

# Predicting Substrates for Matrix Metalloproteinases

## Abstract

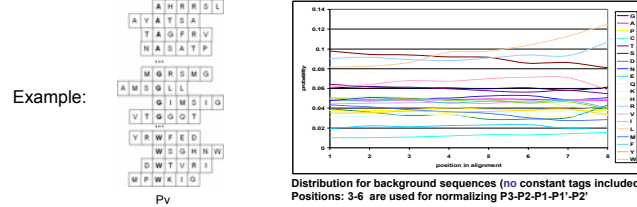
We describe theoretical approach for predicting substrates for matrix metalloproteinases (MMP). We studied seven matrix metalloproteinases representing the three distinct groups in the MMP family: gelatinases (MMP-2,9), transmembrane (MMP-14,15,16,24) and membrane-GPI-linked (MMP-25). The method employs positional weight matrices (PWM) specific for each studied enzyme. The PWMs were derived from statistical analysis of known substrates determined by high throughput phage display experiment and mass spectroscopy analysis. The PWMs were corrected for distribution of amino acids in non-substrate sequences (e.g. background sequences) and used as scoring function in substrate prediction. The prediction power of the PWMs in recognizing position of scissile bonds was augmented by additional filters that use information obtained from analysis of the three-dimensional structures of potential substrates, if it is available. The individual PWMs can be applied for the prediction of differential physiologic and pathologic substrates linking individual family members to particular aspects of biology and pathology, and for the design of selective inhibitors.

## Method – derivation of PWM

**Step I** – derive position probability matrices for set of 230-290 of substrates specific for each MMP – by aligning substrates along their cut sites (P5-P3' positions)

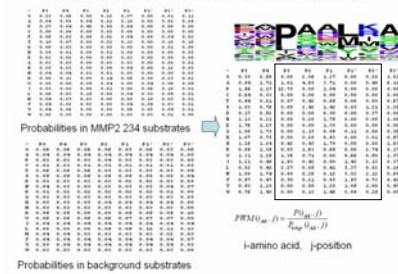
...ADVGGTDYKDDDDKPGGRPTWPSSGGSGETLA-LITASGAET  
...ADVGGTDYKDDDDKPGGRPTWPSSGGSGWIAT-LRTASGAET  
constant tag variable constant tag

**Step II** – normalize each position for each amino acid for background sequence distributions. Such distribution is obtained by aligning 6AA long background sequences using pivot position (Pv).



## Method cont.

Example of PWM for MMP2 (derived without constant tags)



Parameters determined in 10-fold cross validation test:

Offset (treatment of log0 values),

cutoff (score value - below which the sequence is not considered to be a substrate)

position of the "P3" (equivalent to Pv) in the background sequences

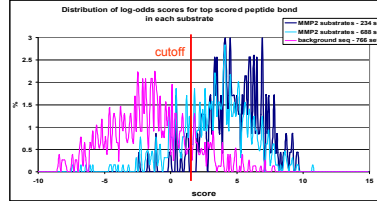
## 10-fold cross validation

	sensitivity	specificity	accuracy	Matthew's corr.coef.	fp_rate	tp_rate	cutoff/offset
MMP2	0.80+-0.08	0.99+-0.01	94.8+-1.4	0.83+-0.04	0.02+-0.01	0.80+-0.08	0.5;-5
MMP9	0.88+-0.06	0.99+-0.01	97.2+-1.6	0.91+-0.05	0.01+-0.01	0.88+-0.06	0.5;-5
MMP14	0.84+-0.03	0.97+-0.01	93.8+-1.4	0.82+-0.03	0.03+-0.01	0.84+-0.03	0.2;-3
MMP15	0.76+-0.07	0.95+-0.03	90.2+-2.2	0.73+-0.06	0.05+-0.03	0.76+-0.07	0.7;-3
MMP16	0.89+-0.04	0.98+-0.02	96.4+-1.8	0.89+-0.06	0.02+-0.02	0.89+-0.04	0.8;-5
MMP24	0.79+-0.12	0.98+-0.01	94.3+-2.8	0.81+-0.10	0.02+-0.01	0.79+-0.12	0.0;-6
MMP25	0.84+-0.08	0.96+-0.02	93.0+-2.9	0.81+-0.08	0.04+-0.02	0.84+-0.07	-0.5;-7

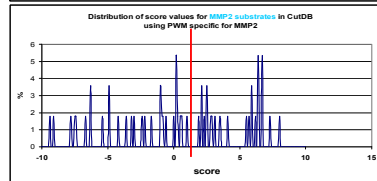
Sensitivity =  $\frac{tp}{(tp+fn)}$ ; fp\_rate =  $\frac{fp}{(fp+tp)}$ ; tp\_rate =  $\frac{tp}{(tp+fn)}$   
Specificity =  $\frac{tn}{(tn+fp)}$ ; Accuracy =  $\frac{(tp+tn)}{(tp+fn+fp+tn)}$   
Mcc =  $\frac{(tp \times tn - fp \times fn)}{\sqrt{((fp+tn)(tp+fn)(fp+tp)(tn+fn))}}$

10-fold cross-validation – performed on set of specific for each enzyme substrates used in phage display experiment.

