Proteome-wide prediction of protease substrates: from specificity profiling to pathways and diseases

Alexey Eroshkin*, Yoshinobu Igarashi, Ying Zhang, Boris Ratnikov, Piotr Cieplak, Andrei Osterman and Jeffrey W. Smith

Web: pmap.burnham.org
Contact: eroshkin@burnham.org

Discovery of novel protease substrates is important for our understanding of the role of proteolytic pathways and networks in norm and disease. High throughput phage display analysis allows us to accurately assess substrate preferences of various proteases at the level of amino acid sequence. This analysis performed for seven human matrix metallopeptidases (MMPs) produced a large set of cleaved peptides characterized by a range of hydrolysis scores for every protease analyzed. A significant fraction (~%30) of ~1300 compared peptides was equally cleaved by all studied MMPs (common set). Differentially cleaved peptides, analyzed by unsupervised NMF clustering approach, form three major clusters: (i) MMP25-specific, (ii) common for MMP2 and MMP9, and (iii) common for the largest group of MMP14, 15, 16 and 24. Four peptide groups were used to construct primary specificity models and to perform proteome-wide prediction of common and differential targets for the respective groups of MMPs using ProfileDB database and tools implemented within the framework of PMAP web-site (pmap.burnahm.org). A comparison of predicted substrates with physiologically relevant proteolytic events described in the literature and captured in our specialized database, CutDB (cutdb.burnham.org) revealed a substantial overlap providing a validation of the approach. A further analysis of functional associations of predicted protein substrates (captured in available pathway resources, such as IPA and GeneGo) led to interesting predictions of previously unrecognized involvement of MMPs in a variety of biological pathways and diseases.

Conclusions:

- Family-based substrate phage provides substrates recognition models
- Substrates recognition models can be used to predict protein substrates in human proteome
- Sets of predicted protein substrates can be associated with pathways; they can reveal the role which protease have in diseases

Hierarchical clustering of phage profiling data reveals highly specific substrate groups

Raw data (part)

Peptide sequence	MMP-24	MMP-16	MMP-14	MMP-15	MMP-9	MMP-2	MMP-25
AA-LTMP	0.38	0.5	0.19	0.58	0.03	0.26	1
AAWAS-I	0.46	0.68	0.36	0.93	0.01	0.9	0.78
AAWG-LN	0.46	0.68	0.36	0.93	0.26	0.07	0.87
AFH-KG	0.46	0.68	0.36	0.93	0.87	0.31	0.54
AFLSH-W	0.23	0.63	0.01	0.86	0.01	1	0.68
AFN-LYS	0.54	0.94	0.35	0.98	1	0.98	1
AFS-LMR	0.15	0.65	0.1	0.45	0.15	0.11	1
AHG-LLL	0.93	1	1	1	0.32	0.17	1
AH-LQLS	0.22	0.29	0.17	0.62	0.1	0.06	0.92
AHYAS-L	0.94	0.96	0.9	0.98	0.1	1	0.99
ALA-LLD	0.72	0.96	0.62	0.99	0.44	0.25	1
ALD-LKR	0.02	0.51	0.12	0.25	0.21	0.17	1
ALG-LLE	0.71	0.69	0.86	1	0.06	0.06	1
ALH-LRE	0.84	0.99	0.71	0.98	0.01	0.03	0.99
A-LHRPG	0.01	0.04	0.01	0.17	0.01	0.32	0.53
ALN-PYS	0.01	0.29	0.01	0.38	0.96	0.63	0.78
A-LREKV	0.37	0.32	0.18	0.44	0.01	0.01	0.89
A-LRVRG	0.93	0.92	0.51	0.94	0.01	0.34	1
AMG-LFR	0.65	0.79	0.77	0.69	0.2	0.05	0.8
ANYSN-I	0.28	0.29	0.01	0.83	0.01	0.8	0.45
APAK-LR	0.01	0.02	0.01	0.55	1	1	1
APDQ-LS	0.01	0.01	0.01	0.01	0.11	0.05	0.65
APMN-VV	0.01	0.25	0.01	0.54	1	0.88	1
APRF-LS	0.59	0.93	0.53	0.99	0.99	1	0.91
APYNH-L	0.11	0.53	0.48	0.96	0.3	0.38	0.28
ARA-MKW	0.03	0.3	0.03	0.28	0.64	0.39	0.37
ARG-IKL	0.55	0.87	0.8	0.7	0.01	0.01	0.26

