Proteome Informatics Research Group

iPRG2010: Informatic Evaluation of Phosphopeptide Identification and Phosphosite Localization Results From Multiple Proteomics Laboratories

Paul A. Rudnick[1], Manor Askenazi[2], Karl R. Clauser[3], William S. Lane[4], Lennart Martens[5], W. Hayes McDonald[6], Karen Meyer-Arendt[7], Brian C. Searle[8] and Jeffrey A. Kowalak[9]

[1]National Institute of Standards and Technology, Gaithersburg, MD, [2]Dana-Farber Cancer Institute, Boston, MA; [3]The Broad Institute of MIT and Harvard, Cambridge, MA; [4]Harvard University, Cambridge, MA; [5]Ghent University, Ghent, Belgium; [6]Vanderbilt University School of Medicine, Nashville, TN; [7]The University of Colorado, Boulder, CO; [8]Proteome Software, Inc. Portland, OR; [9]National Institute of Mental Health, Bethesda, MD

Abstract

Many LC-MS/MS-based phosphoproteomic studies are entering the literature which report thousands of phosphopeptide identifications. Current publication standards often do not require the certainty of the specified localizations to be reported. Since each peptide is often the only peptide observed for a particular protein, both confident identification and accurate localization of the site(s) of phosphorylation to a particular S, T, Y (or H) residue(s) is one of the major challenges in proteome informatics. In this work, the ABRF Proteome Informatics Research Group (iPRG) presents the results of a collaborative study focusing on the analysis of a common phosphoproteomic dataset

Study Goals

1. Evaluate the consistency of reporting phosphopeptide identifications and phosphosite localization across laboratories
2. Characterize the underlying reasons why result sets differ
3. Produce a phosphopeptide benchmark dataset, spectral library and analysis resource

Study Materials

1 Orbitrap XL dataset (3 files)

RAW, mzML, mzXML, MGF, pkl or dta

Data format conversions by ProteoWizard

1 fasta file (SwissProt human v57.1)

1 template (Excel)

1 on-line survey (Survey Monkey)

Study Instructions

1. Retrieve and analyze the dataset using your choice of informatics tools and methods
2. Report the phosphopeptide spectrum matches in the provided template
3. Complete an on-line survey
4. Attach a 1-2 page description of your methodology

Reporting Template

ABRF iPRG 2010 Study Template: Phosphorylated Peptide Analysis

Instructions: Please fill in all REQUIRED fields. After deleting the example rows, create a new row for each *phosphopeptide* spectrum match. Multiple rows MAY be used to report ambiguous phosphosite localizations. Phosphorylated residues MUST be indicated in the 'Peptide Sequence' field, and results should be sorted by 'Peptide Identification Score' from most to least confident. Additional instructions can be found above each field header. **Results should be emailed to 'anonymous.iPRG2010@gmail.com' no later than Jan. 10, 2010.** Please make sure to fill out the REQUIRED survey -------------------->

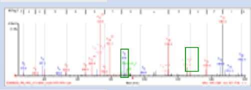
REQUIRED FIELDS

File	Spectrum Identifier	Precursor m/z	Precursor Charge	Peptide Sequence	Accession(s)	Num. Phospho sites	Peptide Identification Certainty	Phosphosite Localization Certainty	Peptide Identification Score
Name of data file (e.g., D20090930_PM_K562_SCX-IMAC_fxn03)	Identifiers should be unique scan numbers from data file but may also refer to a merged range of MS/MS scans (e.g., Scan:19, 2316.19.19.3.dta, 2316.19.19.3.pkl).	Precursor m/z as submited to search engine	Precursor charge reported by search engine	Use lowercase s, t or y (e.g. SLsGSsPCPK) OR a trailing symbol (e.g. SLS#GS#PCPK) OR a string in parentheses (e.g. SLS(ph)GS(ph)PCPK) immediately following each phosphorylated residue. Only phosphorylation of S, T and Y will be compared; all other modifications (e.g., oxidized M) will be ignored. It will be assumed that all modifications indicated on S, T or Y are phosphorylations.	Protein identifier(s) from Fasta file. Use multiple values if peptide is found in multiple proteins, e.g., Q9NZ18; Q9UQ35. Protein inference will not be scored.	Total number of phosphorylations as evidenced by the precursor m/z and MS2 spectrum.	'Y' indicates this match is BETTER than the confidence threshold. 'N' indicates the match is WORSE. Please report BOTH types of identifications in your ranked list. **Is this match above 1% FDR identification threshold (Y|N)?**	Indicate 'Y' if ALL phosphorylations have been confidently localized. 'N' if one or more have not. **Are ALL phosphosites unambiguously localized (Y|N)?**	Peptide identification score reported by search engine (e.g., E-value, p-value, probability, Mascot score, etc.)
D20090930_PM_K	Scan:908	558.7576	2	qGsPVAAGAPAK	Q9NZI8	1	Y	Y	0.0002097
D20090930_PM_K	Scan:2017	710.82233	2	TsPDPSPVSAAPSK	Q13469	1	Y	N	45.41
D20090930_PM_K	Scan:683	692.28891	2	_APQTS(ph)S(ph)SPPPVR_	Q8IYB3	2	Y	N	30.09