## Analysis of background sequences, MMP2 substrates from phage display experiment and known substrates in CutDB

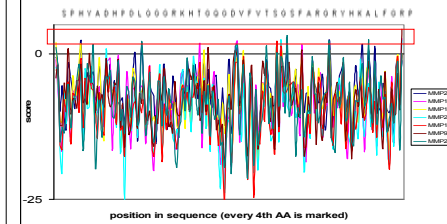


Discrimination between substrates and non-substrates peptide sequences.



Analysis of known substrates collected in CutDB database: <http://cutdb.burnham.org> finds number of sequences that score below cutoff.

## Analysis of MBP - Myelin Basic Protein. No x-ray str. available



Every peptide bond is scored using 5-AA window denoting: P3-P2'

MASQKRPQGRHSGSKYLLATASDTHDARRGFLPRHRDTCLEDSRPFQDGRGAKPRGSGKDSDHIFARTAHYGSLLPKSGHRTQDENVYVHFQVUUNFNIVPRTPRPPSGKRGUULSLERFSWGAEGQRPGFGYGRASDLYKSRKNGFGVDAQTGLSLKFLGHDSDRSGSPHARR

MMP-2 9 14 25 Experimental:  
↓, ↓, ↓ primary hydrolysis site  
↓, ↓, ↓ secondary site

	MMP2	MMP9	MMP14	MMP15	MMP16	MMP24	MMP25	MMP2	MMP9	MMP14	MMP15	MMP16	MMP24	MMP25
SKY-LA	-	-	-	(?)	(?)	(?)	(?)	-5.58	-1.88	-1.32	1.31	-0.76	-5.27	2.31
YGS-LP	-	-	-	(?)	(?)	(?)	+(Pat P2?)	-6.02	-11.33	1.49	-2.35	-1.55	-2.98	-4.16
VHF-FK	+	+	+	(?)	(?)	(?)	+	1.57	-0.71	2.11	-0.77	-1.52	1.57	1.88
GRGLS	+	+	+	(?)	(?)	(?)	+	0.15	0.14	0.85	2.02	2.18	1.81	-0.78
ASD-YK	+	+	+	(?)	(?)	(?)	+	2.12	0.78	-2.42	1.23	-2.19	-1.83	2.35

Experimental sites Predicted sites – in red

## Hydrolysis sites in α-1-antitrypsin – X-ray structure available

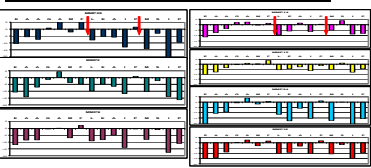


166 171 346 351 356 361  
DQKIVDLVKV RHRUMFLRPLMSIIPP

Differential hydrolysis of α-1-antitrypsin  
Experimental sites: AMF-LE, AIP-MS for MMP25.

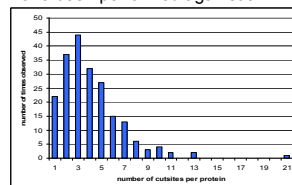
- Prediction: PWM method augmented by additional structural criteria:
- cut site should not be in helical or extended secondary structures
  - Sequence of 5AA should be located on surface
  - AA side chains should be solvent exposed (at least 36%).

Results: all exp. sites have been recognized. Additional sites have been found next to experimentally determined sites and in other loop.



## Potential MMPs substrates in *Thermotoga maritima* proteins

*T.m.* is a model organism for which whole genome has been determined and around 208 of three dimensional structure of proteins has been solved and deposited into PDB by the effort of JCSG center at UCSD. At BIMR there is ongoing project to experimentally and virtually screen *T.m.* proteins for the presence of MMP substrates and determining pathways. Experimentally: 52 proteins is being screened. Computationally: virtual screening of proteins have been performed against all MMPs.



Virtual screening of *T.m.* proteins against MMP14.

	Total no. of detected cleavage sites in 208 proteins using PWMs	No. of cleavage sites after additional structural filtering
MMP2	6161	17 in 13 proteins
MMP9	5350	19 in 16 proteins
MMP14	6414	28 in 20 proteins
MMP15	6790	27 in 17 proteins
MMP16	3328	14 in 12 proteins
MMP24	3638	13 in 12 proteins
MMP25	12067	44 in 32 proteins

Virtual screening of *T.m.* proteins against all MMPs, before and after structural filtering.

## Conclusions:

- The PWM based prediction method is general, applicable to any enzyme.
- The method is validated in 10-fold cross-validation test.
- The method is highly specific
- Additional structural criteria substantially improve cut-sites prediction.

## Future improvement and extensions of the algorithm include:

- predicting secondary structures,
- predicting disordered region,
- creating three-dimensional protein models,
- incorporating hydrolysis scores as additional element of scoring,
- incorporating "virtual gel", to easier correlate predictions with experiment.
- web-based tool for predicting scissile bonds

## Ongoing applications:

Prediction of all potential substrates of MMPs in the whole human genome for pathways annotation.

Acknowledgment: this project is supported by NIH Roadmap – Technology Centers for Networks and Pathways, awarded to BIMR (J.W.Smith) for Center for Proteolytic Pathways. (<http://protease.burnham.org>)