New Advances in Proteomics Informatics for the Trans-Proteomic Pipeline

Luis Mendoza, David Shteynberg,
Michael Hoopmann, Eric Deutsch

ASMS evening Workshop
June 4th, 2019 – A304

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Agenda

- **Introduction to TPP and overview of the latest features in 5.2**
- Proteogenomics & PEFF applications
- Analyzing PTMs with PTMProphet
- Cross-linking analysis with Kojak 2.0
- Deploying the TPP using Docker containers and cloud computing platforms
- Open Discussion and Feedback

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Group presentations at ASMS

For more info:

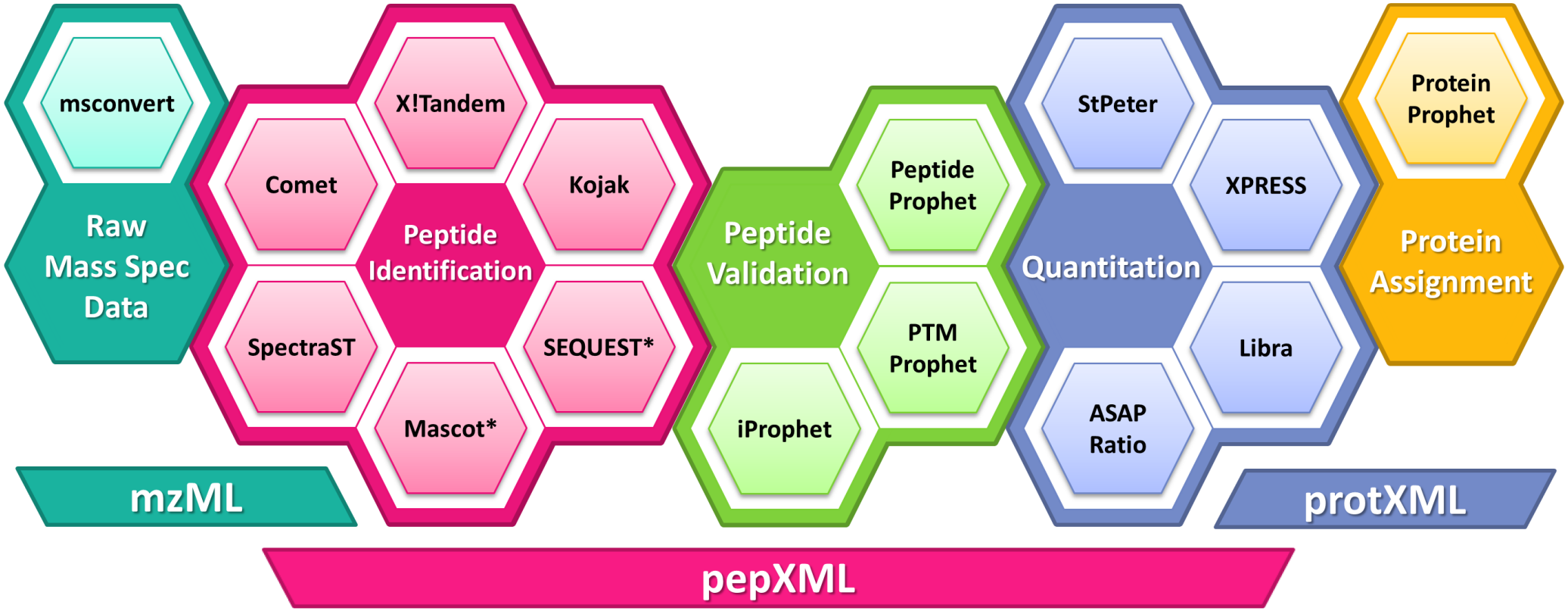
- Kojak 2.0 (MP 044)
- PEFF in TPP (MP 417)
- PTMProphet (WP 400)

Booth #118

www.tppms.org/asms/

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Trans-Proteomic Pipeline (TPP) Overview



