Phosphoproteome analysis by Mass Spectrometry (MS)

A short guide to data exploration and interpretation

Protein Analysis Facility
University of Lausanne
Faculty of Biology and Medicine
Lausanne, Switzerland

www.unil.ch/paf/

UNIL | Université de Lausanne

PAF LAUSANNE PROTEIN ANALYSIS FACILITY

Version 3, 30.7.2023

Background information

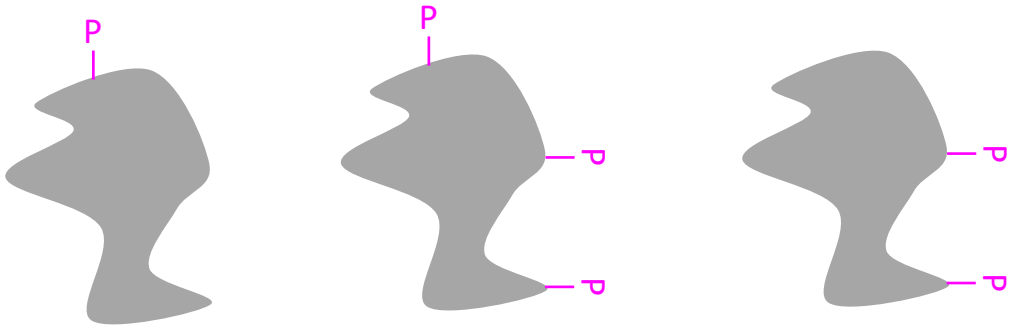
- For general discussion of quantitative proteomics data see our other tutorial (*PAF_Protein Quant_tutorial*) , which can be found on our web site (https://www.unil.ch/paf/)

Phosphoprotein biology

- Phosphorylation can mainly occur on Ser/Thr/Tyr
- Ser / Thr are frequent amino acids ➔ many possible modification sites in every protien sequence.
- Ideally : determine state of all phospho**sites**
- Protein activity is probably determined by combinations of modifications (→ different **proteoforms**), e.g :

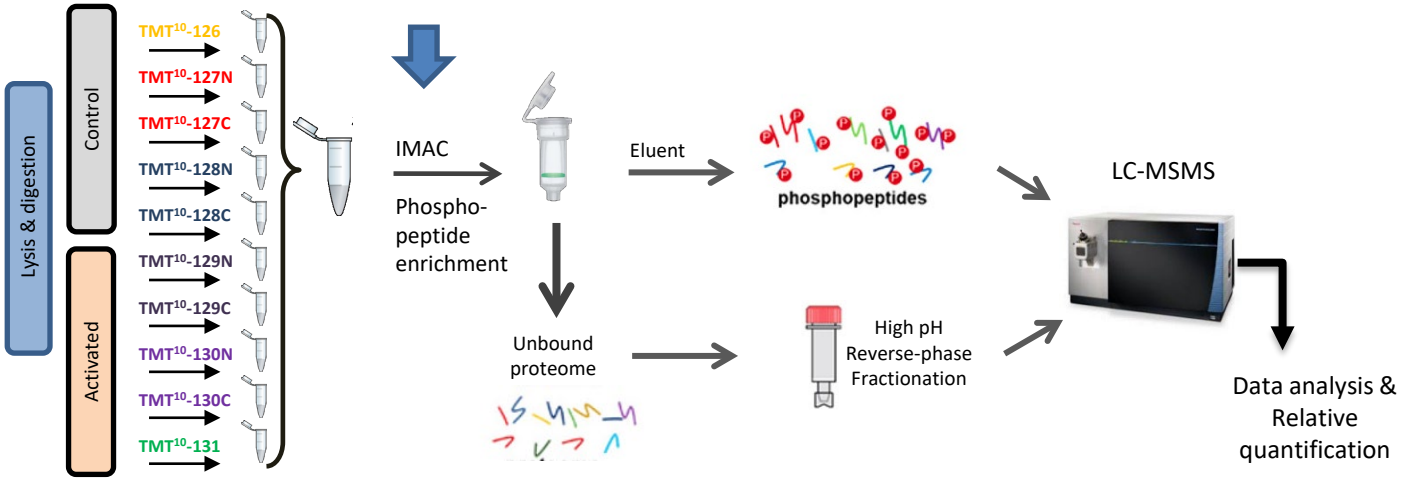
Example :

