Practical 4’-Phosphopantetheine Active Site Discovery from Proteomic Samples

Jordan L. Meier[1], Anand D. Patel[2]*, Sherry Niessen[5], Michael Meehan , Roland Kersten[1], Jane Y. Yang[1], Michael Rothmann[1], Benjamin F. Cravatt[5], Pieter Dorrestein[1, 3-4], Michael D. Burkart[1], and Vineet Bafna[2]*

[1]Department of Chemistry and Biochemistry, [2]Department of Computer Science and Engineering Bioinformatics Program, [3]Department of Pharmacology, and [4]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla CA 92093. [5]The Skaggs Institute for Chemical Biology and Department of Chemical Physiology, The Center for Physiological Proteomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037.

Email: mburkart@ucsd.edu, pdorrestein@ucsd.edu, vbafna@cs.ucsd.edu

Supporting Information

Contents

Supplementary Figures

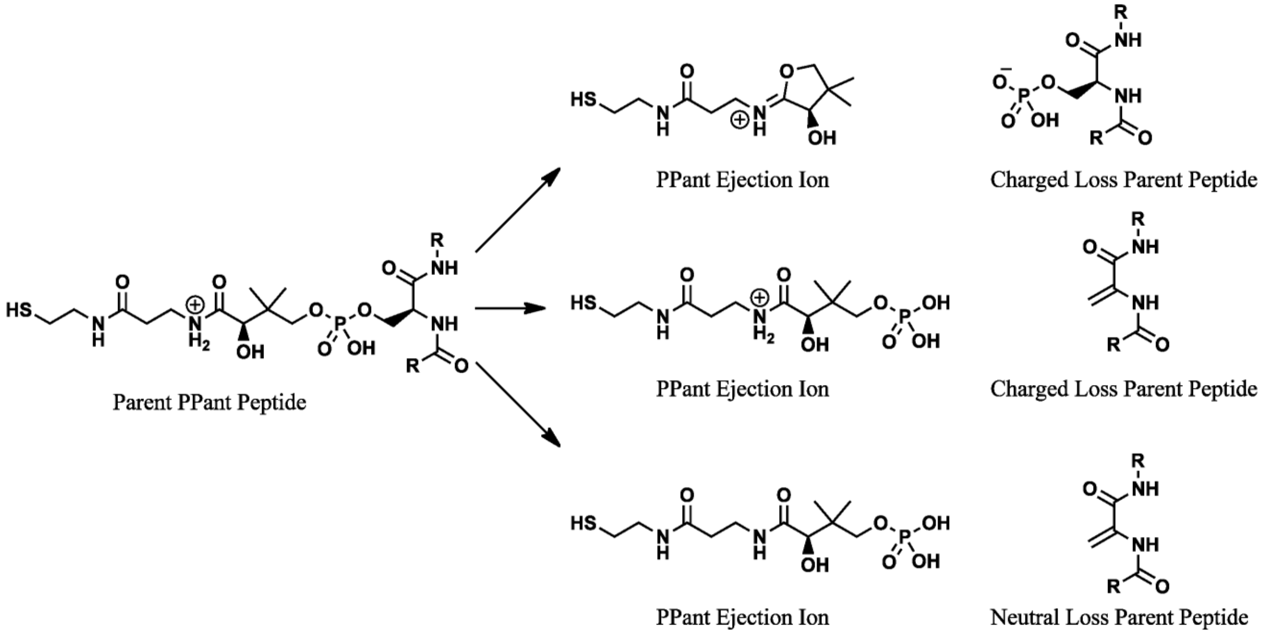
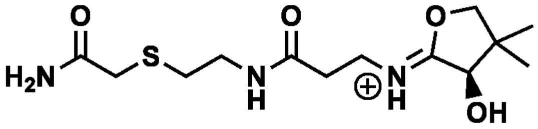
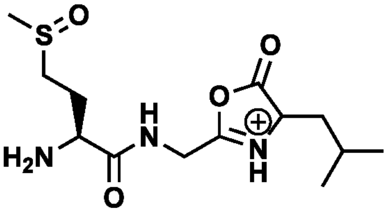

Parent PPant Peptide
PPant Ejection Ion
Charged Loss Parent Peptide
PPant Ejection Ion
Charged Loss Parent Peptide
PPant Ejection Ion
Neutral Loss Parent Peptide

Figure S1: Potential species arising from fragmentation/ejection of PPant peptides.