1305 peptides in MMPs experiment
*Hydrolysis scores, varying from 0 to 1

B

Common peptides: 407
Well hydrolyzed by all proteases from the family (Mean > 0.5): 404 peptides

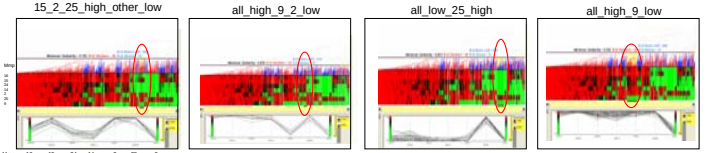
Std dev < 0.2

Std dev >= 0.2

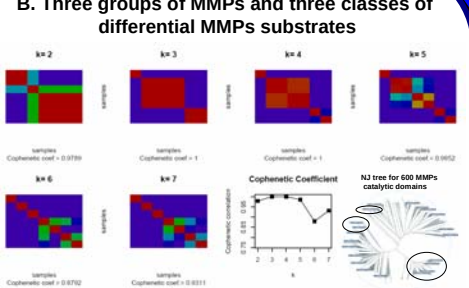
A

Differentially hydrolyzed peptides: 874
The set was submitted to consensus Nonnegative Matrix Factorization clustering (NMF) to define optimal number of protease groups

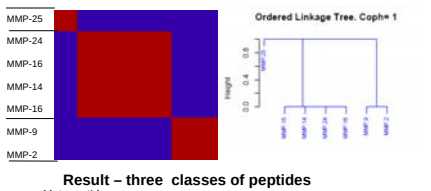
A. Hierarchical clustering of phage substrates



B. Three groups of MMPs and three classes of differential MMPs substrates

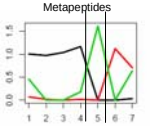


Consensus matrix for 3 factors



Result – three classes of peptides

Metapeptides

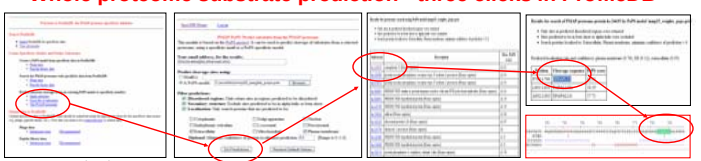


MMPs: 14 15 16 24 25 2 9

Peptide	f0 mmp25	f1 mmp14,15,16,24	f2 mmp-29	Group
AALTMP	0.581	0.51	0.055	0
AAWAS-I	0.736	0.834	0.342	1
AAWG-LN	0.775	0.771	0.005	0
AFH-LG	0.481	0.59	0.064	1
AFLSH-W	0.714	0.594	0.431	0
AFN-LYS	0.857	0.875	0.001	2
AFS-LMR	0.501	0.404	0.038	0
AHG-LLL	0.822	1.316	0.079	1
AH-LQLS	0.342	0.377	0	0

874 peptides

Whole proteome substrate prediction - three clicks in ProfileDB



1. Go to pmap.burnham.org Select tools
2. Input protease specificity Model, filtering conditions
3. Have predicted set of substrates
Optional: Analyze results in details

Filters applied
- **Protein disordered regions**: Only sites in regions predicted to be disordered
- **Secondary structure**: Exclude sites predicted to be in alpha helix or beta sheet
- **Remove "PREDICTED: HYPOTHETICAL"**: Exclude sequences without annotation
- **Cell localization**: Only proteins that are predicted to be:
  - –Extracellular or in plasma membrane (MultiLoc predictor)
  - –Minimum confidence of protein localization prediction: 0.2
  - –Contain signal peptide or/and transmembrane sequence (SignalP + TMHMM)

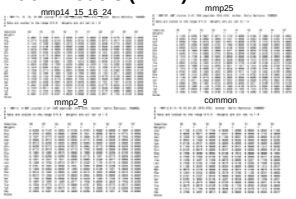
http://protease.burnham.org/www/tools/cgi-bin/specdb/ProfileDB.html

Developing and validating substrate recognition models (PWM)