>Sp|P23528|COF1_HUMAN·Cofilin-1·
MASGVAVSDGVIKVFNDMKVRKSSTPEEVKKRKKAVLFC
LSEDKKNIILEEGKEILVGDVGQTVDDPYATFVKMLPDK
DCRYALYDATYETKESKKEDLVFIFWAPESAPLKSKMIY
ASSKDAIKKKLTGIKHELQANCYEEVKDRCTLAEKLGGS
AVISLEGKPL



Inactive active inactive

TMT-phospho workflow



- Extraction & Digestion
- Tandem Mass Tag (TMT) 10- or 16-plex labeling

J. Navarrete-Perea et al J. Proteome Res. 2018, 1, 17 (6) : 2226-2236

TMT-phospho workflow

- Need 200ug per sample
- Sample prep must preserve phosphorylation
- Batches of 10 or 16 samples

- 3'000-15'000 Phosphosites
- Total protein quantitation optional (recommended)

Ex. *2* conditions, 5 replicates
4-5 weeks

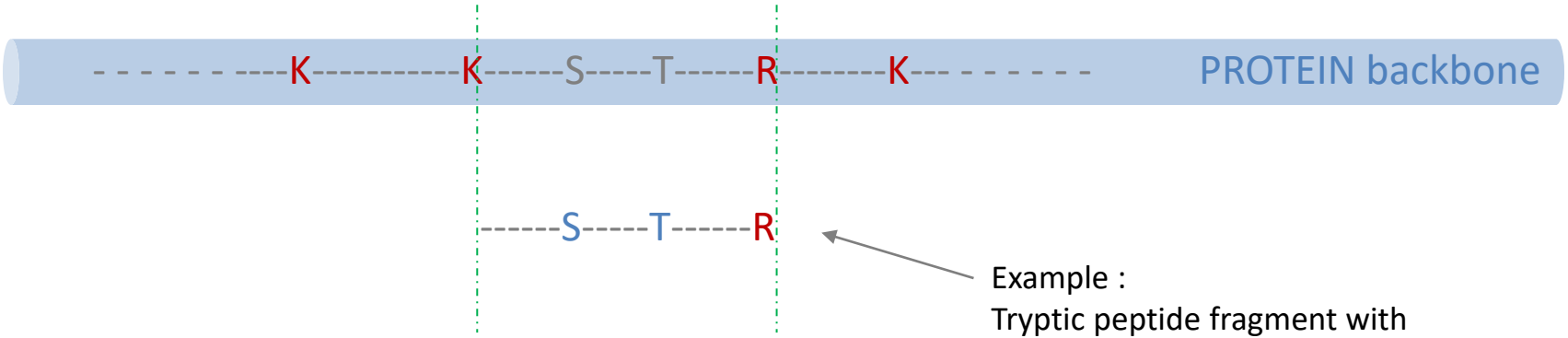
PROS

- P-peptide enrichment is the critical step
- Enrichment performed on all samples together
➔ Reproducible
- P-site localization is consistent (same MS2 spectrum)
- Few missing values (within same TMT mix)
- Multiplexing ➔ speed, efficiency

CONS

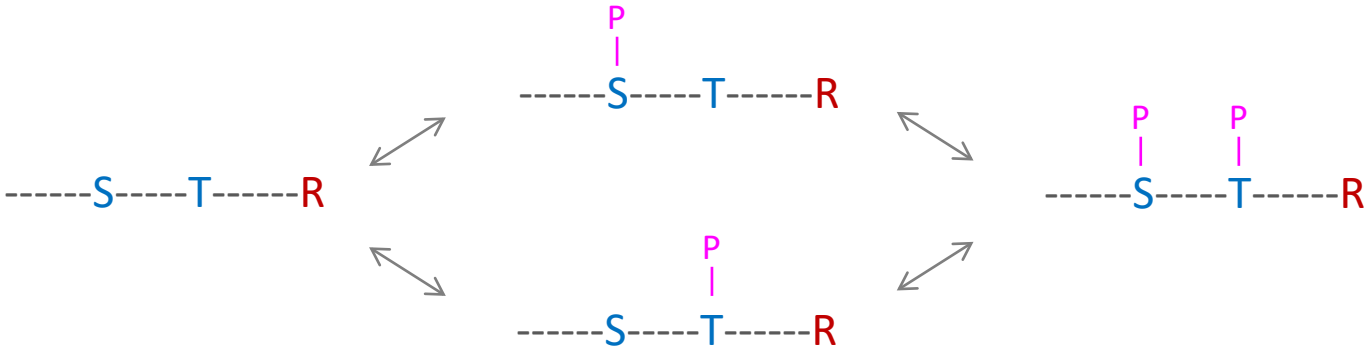
- Lower ID efficiency
- Reagent costs (1 kCHF / 16-plex)