Pantetheine Ejection Fragment
Chemical Formula: $C_{13}H_{24}N_3O_4S$
Exact Mass: 318.15



Met-(sulfoxide)-Gly-Leu b_3 fragment
Chemical Formula: $C_{13}H_{24}N_3O_4S$
Exact Mass: 318.15

Figure S2: Peptide species having the same molecular formula as alkylated PPant ejection fragment. In addition to the pictured peptide fragment, constitutional isomers and Ile/Leu substituted fragments also have identical molecular formulas.

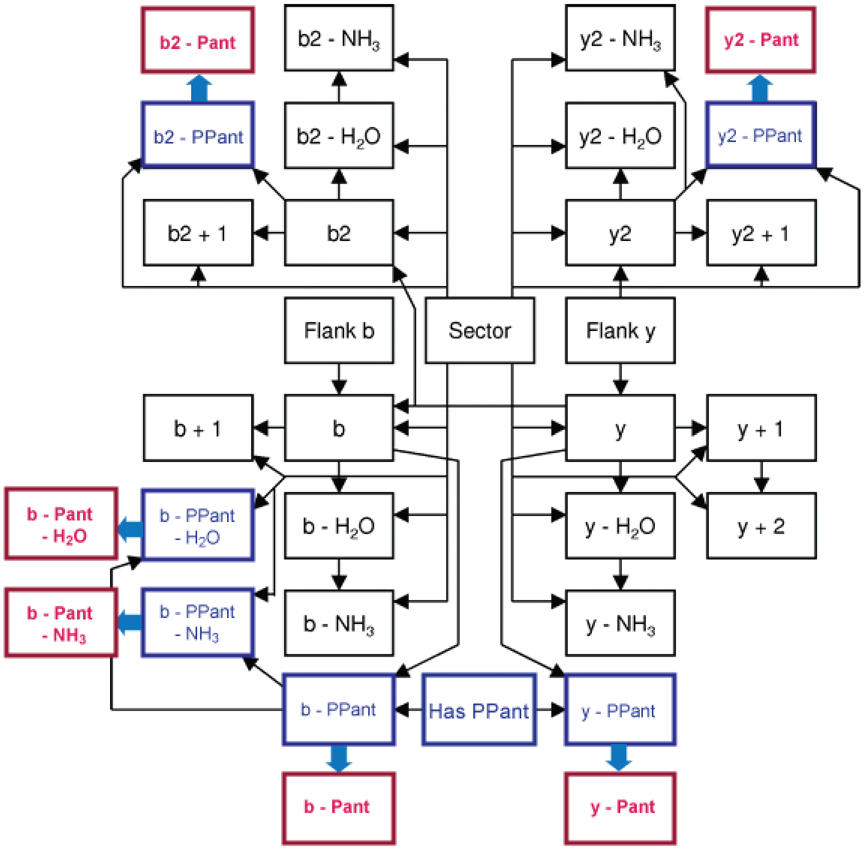


Figure S3: Alterations made to Bayesian network from Payne et al. for InsPecT:PPant. Phosphorylation neutral loss nodes were duplicated as indicated by blue arrows pointing to red nodes. The original phosphorylation nodes (in blue) were modified to a PPant mass offset of -397 and newly created nodes (in red) were altered to a pantetheinyl mass offset of -317. All newly created nodes inherited exactly the same edge relations as the nodes they were derived from. This is a necessary condition as no training is performed on PPant peptide examples.