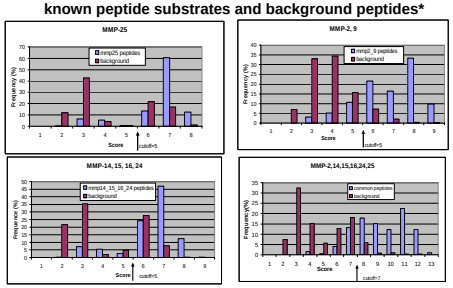
From four sets of peptides to four position weight matrices (PWM)*

- mmp14_15_16_24
- mmp25
- mmp2_9
- common

* - One click job after submission peptides to ProfileDB



Substrate recognition performance evaluated by comparison of known peptide substrates and background peptides*



* Phage display prior to selection. Score cutoff selected to optimize separation of target peptides and background

Predicted MMPs substrates (part)

Known

common

RefSeq ID	GI	PMAP	Name	Score
NP_443177	16445027	hs.20266	BAFF receptor [Homo sapiens]	12.61
NP_055859	27722337	hs.66446	syndecan 2 [Homo sapiens]	12.26
NP_940962	113612068	hs.10242	PREDICTED: similar to syndecan 3 isoform 2 [Homo sapiens]	12.26
NP_004159	55749599	hs.97760	collagen, type XIV, alpha 1 [Homo sapiens]	12.05
NP_060033	56711322	hs.3122	ring finger protein 43 [Homo sapiens]	11.94
NP_000080	4502901	hs.18308	alpha 1 type VII collagen precursor [Homo sapiens]	11.78
NP_071376	62988224	hs.05344	tenascin N [Homo sapiens]	11.76
NP_004370	181410049	hs.18436	collagen, type XII, alpha 1 short isoform precursor [Homo sapiens]	11.75
NP_004361	62414347	hs.16306	collagen, type XII, alpha 1 long isoform precursor [Homo sapiens]	11.75
NP_001129562	113419802	hs.50350	PREDICTED: similar to Sclae carrier family 23 member 1 [Sodium-c	11.75
NP_001129387	113419802	hs.50225	PREDICTED: similar to Sclae carrier family 23 member 1 [Sodium-c	11.75
NP_001129563	113419571	hs.17161	PREDICTED: similar to Sclae carrier family 23 member 1 [Sodium-c	11.75
NP_056238	17653808	hs.09304	neuronal pentraxin receptor [Homo sapiens]	11.65
NP_001128	4502901	hs.18308	alpha 1 type VII collagen precursor [Homo sapiens]	11.63
NP_116586	17511265	hs.17868	G protein-coupled receptor 124 [Homo sapiens]	11.63
NP_056834	110735405	hs.28961	collagen, type V, alpha 3 preproprotein [Homo sapiens]	11.62
NP_005414	4055176	hs.12498	histamine receptor H2 [Homo sapiens]	11.62
NP_001017718	58741022	hs.16275	X Kell blood group precursor-related family, member 7 [Homo sapien	11.57
NP_008838	24797907	hs.11316	pyrroline-5-carboxylate reductase 1 isoform 1 [Homo sapiens]	11.53
NP_005264	7656643	hs.05383	G protein-coupled receptor 162 isoform 1 [Homo sapiens]	11.52
NP_002832	9951613	hs.31465	G protein-coupled receptor 162 isoform 2 [Homo sapiens]	11.52
NP_937995	38146119	hs.34914	hypothetical protein LOC310712 isoform 5 [Homo sapiens]	11.5