Proteomics approaches are based on digestion with trypsin



Example :
Tryptic peptide fragment with
2 potential phosphorylation sites

Mono vs. di-phospho species

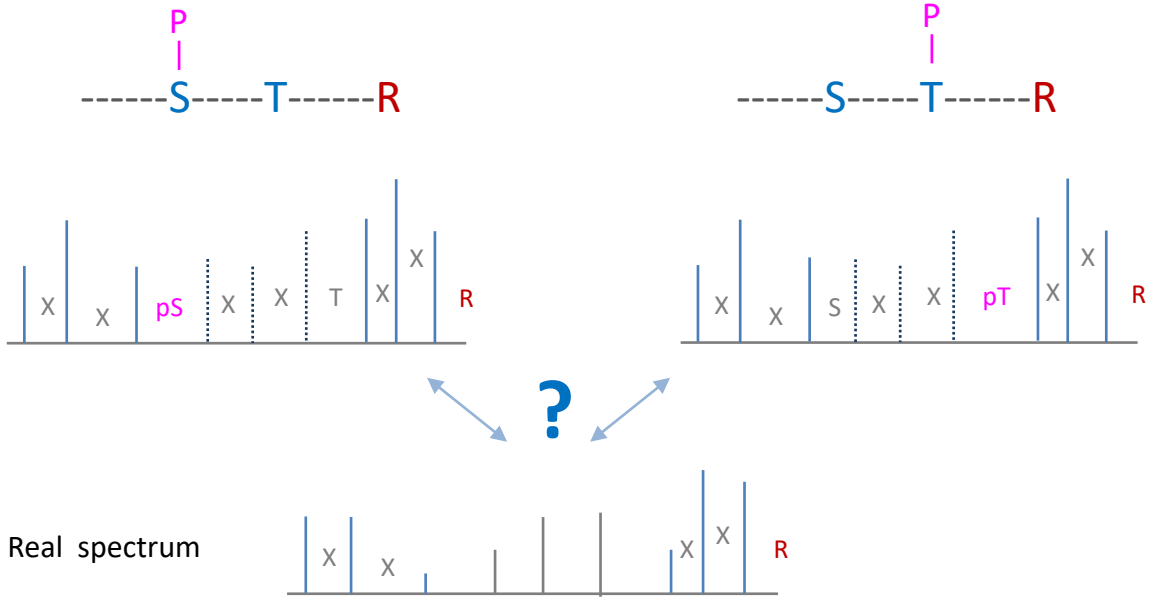


Technical facts :

- Mass Spec detects peptides => P-sites located on same peptide fragment after trypsin digestion are inevitably linked when it comes to detection.
- Identification of a phosphopeptide and exact localization of the modification(s) should, ideally, happen simultaneously. In practice, localisation may be more ambiguous.
- Positional phospho-isomers (the two above) are isobaric (= have exactly the same mass)