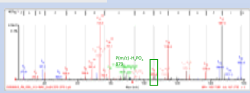
Id Loc

All participants were asked to report required information about each phosphopeptide spectrum match. Importantly, for each identification, participants were asked to report whether or not the match was above a 1% FDR confidence threshold (Id=Y|N) and whether or not ALL phosphosites were confidently localized. See Id and Loc in the template above.

DSAIPVEsDtDDEGAPR



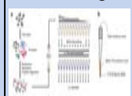
DSAIPVESDtDDEGAPR



Is There Enough Evidence In This Spectrum to Confidently Assign Localization of the Phosphosite?

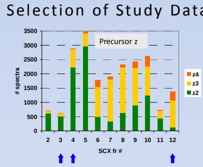
14 of 21 Participants Say No

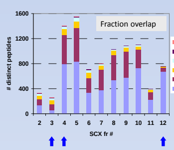
Generating a Phosphopeptide-Enriched Dataset

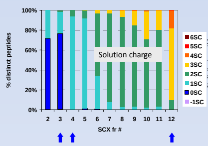


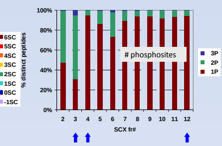
Sample: 7.5x10e7 human K562 human chronic myelogenous leukemia cells, 4mg lysate
Protocol: Villen, J, and Gygi, SP, Nat Prot, 2208, 3, 1630-1638.
Lysis: 8M urea, 75mM NaCl, 50 mM Tris pH 8.2, phosphatase inhibitors
SCX: PolyLC - Polysulfoethyl A 9.4 mm X 200mm, elute: 0-105mM KCl , 30% Acn .
IMAC: Sigma - PhosSelect Fe IMAC beads, bind: 40% Acn, 0.1% formic acid, elute: 500 mM K_2HPO_4 pH 7
MS/MS: Thermo Fisher Orbitrap XL, high-res MS1 scans in the Orbitrap (60k), Top-8 fragmented in LTQ, exclude +1 and precursors w/ unassigned charges, 20s exclusion time, precursor mass error +/- 10 ppm

Preliminary Analysis of All Fractions and Selection of Study Data Files



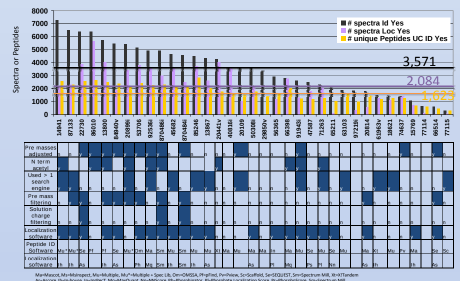






Because of the chemistry of the fractionation protocols, it was expected that the SCX fractions would yield different concentrations of phosphopeptides, multiply phosphorylated peptides and phosphopeptides with higher relative solution (and precursor) charges (below). Fractions 3, 4 and 12 were chosen because they differed in these properties. It was expected that participants using these properties (as selective filters) might be able to advantageously lower false positive identifications in a target-decoy search.

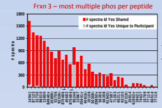
How Many Phosphopeptides By Each Participant?



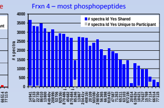
Ma=Mascot, Ms=MsInspect, Mu=Multiple, Mu*=Multiple + Spec Lib, Om=OMSSA, Pf=pFind, Pv=Pview, Sc=Scaffold, Se=SEQUEST, Sm=Spectrum Mill, Xt=X!Tandem
As=Ascore, Ih=In-house, In=InsPecT, Mq=MaxQuant, Nn=NNScore, Ph=Phosphinator, Pl=Phosphate Localization Score, Ps=PhosphoScore, Sm=Spectrum Mill

How Many Id's Were Unique to One Participant?

