., and Mamitsuka, H. (2010). CaMPDB: a resource for calpain and modulatory proteolysis. Genome Inform. *22*, 202–213.

duVerle, D.A., Ono, Y., Sorimachi, H., and Mamitsuka, H. (2011). Calpain cleavage prediction using multiple kernel learning. PLoS One *6*, e19035.

Fukiage, C., Azuma, M., Nakamura, Y., Tamada, Y., Nakamura, M., and Shearer, T.R. (1997). SJA6017, a newly synthesized peptide aldehyde inhibitor of calpain: amelioration of cataract in cultured rat lenses. Biochim. Biophys. Acta *1361*, 304–312.

Gönen, M. and Alpaydin, E. (2011). Multiple kernel learning algorithms. J. Mach. Learn. Res. *12*, 2211–2268.

Gafni, J., Cong, X., Chen, S.F., Gibson, B.W., and Ellerby, L.M. (2009). Calpain-1 cleaves and activates caspase-7. J. Biol. Chem. *284*, 25441–25449.

Gerencser, A.A., Mark, K.A., Hubbard, A.E., Divakaruni, A.S., Mehrabian, Z., Nicholls, D.G., and Polster, B.M. (2009). Real-time visualization of cytoplasmic calpain activation and calcium deregulation in acute glutamate excitotoxicity. J. Neurochem. *110*, 990–1004.

Goll, D.E., Thompson, V.F., Li, H., Wei, W., and Cong, J. (2003). The calpain system. Physiol. Rev. *83*, 731–801.

Gomes, J.R., Lobo, A.C., Melo, C.V., Inacio, A.R., Takano, J., Iwata, N., Saido, T.C., de Almeida, L.P., Wieloch, T., and Duarte, C.B. (2011). Cleavage of the vesicular GABA transporter under excitotoxic conditions is followed by accumulation of the truncated transporter in nonsynaptic sites. J. Neurosci. *31*, 4622–4635.

Hanada, K., Tamai, M., Ohmura, S., Sawada, J., Seki, T., and Tanaka, I. (1978). Isolation and characterization of E-64, a new thiol protease inhibitor. Agric. Biol. Chem. *42*, 523–528.

Hanna, R.A., Campbell, R.L., and Davies, P.L. (2008). Calcium-bound structure of calpain and its mechanism of inhibition by calpastatin. Nature *456*, 409–412.

Hashida, S., Towatari, T., Kominami, E., and Katunuma, N. (1980). Inhibitions by E-64 derivatives of rat liver cathepsin B and cathepsin L *in vitro* and *in vivo*. J. Biochem. *88*, 1805–1811.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, 2nd edition (New York: Springer).

Hirao, T. and Takahashi, K. (1984). Purification and characterization of a calcium-activated neutral protease from monkey brain and its action on neuropeptides. J. Biochem. *96*, 775–784.

Hosfield, C.M., Elce, J.S., Davies, P.L., and Jia, Z. (1999). Crystal structure of calpain reveals the structural basis for Ca^{2+}-dependent protease activity and a novel mode of enzyme activation. EMBO J. *18*, 6880–6889.

Hsu, C.Y., Henry, J., Raymond, A.A., Mechin, M.C., Pendaries, V., Nassar, D., Hansmann, B., Balica, S., Burlet-Schiltz, O., Schmitt, A.M., et al. (2011). Deimination of human filaggrin-2 promotes its proteolysis by calpain 1. J. Biol. Chem. *286*, 23222–23233.

Huang, Z., Hoffmann, F.W., Norton, R.L., Hashimoto, A.C., and Hoffmann, P.R. (2011). Selenoprotein K is a novel target of m-calpain, and cleavage is regulated by Toll-like receptor-induced calpastatin in macrophages. J. Biol. Chem. *286*, 34830–34838.

Ishiura, S., Sugita, H., Suzuki, K., and Imahori, K. (1979). Studies of a calcium-activated neutral protease from chicken skeletal muscle. II. Substrate specificity. J. Biochem. *86*, 579–581.

Ishiura, S., Sano, M., Kamakura, K., and Sugita, H. (1985). Isolation of two forms of the high-molecular-mass serine protease, ingensin, from porcine skeletal muscle. FEBS Lett. *189*, 119–123.