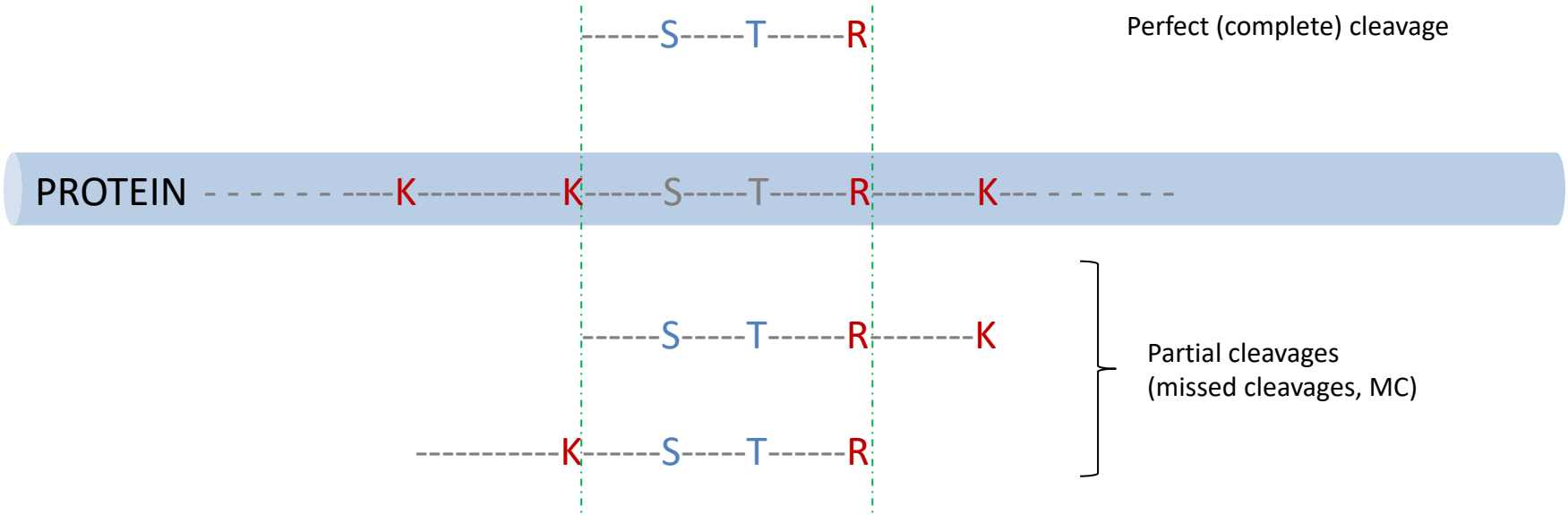
Phosphosite localization : finding the good match

Two Predicted Spectra



- Best fit to calculated model spectra determines highest scoring localisation pattern
- Exact localisation can be based on only one or a few fragments
- The closer to each other the potential P sites, the more difficult it is to discriminate them.