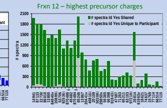
Frxn 3 – most multiple phos per peptide



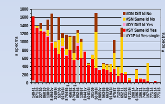
Frxn 4 – most phosphopeptides

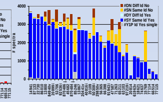


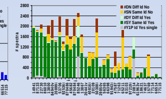
Frxn 12 – highest precursor charges



Some Participants Were Able to Successfully Lower Their Identification Thresholds





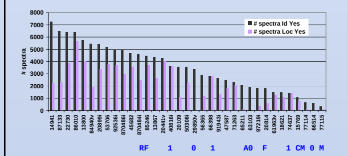


Data Analysis Resource Produced For Participants

YY: Y – identification Y – localization
YN: Y – identification N – localization
NS: N – identification, but top sequence same as matched by most YY & YN participants
ND: N – identification, and top sequence different from most YY & YN participants

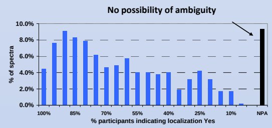
The spreadsheet at the right was generated from the analysis of participant results. It contains per participant results for each spectrum along with the consensus peptide assignments.

Not All Results Could Be Equally Compared For Phosphosite Localizations



Reason Excluded
0 0% localization
1 100% localization
F FDR - very high?
R Replicate submission
M Merged spectra
C Categorization Errors
A Y Loc only when no possible ambiguity

If Participants Agree On the Identity, Do They Agree on the Ability to Confidently Localize All Phosphosites?



Frxn 4 Subset of 472 spectra for which 20/22 participants all agree on Identity

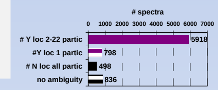
For the reasons indicated below the participant codes, some result sets were not examined during the localization analysis. This reduced the participant dataset count from 35 to 22. The subsequent analysis presented above and below was performed on those 22.

If They Agree On the Identity and That it Can Be Fully Localized (≥2), Do They Agree On the Localization(s)?

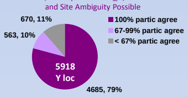
For all of the participants that agree on identity when
- site ambiguity is possible (#S,T,Y > # phos)
- ≥2 participants mark Loc=Y

➢ For 79% (4,685 of 5,918) of the spectra, all participants who mark Loc=Y unanimously agree on the localization of the phosphosites

8050 spectra with ≥ 2/22 Id Yes (Frxn 3, 4, 12)

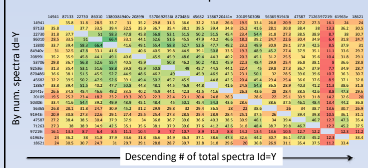


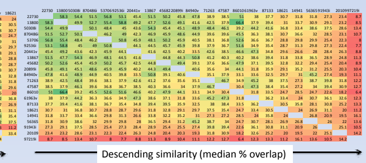
5918/8050 spectra with ≥ 2/22 Loc Yes and Site Ambiguity Possible



How Much Do the Confidently Identified, Confidently Localized Phosphopeptide Identifications Overlap?

n=22





Each set of confidently identified (Id=Y) and localized phosphopeptides (Loc=Y) was compared to those from every other participant. Values shown at left are percentage overlap.

Conclusions

1. The number of phosphopeptide identifications and localizations varied widely
2. Some participants succeeded without the use of localization software
3. Very few identifications were unique to one participant
4. Not all localization results could be uniformly compared, indicating a need for an analysis and/or scoring standard
5. For 79% of the spectra with the same identification agreed upon by all participants and >2 participants willing to mark confidently localized, the participants unanimously agreed on the localization of the phosphosite(s)

Acknowledgements

Phillip Mertins, The Broad Institute – All wet lab work and an analysis
Steve Gygi, Harvard Medical School – Test datasets
Matthew Chambers, Vanderbilt University Medical Center – RAW Data conversions (ProteoWizard)
Steve Stein and Yuri Mirokhin, NIST – spectral library
Renee Robinson, Harvard University -"The Anonymizer"