mmp2_9

RefSeq ID	GI	PMAP	Name	Score
NP_443177	16445027	hs.22065	BAFF receptor [Homo sapiens]	9.688
NP_055859	27722337	hs.66446	syndecan 2 [Homo sapiens]	9.433
NP_940962	113612068	hs.10242	PREDICTED: similar to syndecan 3 isoform 2 [Homo sapiens]	9.237
NP_004159	55749599	hs.97760	collagen, type XIV, alpha 1 [Homo sapiens]	9.237
NP_004361	62414347	hs.16306	collagen, type XII, alpha 1 long isoform precursor [Ho	9.237
NP_056834	110735405	hs.28964	collagen, type V, alpha 3 preproprotein [Homo sapiens	9.201
NP_001848	17402875	hs.01369	alpha 2 type VI collagen isoform 2C2 precursor [Homo	9.148
NP_007716	18041347	hs.04549	putative ermin gene [Homo sapiens]	9.098
NP_937960	38146119	hs.34914	hypothetical protein LOC310712 isoform 5 [Homo sapi	9.098
NP_001129562	113419802	hs.50350	PREDICTED: similar to Sclae carrier family 23 member	9.098
NP_071438	68660910	hs.16494	PREDICTED: similar to FAT tumor suppressor homol	9.098
NP_056499	57202187	hs.06645	syndecan 2 [Homo sapiens]	8.995
NP_945760	113412068	hs.10242	PREDICTED: similar to syndecan 3 isoform 2 [Homo	8.995
NP_004199	4757460	hs.27123	cholinergic receptor, nicotinic, alpha polypeptide 6 [Ho	8.917
NP_001127210	113419840	hs.07121	PREDICTED: similar to RAI14 [Homo sapiens]	8.917
NP_002567	4927102	hs.18633	contactin 2 precursor [Homo sapiens]	8.892
NP_056834	110735405	hs.28961	collagen, type V, alpha 3 preproprotein [Homo sapiens]	8.864
NP_116586	17511265	hs.17868	G protein-coupled receptor 124 [Homo sapiens]	8.799
NP_005413	4005782	hs.12888	histamine receptor H2 [Homo sapiens]	8.737
NP_056822	40716277	hs.04174	transmembrane protein 24 [Homo sapiens]	8.711
NP_005579	28019597	hs.35479	glutamate receptor KA2 precursor [Homo sapiens]	8.711
NP_071376	62988224	hs.05344	tenascin N [Homo sapiens]	8.66

mmp2_9

- 5 of the top 10 scores are known substrates.
- ~40 reported substrates predicted.
- Except for collagens, there are 210 substrates with scores higher than the first documented MMP2 and 3 substrate, serum amyloid A1

Substrates for MMPs	Number of predicted substrates
Common	958
MMP2_and_9	2254
MMP14_15_16_2	1410
MMP25	1559

mmp14, 15, 16, 24

RefSeq ID	GI	PMAP	Name	Score
NP_001003682	51262137	hs.32951	hypothetical protein LOC389474 [Homo sapiens]	7.481
NP_057561	31717239	hs.02242	interleukin 22 receptor, alpha 1 [Homo sapiens]	7.414
NP_009163	6005782	hs.12888	histamine receptor H2 [Homo sapiens]	7.377
NP_060864	45048807	hs.14698	kinesin 1 [Homo sapiens]	7.338
NP_443177	16445027	hs.22065	BAFF receptor [Homo sapiens]	7.338
NP_997264	76769432	hs.33200	FRAS1 related extracellular matrix protein 2 [Homo sapiens]	7.338
NP_071436	62339432	hs.22203	hypothetical protein LOC64115 [Homo sapiens]	7.325
NP_598921	53832009	hs.20132	calcium channel, voltage-dependent, alpha 1H subunit isoform	7.325
NP_001005407	53832011	hs.21807	calcium channel, voltage-dependent, alpha 1H subunit isoform	7.325
NP_004779	4758146	hs.13958	eph homolog 2 [Homo sapiens]	7.3
NP_008856	4504537	hs.14042	S-hydroxynaphthalene (serotonin) receptor 1G [Homo sapiens]	7.3
NP_598910	55763566	hs.67743	collagen, type XIV, alpha 1 [Homo sapiens]	7.274
NP_114095	13994224	hs.11042	enamelin [Homo sapiens]	7.222
NP_997264	76470832	hs.33273	natural cytotoxicity triggering receptor 3 [Homo sapiens]	7.209
NP_079094	56769358	hs.21394	thrombospondin repeat containing 1 isoform 2 [Homo sapiens]	7.17
NP_056534	110735405	hs.28961	collagen, type V, alpha 3 preproprotein [Homo sapiens]	7.132
NP_775765	27731149	hs.11157	hypothetical protein LOC153129 [Homo sapiens]	7.132
NP_001005407	53832011	hs.21807	calcium channel, voltage-dependent, alpha 1H subunit isoform	7.132
NP_598864	16863961	hs.01878	potassium channel, subfamily K, member 10 isoform 1 [Homo	7.106
NP_001019599	112799532	hs.04018	poliovirus receptor related 2 isoform delta precursor [Homo sap	7.091
NP_002847	5360210	hs.04192	poliovirus receptor related 2 isoform alpha precursor [Homo sa	7.091
NP_803863	29284617	hs.05726	sweet taste receptor T1r isoform a [Homo sapiens]	7.091