Isogai, Y., Nemethy, G., Rackovsky, S., Leach, S.J., and Scheraga, H.A. (1980). Characterization of multiple bends in proteins. Biopolymers *19*, 1183–1210.

Kaczmarek, J.S., Riccio, A., and Clapham, D.E. (2012). Calpain cleaves and activates the TRPC5 channel to participate in semaphorin 3A-induced neuronal growth cone collapse. Proc. Natl. Acad. Sci. USA *109*, 7888–7892.

Kopil, C.M., Vais, H., Cheung, K.H., Siebert, A.P., Mak, D.O., Foskett, J.K., and Neumar, R.W. (2011). Calpain-cleaved type 1 inositol 1,4,5-trisphosphate receptor ($InsP_3R1$) has $InsP_3$-independent gating and disrupts intracellular Ca^{2+} homeostasis. J. Biol. Chem. *286*, 35998–36010.

Krigbaum, W.R. and Rubin, B.H. (1971). Local interactions as a structure determinant for globular proteins. Biochim. Biophys. Acta *229*, 368–383.

Krigbaum, W.R. and Komoriya, A. (1979). Local interactions as a structure determinant for protein molecules: II. Biochim. Biophys. Acta *576*, 204–248.

Liu, J., Liu, M.C., and Wang, K.K. (2008). Calpain in the CNS: from synaptic function to neurotoxicity. Sci. Signal *1*, re1.

Liu, M.C., Kobeissy, F., Zheng, W., Zhang, Z., Hayes, R.L., and Wang, K.K. (2011). Dual vulnerability of tau to calpains and caspase-3 proteolysis under neurotoxic and neurodegenerative conditions. ASN Neuro *3*, e00051.

Mamitsuka, H. (2006). Selecting features in microarray classification using ROC curves. Patt. Recog. *39*, 2393–2404.

Mittoo, S., Sundstrom, L.E., and Bradley, M. (2003). Synthesis and evaluation of fluorescent probes for the detection of calpain activity. Anal. Biochem. *319*, 234–238.

Moldoveanu, T., Hosfield, C.M., Lim, D., Elce, J.S., Jia, Z., and Davies, P.L. (2002). A Ca^{2+} switch aligns the active site of calpain. Cell *108*, 649–660.

Moldoveanu, T., Hosfield, C.M., Lim, D., Jia, Z., and Davies, P.L. (2003). Calpain silencing by a reversible intrinsic mechanism. Nat. Struct. Biol. *10*, 371–378.

Moldoveanu, T., Campbell, R.L., Cuerrier, D., and Davies, P.L. (2004). Crystal structures of calpain-E64 and -leupeptin inhibitor complexes reveal mobile loops gating the active site. J. Mol. Biol. *343*, 1313–1326.

Moldoveanu, T., Gehring, K., and Green, D.R. (2008). Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains. Nature *456*, 404–408.

Munoz, V. and Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. Proteins *20*, 301–311.

Naderi-Manesh, H., Sadeghi, M., Arab, S., and Moosavi Movahedi, A.A. (2001). Prediction of protein surface accessibility with information theory. Proteins *42*, 452–459.

Nakai, K., Kidera, A., and Kanehisa, M. (1988). Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng. *2*, 93–100.

Ohno, S., Emori, Y., Imajoh, S., Kawasaki, H., Kisaragi, M., and Suzuki, K. (1984). Evolutionary origin of a calcium-dependent protease by fusion of genes for a thiol protease and a calcium-binding protein? Nature *312*, 566–570.

Ono, Y. and Sorimachi, H. (2012). Calpains: an elaborate proteolytic system. Biochim. Biophys. Acta *1824*, 224–236.

Oobatake, M., Kubota, Y., and Ooi, T. (1985). Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteins. Bull. Inst. Chem. Res. Kyoto Univ. *63*, 82–94.

Panigrahi, A.K., Zhang, N., Mao, Q., and Pati, D. (2011). Calpain-1 cleaves Rad21 to promote sister chromatid separation. Mol. Cell. Biol. *31*, 4335–4347.

Ponnuswamy, P.K., Prabhakaran, M., and Manavalan, P. (1980). Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim. Biophys. Acta *623*, 301–316.

Qian, N. and Sejnowski, T.J. (1988). Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. *202*, 865–884.