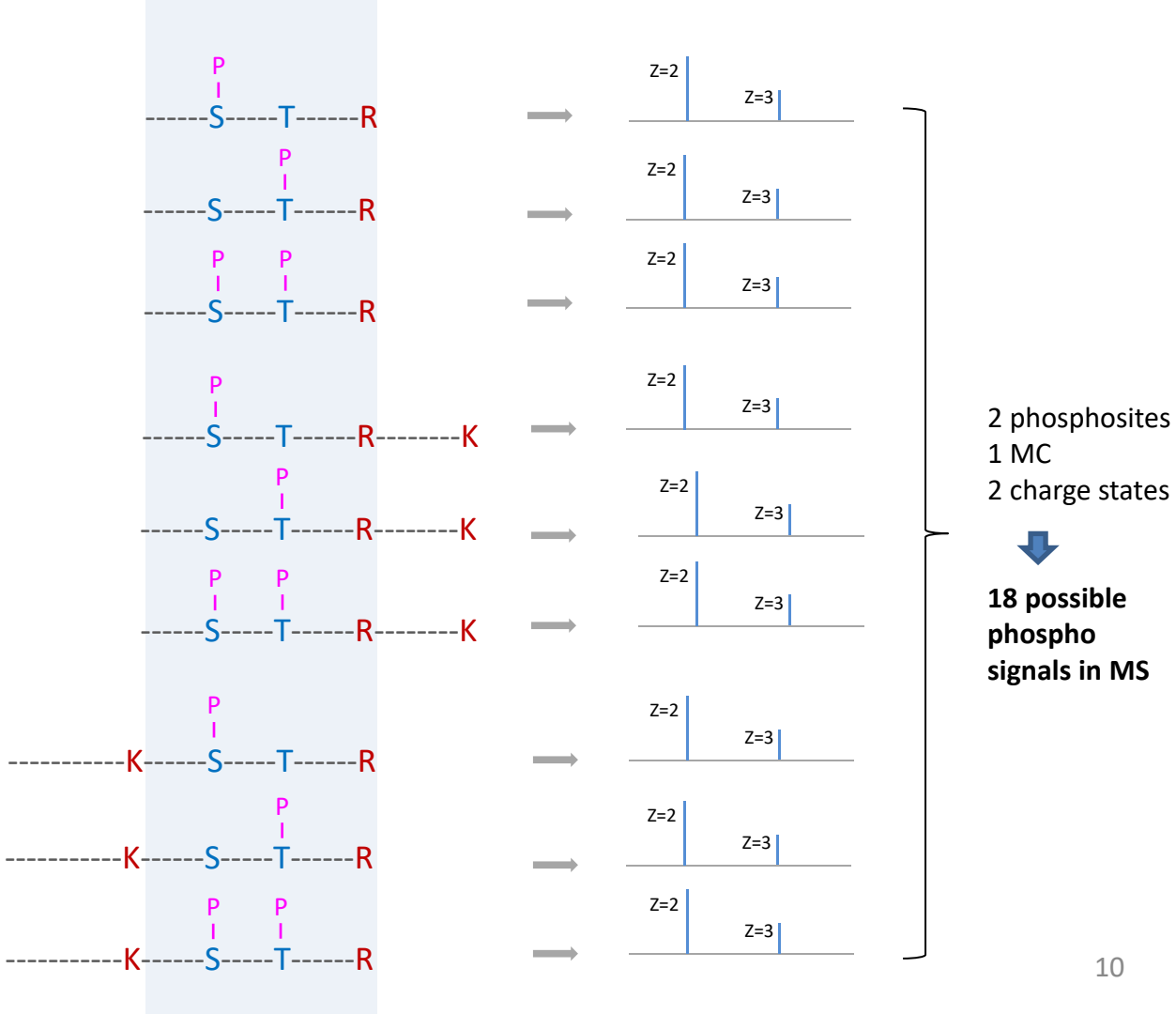
Phospho**sites** vs phosphopeptides: collecting all the evidence - I



FACT : Phospho-enriched samples have above-average percentages of missed cleavages.
Phospho groups are believed to inhibit trypsin cleavage in proximity

Phosphosites vs phosphopeptides: collecting all the evidence -II

- Peptide signals mapping to the same site have to be identified and combined for quantitation.
- Positional phosphoisomers must be discriminated when possible (must have different LC elution and sufficient MS2 information) and their quantitation kept separated .
- **Software packages exists to do this work. Nevertheless, the results are complex and sometimes ambiguous**



Analysis & Interpretation of Phosphoproteomics data

Caveats and things to remember

Sites :

- Phosphosite data (and in general PTM data) is based on identification and quantitation of **single** peptide species. Compared to total protein analysis, which usually results from the combined data for several peptides, it is **less robust** and less reproducible. The complexity of site localisation and the presence of multiple species (previous slide) can add more noise. Thus, obtaining statistically significant quantitative measurements is more challenging for phosphosites. More replicates should be used if this is the goal.

- Although this is unfrequent, phosphosites can **occur in several protein sequences**, making assignment ambiguous (ex. MAPK family).

- Different phosphosites on the **same protein** can (quantitatively) **change in divergent ways**

- **Singly/multiply** phosphorylated forms of the same site can coexist and complicate quantitative conclusions.

Understanding the phosphosite table (e.g. sorted by Gene Name) – **part I**

Fold change **(log2)**

Raw intensities **(log2)**

Nr of phospho groups considered

Pos of P group in protein sequence. Always only one position given (s. below)

Database accession (protein)

