15th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conference on Computational Biology (ECCB)

CALL FOR ABSTRACTS - 3rd ISCB Student Council Symposium 2007

Modeling substrate specificity of human dipeptidyl-peptidase III using Random Forests

Fran Supek[a], Marija Abramić[b], Tomislav Šmuc[a]
[a] Department of Electronics
[b] Department of Organic Chemistry and Biochemistry
Rudjer Boskovic Institute, Zagreb, Croatia

email for correspondence: fran.supek@irb.hr

Summary: Substrate specificity of the human dipeptidyl-peptidase III has been modeled using Random Forests on representations of amino acids by three general physicochemical properties. Site P1 has a strong impact on peptide binding affinity, but does not affect cleavage. A hydrophobic amino acid at site P1' is favorable for both processes.

Human dipeptidyl-peptidase III (DPP III) belongs to the metallopeptidase M49 family and cleaves N-terminal dipeptides with broad specificity. This cytosolic enzyme of eukaryotes[1] may be medically significant as DPP III activity correlates with aggressiveness of ovary carcinomas[2], has a role in pain modulation[3] and in the endogenous defense against oxidative stress[4]. Properties of human DPP III, such as substrate affinity[5], or reactivity of cysteines important for activity[6], differ considerably from rat DPP III, despite very high sequence similarity (~93% identity).

We have attempted to computationally model the properties of peptides that make them good substrates for (A) cleavage or (B) binding by human DPP III. The 'cleavage' dataset consisted of 53 peptides of length 3-7 with cleavage roughly quantified by spot detection after thin-layer chromatography[7] and encoded as 'positive' and 'negative' classes of approximately equal sizes.

The 'binding' dataset consisted of 39 peptides of length 3-7, whose binding affinity, expressed by inhibitory constant (K_i), was determined by treating peptides as alternate substrate inhibitors[8]. Only peptides with very high binding affinities may be relevant as substrates or inhibitors *in vivo*; therefore, we have separated the top 30% (n = 12) binders ($K_i < 3.5$ μM) into a 'positive class' and the others into a 'negative' class.

Each peptide was described by three different amino acid physicochemical properties for the each of the peptide's first three amino acids (P2, P1 and P1'), totaling 9 attributes per peptide. The physicochemical properties were derived by reducing the dimensionality of the Amino Acid Index database [9] using principal components analysis (PCA). The first three principal components (PCs) retained 33%, 15% and 12% of information in the database, and were correlated to hydrophobicity (PC-1), and anticorrelated to: α-helix propensity (PC-2) and abundance in mesophile proteins (PC-3).

A representation of amino acids using only their very general properties, in contrast to noting presence of an exact amino acid at an exact position, may be advantageous with enzymes of broad specificity, and when only a dataset of very limited size is available (danger of overfitting).

The Random Forest (RF) classifier[10] is essentially an ensemble of decision trees built on subsets of the data. RF offers predictive performance in line with the Support Vector Machines[11] while being less sensitive to choice of training parameters, and easily parallelized for speed[12]. Some implementations allow insight into functioning of models via evaluation of attribute importance and computation of 'class prototypes', where the algorithm separates classes into subgroups by how frequently instances share branches in the individual decision trees.

The predictive ability of models for cleavage by, and binding affinity to, DPP III of peptides was moderate – crossvalidation accuracy was 71.7%, and 82.1%, respectively. Amino acid 2 (P1) was not important for determining whether a peptide will be cleaved, however it is highly relevant for determining binding affinity to DPP-III. Amino acid 3 (P1'), especially its hydrophobicity, is important in both cases: highly hydrophobic amino acids (e.g. isoleucine) are likely to improve both cleavage and binding. In modeling binding affinity the fourth amino acid may also be relevant (data not shown).

Table 1. The relative importance ("import") of nine attributes – three physicochemical properties per amino acid in the peptide – for (a) cleavage or (b) binding affinity to DPP III is denoted by diamonds (◆). Each diamond is worth ten in Z-score of attribute importance, as reported by RF. Prototypes in each classification task are described by: class ("pos" or "neg"), number of instances "*n*" in the prototype, and median values of each attribute within the prototype. Representative peptides are approximations derived from attribute medians in prototypes.

		cleavage dataset (*n*=53)					affinity dataset (*n*=39)			
amino acid	PC	import	prototypes				import	prototypes		
			neg *n*=20	**neg *n*=5**	**pos *n*=20**	**pos *n*=6**		**pos *n*=10**	**neg *n*=18**	**neg *n*=6**
1 (P2)	hydro	◆	-0.55	0.49	0.49	0.49	◆◆◆	1.31	0.49	0.49
	helix	n.s.					◆	-0.02	-0.23	0.37
	rarity	◆◆◆	0.01	-0.19	0.01	1.53	◆◆◆	-0.98	0.01	1.53
2 (P1)	hydro	n.s.					◆◆◆◆◆	1.21	-0.55	-1.26
	helix	n.s.					◆◆◆◆	0.37	0.67	2.07
	rarity	n.s.					◆◆◆◆◆	1.53	-0.19	0.43
3 (P1')	hydro	◆◆◆◆◆	0.11	-1.26	1.23	-1.26	◆◆◆◆	1.21	1.21	-0.55
	helix	◆◆◆	0.23	2.07	0.23	2.07	◆◆	0.37	0.23	2.07
	rarity	◆◆◆	-0.19	-0.95	0.43	-0.95	◆◆	0.01	0.43	-0.95
representative peptide for the prototype			TSKQ–?–T	M–?–G	QM–?–IM	Y–?–G		IV–WC–FI	MF–Y–IM	Y–P–P

[1] Abramić M and Vitale Lj., Aminopeptidases in the cytosol of mammalian cells. *Acta Pharmaceutica*. 1994; 44: 71–85.
[2] Šimaga Š *et al.*, Tumor cytosol dipeptidyl peptidase III activity is increased with histological aggressiveness of ovarian primary carcinomas. *Gynecol Oncol*. 2003; 91: 194-200.
[3] Baršun M. *et al*., Human dipeptidyl peptidase III acts as a post-proline-cleaving enzyme on endomorphins. *Biol Chem*. 2007; 388: 343-348
[4] Liu Y. *et al*., A genomic screen for activators of the antioxidant response element. *Proc Natl Acad SCI USA*. 2007; 104: 5205-5210.
[5] Abramić M *et al*., Human and rat dipeptidyl peptidase III: biochemical and mass spectrometric arguments for similarities and differences. *Biol Chem*. 2000; 381: 1233-43.
[6] Abramić M *et al*., Highly reactive cysteine residues are part of the substrate binding site of mammalian dipeptidyl peptidases III. *Int J Biochem Cell Biol*. 2004; 36: 434-46.
[7] Abramić M. *et al*., Dipeptidyl peptidase III from human erythrocytes. *Biol. Chem. Hoppe-Seyler* 1988; 369: 29-38
[8] Chu TG and Orlowski M, Soluble metalloendopeptidase from rat brain: Action on enkephalin-containing peptides and other bioactive peptides. *Endocrinology.* 1985; 116: 1418-1425
[9] Kawashima S *et al*., AAindex: amino acid index database. *Nucleic Acids Res*. 1999; 27: 368-369.
[10] Breiman L, Random Forests. *Machine Learning*. 2001; 45: 5-32
[11] Qi Y *et al*., Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins,* 2006; 63: 490-500.
[12] Šmuc T and Topić G, PARF: Parallel Random Forests. http://www.parf.irb.hr/