Richard, I., Broux, O., Allamand, V., Fougerousse, F., Chiannilkulchai, N., Bourg, N., Brenguier, L., Devaud, C., Pasturaud, P., Roudaut, C., et al. (1995). Mutations in the proteolytic enzyme calpain 3 cause limb-girdle muscular dystrophy type 2A. Cell *81*, 27–40.

Robson, B. and Suzuki, E. (1976). Conformational properties of amino acid residues in globular proteins. J. Mol. Biol. *107*, 327–356.

Rosser, B.G., Powers, S.P., and Gores, G.J. (1993). Calpain activity increases in hepatocytes following addition of ATP. Demonstration by a novel fluorescent approach. J. Biol. Chem. *268*, 23593–23600.

Saido, T.C., Shibata, M., Takenawa, T., Murofushi, H., and Suzuki, K. (1992). Positive regulation of mu-calpain action by polyphosphoinositides. J. Biol. Chem. *267*, 24585–24590.

Sakai, K., Akanuma, H., Imahori, K., and Kawashima, S. (1987). A unique specificity of a calcium activated neutral protease indicated in histone hydrolysis. J. Biochem. *101*, 911–918.

Sasaki, T., Kikuchi, T., Yumoto, N., Yoshimura, N., and Murachi, T. (1984). Comparative specificity and kinetic studies on porcine calpain I and calpain II with naturally occurring peptides and synthetic fluorogenic substrates. J. Biol. Chem. *259*, 12489–12494.

Shao, H., Chou, J., Baty, C.J., Burke, N.A., Watkins, S.C., Stolz, D.B., and Wells, A. (2006). Spatial localization of m-calpain to the plasma membrane by phosphoinositide biphosphate binding during epidermal growth factor receptor-mediated activation. Mol. Cell. Biol. *26*, 5481–5496.

Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B. (2006). Large scale multiple kernel learning. J. Mach. Learn. Res. *7*, 1531–1565.

Sorimachi, H., Hata, S., and Ono, Y. (2011a). Calpain chronicle: an enzyme family under multidisciplinary characterization. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. *87*, 287–327.

Sorimachi, H., Hata, S., and Ono, Y. (2011b). Impact of genetic insights into calpain biology. J. Biochem. *150*, 23–37.

Stabach, P.R., Cianci, C.D., Glantz, S.B., Zhang, Z., and Morrow, J.S. (1997). Site-directed mutagenesis of αII spectrin at codon 1175 modulates its μ-calpain susceptibility. Biochemistry (Mosc). *36*, 57–65.

Strobl, S., Fernandez-Catalan, C., Braun, M., Huber, R., Masumoto, H., Nakagawa, K., Irie, A., Sorimachi, H., Bourenkow, G., Bartunik, H., et al. (2000). The crystal structure of calcium-free human m-calpain suggests an electrostatic switch mechanism for activation by calcium. Proc. Natl. Acad. Sci. USA *97*, 588–592.

Takahashi, K. (1990). Calpain substrate specificity. In: Intracellular calcium dependent proteolysis. R.L. Mellgren and T. Murachi, eds. (Boca Raton, FL: CRC Press), pp. 571–598.

Takano, J., Mihira, N., Fujioka, R., Hosoki, E., Chishti, A.H., and Saido, T.C. (2011). Vital role of the calpain-calpastatin system for placental-integrity-dependent embryonic survival. Mol. Cell. Biol. *31*, 4097–4106.

Tamai, M., Matsumoto, K., Omura, S., Koyama, I., Ozawa, Y., and Hanada, K. (1986). *In vitro* and *in vivo* inhibition of cysteine proteinases by EST, a new analog of E-64. J. Pharmacobiodyn. *9*, 672–677.

Tompa, P., Emori, Y., Sorimachi, H., Suzuki, K., and Friedrich, P. (2001). Domain III of calpain is a Ca^{2+}-regulated phospholipid-binding domain. Biochem. Biophys. Res. Commun. *280*, 1333–1339.

Tompa, P., Buzder-Lantos, P., Tantos, A., Farkas, A., Szilagyi, A., Banoczi, Z., Hudecz, F., and Friedrich, P. (2004). On the sequential determinants of calpain cleavage. J. Biol. Chem. *279*, 20775–20785.