mmp25

RefSeq ID	GI	PMAP	Name	Score
NP_003729	16932657	hs.12776	dachsous 1 precursor [Homo sapiens]	8.158
NP_002548	32038720	hs.38732	reductase glycoprotein 1 precursor [Homo sapiens]	7.819
NP_114095	13994224	hs.11042	enamelin [Homo sapiens]	7.763
NP_006564	10453892	hs.39575	triadin [Homo sapiens]	7.613
NP_004778	4758146	hs.13958	eph homolog 2 [Homo sapiens]	7.594
NP_078866	83307077	hs.28566	mucin 16 [Homo sapiens]	7.462
NP_742054	26051273	hs.09702	voltage-gated potassium channel, subfamily H, member 2 isoform c	7.5
NP_112200	13959602	hs.10532	organic anion transporter polypeptide-related protein 4 [Homo sapie	7.5
NP_004172	5453810	hs.16072	nestin 2-like [Homo sapiens]	7.462
NP_598851	53832009	hs.20132	calcium channel, voltage-dependent, alpha 1H subunit isoform a [H	7.462
NP_001005407	53832011	hs.21807	calcium channel, voltage-dependent, alpha 1H subunit isoform b [H	7.462
NP_001998	19705556	hs.32358	cadherin EGF LAG seven-pass G-type receptor 3 [Homo sapiens]	7.406
NP_057438	30777594	hs.03134	solute carrier organic anion transporter family member 4A1 [Homo s	7.406
NP_001901	22538668	hs.09209	Smith-Magenis syndrome chromosome region, candidate 7 [Homo	7.406
NP_683684	22538660	hs.09631	Smith-Magenis syndrome chromosome region, candidate 7 [Homo	7.406
NP_375896	113412370	hs.20333	PREDICTED: similar to DENN/MADD domain containing 4B isoform	7.368
NP_001001479	47937222	hs.29004	solute carrier family 35, member E4 [Homo sapiens]	7.368
NP_997262	46459834	hs.00058	hypothetical protein LOC400960 [Homo sapiens]	7.312
NP_004864	10863961	hs.01878	potassium channel, subfamily K, member 10 isoform 1 [Homo sapi	7.312
NP_076564	68798858	hs.21394	thrombospondin repeat containing 1 isoform 2 [Homo sapiens]	7.312
NP_002231	4504843	hs.31646	potassium inwardly-rectifying channel, J4 [Homo sapiens]	7.312
NP_001003682	51262137	hs.32951	hypothetical protein LOC389474 [Homo sapiens]	7.274

Substrates overlap	Common	MMP2_9	MMP14_15_16_24	MMP25
Common		880	798	825
MMP2_and_9	880		1198	1310
MMP14_15_16_24	798	1198		1377
MMP25	825	1310	1377	

Percent overlap	Common	MMP2_9	MMP14_15_16_24	MMP25
Common	100.00	91.86	83.30	86.12
MMP2_and_9	39.04	100.00	53.15	58.12
MMP14_15_16_24	56.60	84.96	100.00	97.66
MMP25	52.92	84.03	88.33	100.00

Pink color – known substrates; substrates for mmp15, 16, 24 and 25 are less studied than substrates for mmp2, 9 and 14

Substrates association with disease is validated in Epilepsy

Similarity and differences in disease association found

MMP25

MMP25 - Reatoad Diseases
Schizophrenia
Schizophrenia and Disorders with Psychotic
Depression
Behavioral Symptoms
Behavior
Vascular Diseases
Thrombosis
Embolism and Thrombosis
Cardiovascular Diseases
Pituitary ACTH Hypersecretion
Psychophysiology
Behavioral Disciplines and Activities
Behavioral Sciences
Psychological Phenomena and Processes
Epilepsy
Musculoskeletal Diseases
Pathologic Processes
Wounds and Injuries
Bone Diseases
Hyperpituitarism