Probabilities (0…1) of localization on each site

Intensity 12797 TiO_x-y_Intensity 12796 TiO	Intensity 12796 TiO	Intensity 12797 TiO	Amino acid	Charge	Identification ty	Identification typ	Multiplicity	Localization prob	PEP	Score	Delta score	Score for localization	Mass error [ppm]	Intensity	Position	Localization prob 12796 TiO	Localization prob 12797 TiO	id	Proteins	Positions within proteins	Leading proteins	Protein	Protein names	Gene names	Sequence window	Phospho (STY) Probabilities
4.27	-3.82	0.45	T	2	By	By	___1	0.81	0.001	69.6	42.49	69.6	0.033	22.5	604	0.00	0.81	4883	Q3U	604	Q3U	Q3UHJ0	AP2-associated protein kinase 1	Aak1	GQIQAPV	VQT(0.189
1.50	-1.85	-0.36	T	2	By	By	___1	1.00	#####	147.7	137	147.7	-0.41	24.8	618	1.00	1.00	4882	Q3U	618	Q3U	Q3UHJ0	AP2-associated protein kinase 1	Aak1	TTPPPTIQ	VGSLT(1)P
-0.43	1.29	0.86	T	2	By	By	___2	1.00	#####	147.7	137	147.7	-0.41	24.8	618	1.00	1.00	4882	Q3U	618	Q3U	Q3UHJ0	AP2-associated protein kinase 1	Aak1	TTPPPTIQ	VGSLT(1)P
-0.43	1.29	0.86	S	2	By	By	___2	0.59	#####	113.7	99.61	113.7	0.018	24.5	621	0.59	0.58	1539	Q3U	621	Q3U	Q3UHJ0	AP2-associated protein kinase 1	Aak1	PPTIQGQK	VGSLT(1)P
0.65	-1.64	-0.99	S	3	By	By	___1	1.00	#####	99.39	95.73	99.39	0.185	22	88	1.00	1.00	1441	Q3TI	88	Q3TH	Q3THG9	Alanyl-tRNA editing protein Aarsd1	Aarsd1	TRRGAQA	GAQADHF
0.01	1.27	1.28	S	2	By	By	___1	1.00	#####	135.1	121.1	135.1	0.34	25.3	409	1.00	1.00	1443	Q3TI	409	Q3TH	Q3THG9	Alanyl-tRNA editing protein Aarsd1	Aarsd1	RAEAQALI	RAEAQALI
-0.35	1.00	0.65	S	3	By	By	___1	0.99	#####	99.85	94.01	99.85	0.171	24.2	659	0.99	0.00	617	P067	659	P067	P06795	Multidrug resistance protein 1B	Abcb1b	QSDTDASI	GNEIEPGN
0.13	-1.38	-1.25	S	2	By	By	___1	0.55	#####	81.43	68	67.46	0.007	22	289	0.50	0.55	415	O35	289	O353	O35379	Multidrug resistance-associated protein 1	Abcc1	RIVYAPPK	GS(0.555)S
0.61	3.38	4.00	S	2	By	By	___1	1.00	#####	171.1	157.9	150.6	-0.17	28.4	107	1.00	1.00	2343	Q6P	107	Q6P5	Q6P542	ATP-binding cassette sub-family F membe	Abcf1	ERVLMERL	QLSVPAS(
-0.04	4.12	4.08	S	3	By	By	___1	1.00	#####	133.2	120.9	46.58	0.083	27.7	138	1.00	1.00	2342	Q6P	138	Q6P5	Q6P542	ATP-binding cassette sub-family F membe	Abcf1	KAKGGNV	GGNVFEA
0.84	0.94	1.79	S	3	By	By	___1	1.00	#####	168.8	161.8	168.8	-0.24	25	194	1.00	1.00	2345	Q6P	194	Q6P5	Q6P542	ATP-binding cassette sub-family F membe	Abcf1	EKSKGKAK	SKPAAAD
0.69	4.03	4.71	S	2	By	By	___1	1.00	#####	217.7	217.7	186.1	-0.12	29.2	475	1.00	1.00	3419	Q8K	475	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	DKQERQS	TLS(1)PTP
-1.05	-0.74	-1.79	S	2	By	By	___2	1.00	#####	217.7	217.7	186.1	-0.12	29.2	475	1.00	1.00	3419	Q8K	475	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	DKQERQS	TLS(1)PTP
-1.05	-0.74	-1.79	S	3	By	By	___2	0.77	3E-04	70.45	70.45	70.45	0.053	22.3	479	0.77	0.00	3420	Q8K	479	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	RQSLGESP	T(0.037)LS
0.41	1.64	2.05	S	2	By	By	___1	1.00	#####	159	135.8	159	0.054	25.2	496	0.97	1.00	3418	Q8K	496	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	EGYQDVR	S(0.002)T(
-0.49	0.96	0.46	S	2	By	By	___1	0.50	#####	86.29	71.03	86.29	0.193	24.2	670	0.50	0.50	3416	Q8K	670	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	GPPSLAAV	S(0.5)S(0.5
0.54	2.12	2.66	S	3	By	By	___1	0.96	#####	153.1	117.6	153.1	-0.04	26.4	671	0.87	0.96	3417	Q8K	671	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	PPSLAAVC	RS(0.041)S
1.09	-1.37	-0.29	S	2	By	By	___1	1.00	3E-04	86.9	71.75	86.9	0.292	22.5	738	1.00	1.00	3421	Q8K	738	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	SPLHSASH	TSS(1)LPG

