Modification Specific Modeling in PeptideProphet Improves Validation of Rare PTM Containing Peptides

¹Institute for Systems Biology, Seattle WA ²University of Washington, Seattle WA

Overview

The Trans-Proteomic Pipeline (TPP) has been a gold-standard, open source analysis tool for proteomics data for well over a decade, and its major component PeptideProphet was indeed first published 20 years ago! In this abstract we describe a new model in PeptideProphet for validating PSM data searched with many variable modifications, some of them rare and others less rare, in a single search.



Comet searches were performed using variable modifications shown in the To test the feasibility of detecting rare PTMs we performed native bioactivation of raloxifene in insect cell microsomes, generating parameters below. The database used for searching was composed of a highly complex sample made by co-expressing native human CPY3A4, cytochrome P450 reductase and cytochrome b5 in UniProt protein sequences of the organism Spodoptera frugiperda, plus insect cells. Raloxifene metabolism produces several electrophilic species one of which forms protein adducts of mass 471.1504 human P450 enzymes (e.g. CYP3A4), human P450 reductase, cytochrome Da. We incubated raloxifene with the insect cell microsomes resulting in raloxifene metabolism and protein adduct formation. We b5, yeast enclase, and common contaminants. Two sets of independently collected data on unexposed (solvent only), light (d0) raloxifene exposed, heavy (d4) raloxifene exposed and a mixture of d0/d4 randomized decoy sequences were appended to the target database. The raloxifene exposed samples. Comet searches were performed allowing for variable modifications of 471.1504 (d0 raloxifene decoy sequences were generated using DeBruijn repeat-preserving diquinone methide metabolite) and 475.1755 (d4 raloxifene diquinone methide metabolite) on cysteine, tryptophan and tyrosine, randomization, provided by software within the TPP. The decoy sequences 15.9949 on methionine (oxidation) and 57.021464 on cysteines (carbamidomethyl). were randomly interleaved in the fasta database used for the Comet search.

Dataset
AZ944
AZ945
AZ946
AZ945-6-mix

Variable Modification Count (VMC)

The new VMC model assists PeptideProphet to better classify PSMs containing variable modified PSMs are more likely to occur among random results than among correct results.

- xinteract option -OV
- PeptideProphetParser option VMC
- Computes a different VMC count and model separately

Using VMC does not negatively imp PeptideProphet's FDR estimate a all PSMs in the dataset, as demonstra using independent entrapment decova

By charge state (same as original) variable PTM type (New!)

As a result, PeptideProphet is better able to control FDRs and error rates on rare-PTM-containing p entrapment decoys employed during the search, while preserving rare PTM containing PSMs with evidence.



David D. Shteynberg¹, Michael R. Hoopmann¹, Eric W. Deutsch¹, Luis Mendoza¹, Jimmy Eng², Alex Zelter², Nina Isoherranen², and Robert L. Moritz¹