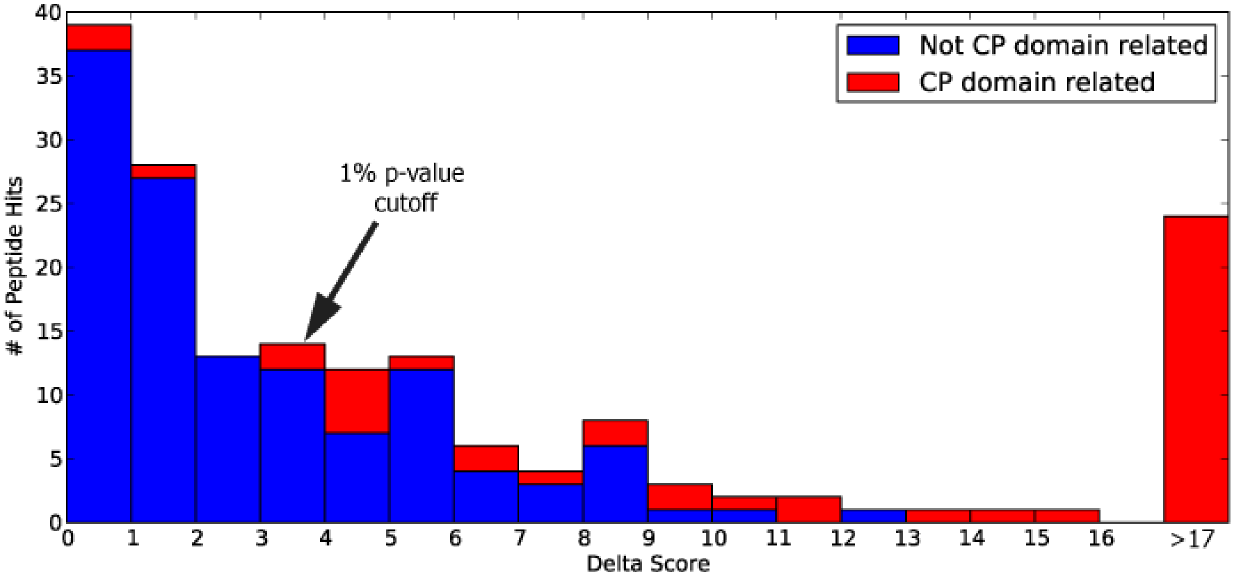


Figure S4: Delta score distributions of CP domain related peptide hits and non-CP domain peptide hits. Each peptide hit was the best match for a MS^3 PPant signature confirmed PPant peptide spectrum.