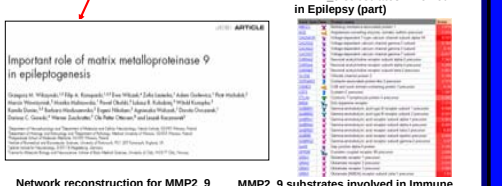
MMP2 and 9

MMP2 and 9 - Related Diseases
Epilepsy
Neoplastic Diseases
Pathological Conditions, Signs and Symptoms
Cardiovascular Diseases
Eclampsia
Pathologic Processes
Schizophrenia
Schizophrenia and Disorders with Psychotic Features
Pre-Eclampsia
Wounds and Injuries
Edema
Skin Diseases, Vascular
Myocardial Infarction
Hypertension, Pregnancy-Induced
Inflammation
Abortion, Spontaneous
Hypertension
Thrombosis
Bone Diseases
Brain Diseases

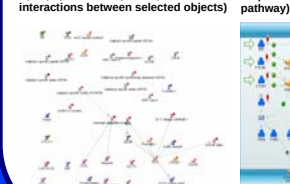
MMP2_9 substrates (part) involved in *Wounds and Injuries* (known relation)

Gene Symbol	Protein name	Score
ABCC1	Canalicular multispecific organic anion transporter 2	0.793
ABCC2	Canalicular multispecific organic anion transporter 2	0.793
ADAM12	ADAM12 precursor	0.525
ADAMTS13	ADAMTS-13 precursor	0.525
ADRA2A	Alpha-2A adrenergic receptor	0.525
ADRB2	Beta-2 adrenergic receptor	0.525
AGER	Advanced glycosylation end product-specific receptor precursor	0.525
AGT	Angiotensinogen precursor	0.525
AGTR1	Type-1 angiotensin II receptor	0.525
AHNAK	Neuroblast differentiation-associated protein AHNAK	0.525
AIF1	Allograft inflammatory factor 1 precursor	0.525
AKT1	Beta-mercaptopyruvate sulfurtransferase	0.525
ANG	Angiogenin precursor	0.525
ANXA5	Annexin A5 (Annexin V) (Endonexin II)	0.525
APOA1	Apolipoprotein A-I precursor [Contains: Complement C3 beta-chain	0.525
APOE	Apolipoprotein E precursor (Contains: Glycoprotein	0.525
CAMP	Antimicrobial cationic peptide precursor	0.525
CASP3	Caspase-3 precursor (Apopain) (CPP32)	0.525
CCL2	C-C chemokine receptor type 2	0.525
CCR2	C-C chemokine receptor type 2	0.525
CD14	Monocyte differentiation antigen CD14 precursor	0.525
CD36	Platelet glycoprotein 4 precursor (CD36 antigen)	0.525
CD4	T-cell surface glycoprotein CD4 precursor	0.525
CD40	Tumor necrosis factor receptor superfamily member	0.525
CD44	CD44 antigen precursor	0.525
CD68	Macrosialin (CD68 antigen)	0.525
CD81	Tetraspanin (Cell surface glycoprotein	0.525
CD9	Complement component 1	0.525
CTGF	Connective tissue growth factor precursor	0.525
CTNNA1	Catenin alpha-1 (Cadherin-associated protein	0.525
CXCL12	Stromal cell-derived factor 1 precursor	0.525
CXCR4	Chemokine receptor 4	0.525
DCN	Decorin precursor	0.525
EDN1	Endothelin-1 precursor	0.525
EGF	EGF precursor	0.525
EGFR	Epidermal growth factor receptor	0.525
ELN	Elastin precursor (Tropoelastin)	0.525
ENG	Endoglin precursor	0.525

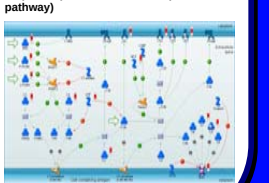
MMP2_9 substrates involved in Epilepsy (part)



Network reconstruction for MMP2_9 Epilepsy – related genes (direct interactions between selected objects)



MMP2_9 substrates involved in Immune response (Lectin Induced complement pathway)



MMP2_9 substrates marked by red dot

MMP2_9 substrates marked by red bar