Raloxifene Adducts d0/d4

Label					
No label		Dataset	PSMs (Peptides)	ENTRAP PSMs (Peptides)	PSM Error Rate
d0, light label			with Adduct	with Adduct	(Peptide Error Rate)
d4, heavy label	Percolator v.3.06.0 q ≤ 0.01	AZ944	d0: 65 (49) d4: 94 (61) 32 (24)	67% (73%)
d0/d4, light/heavy mix		AZ945	d0: 89 (43) d4: 48 (26)	32 (14)	72% (65%)
		AZ946	d0: 45 (36) d4: 95 (52)	25 (15)	55% (53%)
Mode	noVMC iProphet 1% FDR	Dataset	PSMs (Peptides) with Adduct	ENTRAP PSMs (Peptides) with Adduct	PSM Error Rate (Peptide Error Rate)
		AZ944	d0: 11 (11) d4: 15 (11)	1	12% (30%)
		AZ945	d0: 64 (30) d4: 19 (15)	1	3.6% (6.7%)
cations. Intuitively, such		AZ946	d0: 4 (3) d4: 67 (36)	2 (1)	8.5% (7.7%)
0.012	-	AZ945-6	d0: 27 (19) d4: 48 (30)	1 (1)	4.0% (6.1%)
0.01					
		Dataset	PSMs (Peptides) with Adduct	ENTRAP PSMs (Peptides) with Adduct	PSM Error Rate "ε" Range [§] (Peptide Error Rate Range)
ted y 0.004	VMC iProphet 1% FDR	AZ944	d0: 0 (0) d4: 1 (1)	0	0%≤ε<75% (0%≤ε<75%)
0.002		AZ945	d0: 55 (23) d4: 0	0	0%≤ε<5.4% (0%≤ε<12.5%)
PeptideProphet (VMC) PeptideProphet (no VMC)	,	AZ946	d0: 0 d4: 48 (21	.) 0	0%<ε<6.1% (0%<ε<13.6%)
0 0.002 0.004 0.006 0.008 0.0 Decoy-estimated Error Rate	01	AZ945-6	d0: 15 (8) d4: 28 (14)) 0	0%≤ε<6.8% (0%≤ε<13%)
peptides as indicated by	15000			§ Assumin	g one imaginary entrapment PSM
strong complementary	45000]			
r strong complementary	40000				
	40000				
+3 Models	35000			and the second se	
var mod count type [vmcC57pt021] /mc=0 0.825, vmc=1 0.156, vmc=2 0.016, vmc=3 0.003)	(0				
vmc=0 0.858, vmc=1 0.123, vmc=2 0.017, vmc=3 0.003) var mod count type [vmcC471pt150]	분 30000	PSM Validation AZ944:			
/mc=0 0.999, vmc=1 0.001, vmc=2 0.000, vmc=3 0.000)	ect arooo	ROC* comparing performance of:			
var mod count type [vmcC475pt175]	5 25000 ·			 IProphet (with and PoptideProphet (with and 	without VMC),
/mc=0 0.998, vmc=1 0.002, vmc=2 0.000, vmc=3 0.000) /mc=0 0.963, vmc=1 0.035, vmc=2 0.002, vmc=3 0.000)	- 5 20000			 PeptidePropriet (w Percolator (v3.06.0 	a < 0.01
var mod count type [vmcM15pt995]	per			(, - <u> </u> ,
/mc=0 0.752, vmc=1 0.208, vmc=2 0.036, vmc=3 0.004)	<u>5</u> 15000		* FL	DR is computed using entrapment dec	oys, comprising 1/3 of all "targets"
var mod count type [vmcW471pt150] /mc=0 0.999, vmc=1 0.001, vmc=2 0.000, vmc=3 0.000)	2				
vmc=0 0.974, vmc=1 0.026, vmc=2 0.001, vmc=3 0.000) var mod count type [vmcW475pt175]	10000		VMC (iPro	DECOY)	
/mc=0 0.999, vmc=1 0.001, vmc=2 0.000)	E000		VMC (PepPro	DECOY)	
var mod count type [vmcY471pt150]	5000		no VMC (PepPro	DECOY)	
/mc=0 1.000, vmc=1 0.000, vmc=2 0.000, vmc=3 0.000) /mc=0 0.923, vmc=1 0.072, vmc=2 0.004, vmc=3 0.000)	0		Percolator 3.06 (q<0.0)	1 DECOY)	
var mod count type [vmcY475pt175]		0.002 0.0	04 0.006 0.008 0.01	0.012 0.014	
/mc=0 0.916, vmc=1 0.077, vmc=2 0.006, vmc=3 0.000)	-		FDR		