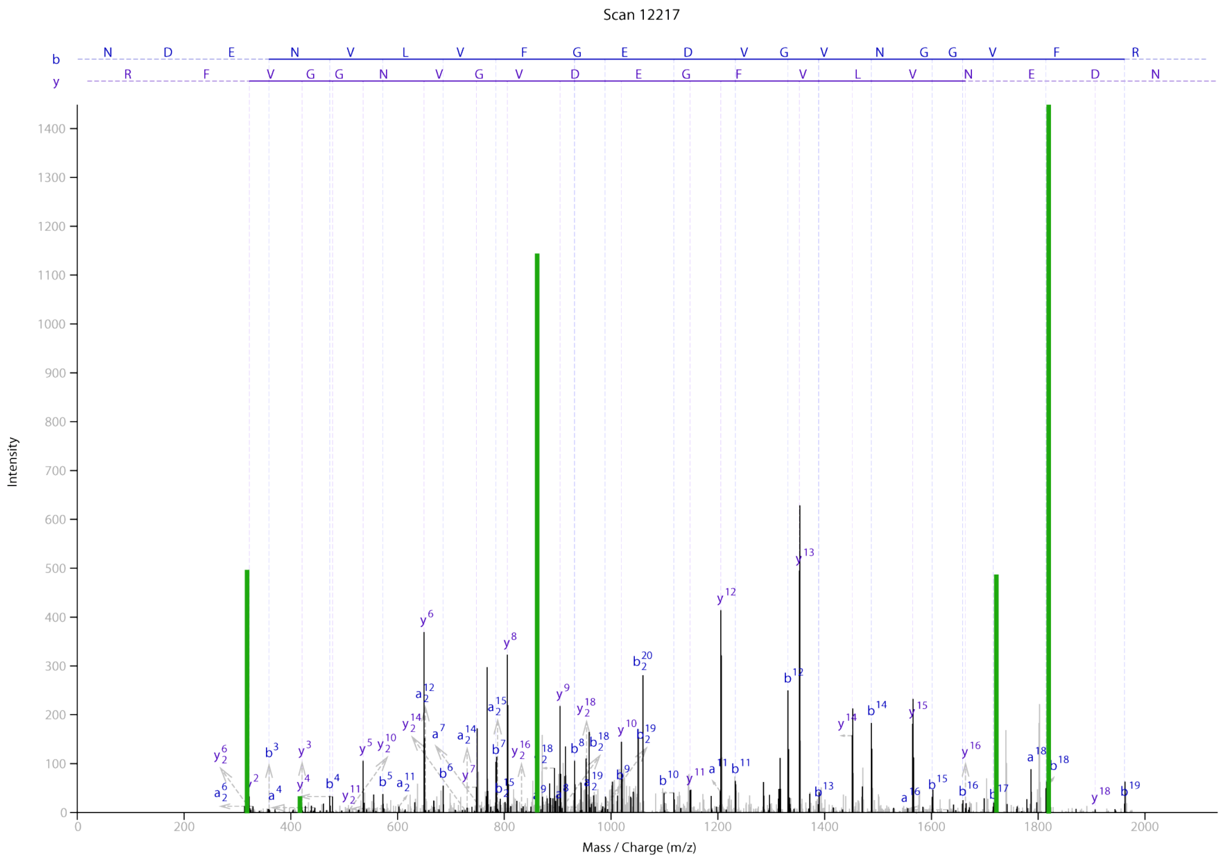


Figure S5: Example of mixture spectra (fp5ms3318-02.mzXML, scan 12217). InsPecT annotated spectrum as 2+ K.NDENVLVFGEDVGVNGGVFR.A from protein PdhB, pyruvate dehydrogenase (E1 beta subunit). Green peaks were unannotated by InsPecT and correspond to expected characteristic PPant ions.

Table S1: CCMS LiveSearch InsPecT search results for FP-biotin enrichment of *B. subtilis* 168 proteome. See tab-delimited text file, *bsubtilis_livesearch_SLR_fdr01.txt*.

Table S1.1: CCMS LiveSearch InsPecT search results for FP-biotin enrichment of *B. subtilis* 168 filtered for MS3 validated hits. See tab-delimited text file, *bsubtilis_livesearch_SLR_fdr01_pp-validated.txt*.

Table S2: Annotated Characteristic PPant Ions. Exact *m/z* values for each expected ion type are calculated using the parent PPant peptide's monoisotope parent mass, $\boldsymbol{M}$, and charge, $\boldsymbol{z}$.

$$\boldsymbol{M} = precursor\ m/z \times \boldsymbol{z} - (neutron\ mass)\boldsymbol{i} - (proton\ mass)\boldsymbol{z}$$

$\boldsymbol{z}$ is the peptide charge, $\boldsymbol{i}$ is the peptide isotope.

Typical proteomic MS runs involve the artificial addition of a substrate to the thiol end of PPant (i.e. iodoacetamide treatment is required to break disulfide bridges with a resulting 57 mass addition to thiol groups). This substrate mass is represented as $\boldsymbol{s}$, which is 0 for PPant with no substrate additions.

Expected and Observed values are based on the +2 peptide "FFDLGGHS*LLAVQLMSR" (scan 13673, run fp3-04, Figure 4), where $\boldsymbol{M}$=2287.784, $\boldsymbol{i}$=0, $\boldsymbol{z}$=2, and $\boldsymbol{s}$=57.