- All intensity values are **in log2** scale. We recommend to read first our general tutorial on quantitative proteomics (https://www.unil.ch/paf/)

- We typically perform statistical tests (T-test) to determine sites that change in a significant way (p- and q-values; not done in this table). Independently of the p-value obtained, the test output includes the difference in the averages of the conditions considered. This is in columns named «**Student's T-test difference…»** and is equivalent to a fold-change value. These values, too are **log2 scale.**

- For additional complexities specific to the phosphosites table : see next slide

Understanding the phosphosite table (e.g. sorted by Gene Name) – **part II**

Nr of phospho Groups considered

Pos of P group in protein Sequence. Always only one position given (s. below)

Probabilities (0…1) of localization on each site

Intensity 12797 TiO_x-y_Intensity 12796 TiO	Intensity 12796 TiO	Intensity 12797 TiO	Amino acid	Charge	Identification ty	Identification typ	Multiplicity	Localization prob	PEP	Score	Delta score	Score for localization	Mass error [ppm]	Intensity	Position	Localization prob 12796 TiO	Localization prob 12797 TiO	id	Proteins	Positions within proteins	Leading proteins	Protein	Protein names	Gene names	Sequence window	Phospho (STY) Probabilities
4.27	-3.82	0.45	T	2	By M	By M	__1	0.81	0.001	69.6	42.49	69.6	0.033	22.5	604	0.00	0.81	4883	Q3U	604	Q3UH	Q3UHJ0	AP2-associated protein kinase 1	Aak1	GQIQAPV	VQT(0.189
1.50	-1.85	-0.36	T	2	By M	By M	__1	1.00	#####	147.7	137	147.7	-0.41	24.8	618	1.00	1.00	4882	Q3U	618	Q3UH	Q3UHJ0	AP2-associated protein kinase 1	Aak1	TTPPPTIQ	VGSLT(1)F
-0.43	1.29	0.86	T	2	By M	By M	__2	1.00	#####	147.7	137	147.7	-0.41	24.8	618	1.00	1.00	4882	Q3U	618	Q3UH	Q3UHJ0	AP2-associated protein kinase 1	Aak1	TTPPPTIQ	VGSLT(1)F
-0.43	1.29	0.86	S	2	By M	By M	__2	0.59	#####	113.7	99.61	113.7	0.018	24.5	621	0.59	0.58	1539	Q3U	621	Q3UH	Q3UHJ0	AP2-associated protein kinase 1	Aak1	PPTIQGQK	VGSLT(1)F
0.65	-1.64	-0.99	S	3	By M	By M	__1	1.00	#####	99.39	95.73	99.39	0.185	22	88	1.00	1.00	1441	Q3TI	88	Q3TH	Q3THG9	Alanyl-tRNA editing protein Aarsd1	Aarsd1	TRRGAQA	GAQADHF
0.01	1.27	1.28	S	2	By M	By M	__1	1.00	#####	135.1	121.1	135.1	0.34	25.3	409	1.00	1.00	1443	Q3TI	409	Q3TH	Q3THG9	Alanyl-tRNA editing protein Aarsd1	Aarsd1	RAEAQALI	RAEAQALI
-0.35	1.00	0.65	S	3	By M	By M	__1	0.99	#####	99.85	94.01	99.85	0.171	24.2	659	0.99	0.00	617	P067	659	P067	P06795	Multidrug resistance protein 1B	Abcb1b	QSDTDASI	GNEIEPGN
0.13	-1.38	-1.25	S	2	By M	By M	__1	0.55	#####	81.43	68	67.46	0.007	22	289	0.50	0.55	415	O35	289	O353	O35379	Multidrug resistance-associated protein 1	Abcc1	RIVYAPPK	GS(0.555)S
0.61	3.38	4.00	S	2	By M	By M	__1	1.00	#####	171.1	157.9	150.6	-0.17	28.4	107	1.00	1.00	2343	Q6P	107	Q6P5	Q6P542	ATP-binding cassette sub-family F membe	Abcf1	ERVLMERI	QLSVPAS(
-0.04	4.12	4.08	S	3	By M	By M	__1	1.00	#####	133.2	120.9	46.58	0.083	27.7	138	1.00	1.00	2342	Q6P	138	Q6P5	Q6P542	ATP-binding cassette sub-family F membe	Abcf1	KAKGGNV	GGNVFEA
0.84	0.94	1.79	S	3	By M	By M	__1	1.00	#####	168.8	161.8	168.8	-0.24	25	194	1.00	1.00	2345	Q6P	194	Q6P5	Q6P542	ATP-binding cassette sub-family F membe	Abcf1	EKSKGKAK	SKPAAADS
0.69	4.03	4.71	S	2	By M	By M	__1	1.00	#####	217.7	217.7	186.1	-0.12	29.2	475	1.00	1.00	3419	Q8K	475	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	DKQERQS	TLS(1)PTP
-1.05	-0.74	-1.79	S	2	By M	By M	__2	1.00	#####	217.7	217.7	186.1	-0.12	29.2	475	1.00	1.00	3419	Q8K	475	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	DKQERQS	TLS(1)PTP
-1.05	-0.74	-1.79	S	3	By M	By M	__2	0.77	3E-04	70.45	70.45	70.45	0.053	22.3	479	0.77	0.00	3420	Q8K	479	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	RQSLGESP	T(0.037)LS
0.41	1.64	2.05	S	2	By M	By M	__1	1.00	#####	159	135.8	159	0.054	25.2	496	0.97	1.00	3418	Q8K	496	Q8K4	Q8K4G5	Actin-binding LIM protein 1	Ablim1	EGYQDVR	S(0.002)T(