Comet + TPP

# comet version 2021.02 rev. 0					
# Generated via Petunia/TPP by dsh	stevnh on Mon May 22 13:31:24 2023				
# Comet MS/MS search engine parame	eters file.				
# Everything following the '#' sym	abol is treated as a comment.				
database_name = /proteomics/dshtey	/nb/data/AlexZelter/Jan2023/AZ944_3pepCnt_fixed_DECOY.fasta				
decoy_search = 0	# 0=no (default), 1=concatenated search, 2=separate search				
peff_format = 0	# 0=no (normal fasta, default), 1=PEFF PSI-MOD, 2=PEFF Unimod				
peff_obo =	# path to PSI Mod or Unimod OBO file				
num_threads = 72	# 0≈poll CPU to set num threads; else specify num threads directly (max 128)				
# masses					
peptide_mass_tolerance = 16					
peptide_mass_units = 2	# 0=anu, 1=nnu, 2=ppm				
mass_type_parent = 1	# 0=average masses, 1=monoisotopic masses				
mass_type_fragment = 1	# 0=average masses, 1=monoisotopic masses				
precursor_tolerance_type = 0	# 0=MH+ (default), 1=precursor m/z; only valid for anu/mmu tolerances				
Taocobe_error = 1	a endri' vediv (eva elinel) endriva nedvivia deladidate (ini selite venevull)				
# search enzyme					
*					
search_enzyme_number = 1	# choose from list at end of this params file				
search_enzyme2_number = 0	# second enzyme; set to 0 if no second enzyme				
num_enzyme_termini = 1	# 1 (semi-digested), 2 (fully digested, default), 8 C-term unspecific , 9 N-term unspecific				
allowed_missed_cleavage = 3	# maximum value is 5; for enzyme search				
a lin an O undable medificanting a					
# Op to 9 variable modifications a	ire supported aniabla/dia hinanyi yaay mode nan nantidai yteem distancai yn/c_termi yraaninadi ynaiteal lossi				
# e.g. 70 066331 STV 0.3 -1.0					
#					
variable_mod01 = 15.9949 M 0 3 -1	(0,0) [vaniable model = 15 0040 M 0 3 -1 0 0 0 0				
variable_mod02 = 57.021464 C 0 3 -	10000 Val 1aute moust - 13,9949 M 8 3 -1 8 8 8.8				
variable_mod03 = 471.1504 CWY 0 3	-1 0 0 0.0				
variable_mod04 = 475.1755 CWY 0 3	10000 [worish] = mod 0.2 = 57 0.21464 C 0.2 1 0 0 0 0				
#variable_mod04 = 0.0 X 0 3 -1 0 0	\therefore Val. Table 0 0 2 2 0 0 0 0 0				
1 March 4 M M M M M M M M M M M M M M M M M					
wariable medec - 0.0 X 0.3 -1.0 0					
variable_mod06 = 0.0 X 0 3 -1 0 0 variable_mod07 = 0.0 X 0 3 -1 0 0					
variable_mod06 = 0.0 X 0 3 -1 0 0 variable_mod07 = 0.0 X 0 3 -1 0 0 variable_mod07 = 0.0 X 0 3 -1 0 0 variable_mod08 = 0.0 X 0 3 -1 0 0	variable mod03 = 471.1504 CWY 0 3 -1 0 0 0.0				
variable_mod06 = 0.0 X 0 3 -1 0 0 variable_mod07 = 0.0 X 0 3 -1 0 0 variable_mod07 = 0.0 X 0 3 -1 0 0 variable_mod08 = 0.0 X 0 3 -1 0 0 variable_mod09 = 0.0 X 0 3 -1 0 0	variable_mod03 = 471.1504 CWY 0 3 -1 0 0 0.0				
Variable_mod06 = 0.0 X 0 3 -1 0 0 variable_mod07 = 0.0 X 0 3 -1 0 0 variable_mod08 = 0.0 X 0 3 -1 0 0 variable_mod09 = 0.0 X 0 3 -1 0 0 variable_mod09 = 0.0 X 0 3 -1 0 0 max variable_mod09 = 0.0 X 0 3 -1 0 0	variable_mod03 = 471.1504 CWY 0 3 -1 0 0 0.0				
Variable_mod06 = 0.0 X 0 3 -1 0 0 Variable_mod07 = 0.0 X 0 3 -1 0 0 Variable_mod07 = 0.0 X 0 3 -1 0 0 Variable_mod09 = 0.0 X 0 3 -1 0 0 max_variable_mod5_in_peptide = 5 require_variable_mod_= 0	variable_mod03 = 471.1504 CWY 0 3 -1 0 0 0.0				

The Comet search results were validated using the TPP software on the PSM, peptide and modifications levels, both with and without the use of the novel modification-specific Variable Modification Count (VMC) model.

Conclusions

- The new Variable Modification Count (VMC) model improves classification of PSMs without negatively impacting performance of TPP classifiers PeptideProphet and iProphet.
- We demonstrate the application of the VMC model in the identification of protein adducts.
- VMC greatly improves TPP classification of PSMs modified by rare PTMs, as confirmed by both entrapment decoys and by the prior knowledge of d0/d4 sample type.
- Tested using entrapment decoys, PeptideProphet with or without the VMC model outperforms Percolator. Furthermore, iProphet boosts the performance of PeptideProphet with or without VMC.

Support & Information

Support provided by:

- NIH, NIGMS grants: R01GM087221
- NIAID grant: R21AI133335
- NHLBI grant: R01HL133135



TPP Resources:

www.tppms.org