	Ion type	Expected m/z Calculation	Expected m/z	Observed m/z
Ejected Ions	1) Pant	$260.120 + 1.008 + \boldsymbol{s}$	318.127	318.248
	2) PPant	$340.086 + 1.008 + \boldsymbol{s}$	398.093	-
	3) PPant + H_2O	$340.086 + 18.015 + 1.008 + \boldsymbol{s}$	416.109	416.266
Intact Peptide with Neutral Loss Ions	4) M-PPant-H_2O	$\frac{\boldsymbol{M} - [340.086 + 18.015 + \boldsymbol{s}] + \boldsymbol{z} \times 1.008}{\boldsymbol{z}}$	926.342	937.277
	5) M-PPant	$\frac{\boldsymbol{M} - [340.086 + \boldsymbol{s}] + \boldsymbol{z} \times 1.008}{\boldsymbol{z}}$	945.350	946.229
	6) M-Pant-H_2O	$\frac{\boldsymbol{M} - [260.120 + 18.015 + \boldsymbol{s}] + \boldsymbol{z} \times 1.008}{\boldsymbol{z}}$	976.325	977.261
	7) M-Pant	$\frac{\boldsymbol{M} - [260.120 + \boldsymbol{s}] + \boldsymbol{z} \times 1.008}{\boldsymbol{z}}$	985.333	985.965
Intact Peptide with Charged Loss Ions	8) M-PPant-H_2O	$\frac{\boldsymbol{M} - [340.086 + 18.015 + \boldsymbol{s}] + (\boldsymbol{z} - 1) \times 1.008}{(\boldsymbol{z} - 1)}$	1872.684	1872.904
	9) M-PPant	$\frac{\boldsymbol{M} - [340.086 + \boldsymbol{s}] + (\boldsymbol{z} - 1) \times 1.008}{(\boldsymbol{z} - 1)}$	1890.699	1891.053
	10) M-Pant-H_2O	$\frac{\boldsymbol{M} - [260.120 + 18.015 + \boldsymbol{s}] + (\boldsymbol{z} - 1) \times 1.008}{(\boldsymbol{z} - 1)}$	1952.650	1951.898
	11) M-Pant	$\frac{\boldsymbol{M} - [260.120 + \boldsymbol{s}] + (\boldsymbol{z} - 1) \times 1.008}{(\boldsymbol{z} - 1)}$	1970.665	1970.980

Table S3: CP active site peptide identification search results for MS3 validated spectra, filtered by a 5% CP domain protein match p-value cutoff and a delta score cutoff of 3.84 (based on a 1% p-value using the empirical distribution of delta scores in the top 50 peptide interpretations of all spectra). See tab-delimited text file, *out_identifyiso_out_detect_ms3_bsubtilis_d0384cpp005.txt*.

Table S4: List of PPant positive (1 in the train column) and negative (-1 in the train column) spectra used for SVM learning. See tab-delimited text file, *svm_training_data.txt*.

Table S5: ProteomeCommons.org hosts all MS data used in this paper. Below are the Tranche Hashes to download data.

MS Data	Hash
PikAIV	plawkTZIMb6fm6OaaMP3K9SRX7ZazTA54Qoma0VF/3j2mYvQ49O8NFtYtws1LHD1kVDG81DV2ABEkT3SRJASpXXAx+UAAAAAAAABfQ== HHyfiJd9qFdsy1KM9bjjGGOI2ER9Lsvt1Gf+Vd/Qq4dqXA05nph8nNrHwiYaZSea2sckS2KPaiPksCPWaX4v/VZOhf0AAAAAAAABdw== UPcoGpM1zqTmNqUYe9XzjJcpdMQCb12OU+JaPO7yu/mHEaEzbJBLb2DvJOBDpxxN1Q5k7GQ86bl5HvprSjWcMOfgzZ8AAAAAAAABbA== fDJlnXftKRmegCRxyFHu1pCZYEuU8MoMPZWGo0AMsWoXPgBBxVFMsta2M1PZj4W2VbO1bd8PuM0uVNs+KlvfhM72iXAAAAAAAAABnA==
B. Subtilis FP-enrichment	WUeR/4G7in+Yo6F9XIE9hl+UTYy5piSTHntvOKDNTxBB5Qez4T310D1hs96J0yTOaT5nxDaxue2isY61DBxhn6KL9FMAAAAAAAAA5mQ==
YbbR	WR9PrkNNr/OaENvsCm5KMpBH0Cm+z+Hw1r8JjPdM2bsaWW0jQfPLdFABtyBbN1edLJ4pr5o6l+AHLN33MdGOqRH/WKMAAAAAAAABuQ==
CouN5	0GkW5MFILmhQsiqmKyOLG2uzk8XcLLg0/90E7zfRkxsf78Cwbv0lQxnKbp94LCS9PrqsPwa6hFVuTEj/a/jiqNiEmgsAAAAAAAABwA==
Nrps	nH84rcA0wa3FN6xqAg3unkJJnnSpp7+AoANuQZsFfpl0FPM711skU7jTs98j2jkQNw1gB66KbSS2Svw0POj7iSNqa1cAAAAAAAAB0A==