Tozser, J., Bagossi, P., Weber, I.T., Louis, J.M., Copeland, T.D., and Oroszlan, S. (1997). Studies on the symmetry and sequence context dependence of the HIV-1 proteinase specificity. J. Biol. Chem. *272*, 16807–16814.

Van den Bosch, L., Van Damme, P., Vleminckx, V., Van Houtte, E., Lemmens, G., Missiaen, L., Callewaert, G., and Robberecht, W. (2002). An alpha-mercaptoacrylic acid derivative (PD150606) inhibits selective motor neuron death via inhibition of kainate-induced Ca^{2+} influx and not via calpain inhibition. Neuropharmacology *42*, 706–713.

Wang, K.K., Nath, R., Posner, A., Raser, K.J., Buroker-Kilgore, M., Hajimohammadreza, I., Probert, A.W. Jr., Marcoux, F.W., Ye, Q., Takano, E., et al. (1996). An α-mercaptoacrylic acid derivative is a selective nonpeptide cell-permeable calpain inhibitor and is neuroprotective. Proc. Natl. Acad. Sci. USA *93*, 6687–6692.

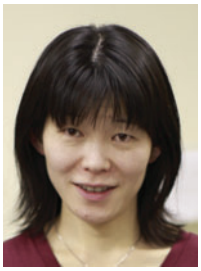
Wang, F., Xia, P., Wu, F., Wang, D., Wang, W., Ward, T., Liu, Y., Aikhionbare, F., Guo, Z., Powell, M., et al. (2008). *Helicobacter pylori* VacA disrupts apical membrane-cytoskeletal interactions in gastric parietal cells. J. Biol. Chem. *283*, 26714–26725.

Wolf, B.B., Goldstein, J.C., Stennicke, H.R., Beere, H., Amarante-Mendes, G.P., Salvesen, G.S., and Green, D.R. (1999). Calpain functions in a caspase-independent manner to promote apoptosis-like events during platelet activation. Blood *94*, 1683–1692.

Wu, H.Y., Tomizawa, K., Oda, Y., Wei, F.Y., Lu, Y.F., Matsushita, M., Li, S.T., Moriwaki, A., and Matsui, H. (2004). Critical role of calpain-mediated cleavage of calcineurin in excitotoxic neurodegeneration. J. Biol. Chem. *279*, 4929–4940.

Hiroyuki Sorimachi was born in Yokohama and started working on calpains in 1988 at the Tokyo Metropolitan Institute of Medical Science (Rinshoken) under the supervision of late Prof. Koichi Suzuki, after receiving his BSc in Biochemistry from the University of Tokyo. In 1992, he moved to the University of Tokyo as a Prof. Suzuki's assistant professor, and obtained his PhD at the same University. He continued studying calpains, and in 1997 moved to the Graduate School of Agriculture and Life Sciences at the University of Tokyo as an associate professor. In 2004, he returned to Rinshoken, which has been reorganized to Igakuken since 2011, as a project leader of Calpain Project. His research interests include biochemistry and genetics of all kinds of calpains.

Hiroshi Mamitsuka was born in a small port town near Tokyo, Japan. He obtained his BSc in Biophysics and Biochemistry from the University of Tokyo, while his PhD degree was in Information Sciences from the same university. He joined Kyoto University after working on research in machine learning and data mining with NEC Research Laboratories for more than 10 years. Currently he is a professor at the Bioinformatics Center at the Institute for Chemical Research of Kyoto University and jointly appointed at the Graduate School of Pharmaceutical Sciences of the same university. His research interest is in machine learning and data mining over graphs, particularly biological networks, and is now extended to machine learning-based systems biology.



Yasuko Ono obtained her PhD in Protein Science from the University of Tokyo in 1999. She then explored the research area focusing on muscle-specific calpain (at the University of Tokyo) and other muscle regulatory proteins (at the University of Arizona, USA) as a post-doctoral fellow supported by the Japan Society for the Promotion of Science (JSPS) and the Japan Science and Technology Agency (JST). Since 2004, she has been a researcher in Calpain Project organized by Dr. Hiroyuki Sorimachi in Rinshoken, which is now known as Tokyo Metropolitan Institute of Medical Science (Igakuken). Her research interests include biochemistry and genetics of muscle-specific calpain and PalB-type calpain, and the creation of analytical tools for calpains.