??? What is going on here ??? :

Two phosphorylations : pos 475 and 479; these sites happen to be on same tryptic peptide

P-sites actually detected by MS :

1) mono-phospho peptide, P group group on 475 : increasing in sample 12797 (+0.69)
2) di-phospho peptide , P group on 475 and 479 : decreasing in 12797 (-1.05)
3) phosphosite on pos. 479 is only listed once as a 2xP site→ a mono-P peptide with modification at 479 was not identified

When the table is in this format, the site 475 is listed twice , once with «Multiplicity» = __1 and once with «Multiplicity» = __2. This is only to make clear that this site is reported in two lines, quantitated separately. Note : for both lines with «Multiplicity» = __2, the same quantitative values are present (-1.05) as they are linked. Indeed, they derive from the same quantitated , 2x phosphorylated peptide.

What we observe here is probably a dephosphorylation : the mono-P peptide increases due to a decrease of the 2xP species

Analysis & Interpretation of Phosphoproteomics data

- Biological Annotation can be added and used for interpretation purposes :
  - GOBP, GOCC, GOMF , KEGG, Reactome, … : standard annotation of proteins
  - Linear kinase **motifs** : (added by Perseus software) based on sequence window (+/-15 AA) surrounding the phospho-site. *Although this can be very useful, such assignments are often quite permissive, as almost any site gets one or more kinase assigned*

- Annotation enrichment analysis : see tutorial on quantitative proteomics* for background information. Enrichment analysis can be performed on phosphosite tables, on the basis of fold –changes. However the situation is different relative to protein tables, because a single protein may be present with many phosphosites, introducing a bias. To calculate this enrichment stringently an adjustment per protein can be done (e.g. in Perseus). Without this, the analysis can nevertheless provide indications on activated pathways, if we consider phosphosites as distinct entities that can be regulated independently (though this is debatable).

- Unfortunately our software at the moment does not have an easy function to export automatically the list of protein ID's/gene names/phosphosites that are linked to each annotation term. This can be done manually for selected terms using Excel. See slides in *

* PAF_Protein Quant_tutorial_vxx.pdf