InsPecT:PPant Database Search for Peptide Identification

In addition to the *B. subtilis* spectra, we evaluated InsPecT:PPant on the three recombinant proteins YbbR, Coun5, and Strop_4416 and found the delta scores of the true peptide sequence significantly outscored any peptide hits from the background peptide database (delta scores >40). The known peptide sequences of YbbR and CouN5 were added into the *B. Subtilis* proteome to run the database search protocol. The *Salinispora tropica* proteome from NCBI already included the Strop_4416 peptide and no special sequence additions were required for the database search. Also, these peptides were not treated with iodoacetamide and therefore had no substrate mass addition to PPant (Table S2, ***s***=0).

DS+ppantLEFIASKLA - YbbR
GILNS+ppantLNTAILVAH – CouN5
HDNFFDLGGHS+ppantLLAASLATR – Nrps_4416

MS2 PPant peptide Detection SVM and Alternative Methods

After PPant ion type features are annotated, we utilize these feature annotations to score whether a MS^2 spectrum represents a PPant peptide. We applied 5 different metrics to assess which are better detectors of PPant peptides in MS^2. Naively, the features can be combined into a score and applying a score threshold determines which spectra are PPant peptides. One common method is to assess the percent of relative abundance in the spectrum explained by feature (**Explained Intensity**). Relative abundance is known to vary greatly between replicate spectra and this metric typically does not perform well (observed in Figure 4). Similarly, we analyzed the number of expected PPant features found in the spectrum (**Number of Ions Present**). A better measure for intensity is to use the feature's rank in the spectrum rather than the explicit relative abundance value. This motivated us to use the sum of the feature intensity ranks (**Sum of Ion Intensity Rank**) as a score. The total number of spectrum peaks is used as a default value for expected PPant features not observed in the spectrum. The last metric evaluated was the sum of PPM error in observed features (**Sum of PPM Error**). The default value for expected features not observed is 800, the maximum tolerance given to annotate spectrum peaks. These metrics yielded some detection power, with the intensity rank performing the best.

While the intensity rank metric performs fairly well, we sought to increase our detection capability using SVM. This kernel-based method defines "support vectors" from two classes of examples that are used to predict the class of datum. SVM requires the optimal selection of parameters, kernel, and most importantly a feature set. We evaluated 3 normalized feature sets, 1) intensity rank features and ppm error features (22 features), 2) intensity rank features (11 features), and 3) intensity rank features and ppm error features without parent peptide neutral PPant ejection (14 features). Each feature value was normalized with the appropriate maximum value; divided by total number of spectrum peaks for intensity rank features and divided by 800 for ppm error features. We used the SVMLight implementation maintained by Thorsten Joachims to train classifiers and estimate leave one out errors.[1] Feature set 1) had the best performance by a small margin under the parameters of radial-basis kernel with a gamma of 0.5 and cost-factor of 1.0. We evaluated the linear, polynomial, and radial basis function kernel with cost-factors 2^i for i in [-5, 15]. Also, we evaluated degrees in the range of 2, 3, and 4, for the polynomial kernel and gammas in the range of 2^i for i in [-15, 2] for the radial-basis kernel. The optimal parameter set was selected by the SVM training that produced the best precision, recall, and error from SVMLight's leave-one-out estimate. As seen in Figure 4, the SVM approach outperforms **Sum of Ion Intensity Rank** metric.

References:

1. Joachims, T., *Estimating the Generalization Performance of a SVM Efficiently* Morgan Kaufmann Publishers Inc.: San Francisco, 2000; p 431-438.
2. Joachims, T., *Making Large-Scale SVM Learning Practical* LS8-Report, 24, Universität Dortmund, LS VIII-Report, 1998.