Free and **open source** suite of **software tools** and **file formats** that facilitates and standardizes proteomics analysis

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

TPP components

- PSM and Peptide validation
- Protein inference
- Quantification
- Spectral Libraries: build and search
- Combine search engines
- Cloud capabilities
- etc…

www.tppms.org

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

New features in 5.2.0 (March‘19)

- **tpp2mzid**, a new tool for converting *PepXML* and *ProtXML* to *mzIdentML*.
- New **ProteoMapper** tools: **clips** and **promast**, for fast and efficient mapping of peptide sequences to a proteome
  - Supports variants via *PEFF* format
  - Able to do fuzzy matching of sequences, including isobaric (with and without mass modifications)
- New quantitation tool **QUANTIC**, which uses matched ion intensities of identified (MS/MS) spectra
- New tool to generate quantitative **volcano plots** on Proteins and Peptides
- Added a simple utility to **compute p-values** for some common distributions using GSL
- Support for *Magnum* and *Morpheus* search engines
- Expanded UI support for up to 12 channels of isobaric labeling, including **TMT-11**

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Agenda

- Introduction to TPP and overview of the latest features in 5.2
- **Proteogenomics & PEFF applications**
- Analyzing PTMs with PTMProphet
- Cross-linking analysis with Kojak 2.0
- Deploying the TPP using Docker containers and cloud computing platforms
- Open Discussion and Feedback

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

What is PEFF?

- PEFF is a “unified format for protein sequence databases to be used by sequence search engines and other associated tools, to enable consistent extraction, display and processing of information such as **post-translational modifications, mutations** and other processing events,” among others.

- The format is plain text, largely FASTA-like for backwards compatibility.

psidev.info/peff

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PEFF Basics

"\" Keywords

```
>nxp:NX_Q5EE01-1 \PName=Centromere protein W isoform Iso 1
\GName=CENPW \NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=88
\SV=61 \EV=265 \PE=1 \VariantSimple=(4|L)(6|M)(6|V)(8|P)(8|F)
(11|R)(19|H)(19|C)(20|D)(24|Q)(28|L)(28|P)(31|R)(32|*)(40|N)(41|F)
(45|V)(47|F)(52|R)(53|*)(53|Q)(57|D)(59|G)(63|F)(64|V)(12|H)(26|C)
(62|T)(63|S)(74|R)(78|T)(80|M)(86|I)(86|G) \Processed=(1|88|mature
protein)
MALSTIVSQRKQIKRKAPRGFLKRVFKRKKPQLRLEKSGDLLVH
LNCLLFVHRLAEESRTNACASKCRVINKEHVLAAAKVILKKSRG
>nxp:NX_Q5EE01-1 \DbUniqueId=NX_Q5EE01-1 ...
...
```

Sequence

PEFF entry

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Single Amino Acid Variants in PEFF

```
>nxp:NX_Q5EE01-1 \PName=Centromere protein W isoform Iso 1
\GName=CENPW \NcbiTaxId=9606 \TaxName=Homo Sapiens \Length=88
\SV=61 \EV=265 \PE=1 \VariantSimple=(4|L)(6|M)(6|V)(8|P)(8|F)
(11|R)(19|H)(19|C)(20|D)(24|Q)(28|L)(28|P)(31|R)(32|*)(40|N)(41|F)
(45|V)(47|F)(52|R)(53|*)(53|Q)(57|D)(59|G)(63|F)(64|V)(12|H)(26|C)
(62|T)(63|S)(74|R)(78|T)(80|M)(86|I)(86|G) \Processed=(1|88|mature
protein)
MALSTIVSQRKQIKRKAPRGFLKRVFKRKKPQLRLEKSGDLLVH
LNCLLFVHRLAEESRTNACASKCRVINKEHVLAAAKVILKKSRG
>nxp:NX_Q5EE01-1 \DbUniqueId=NX_Q5EE01-1 ...
...
```

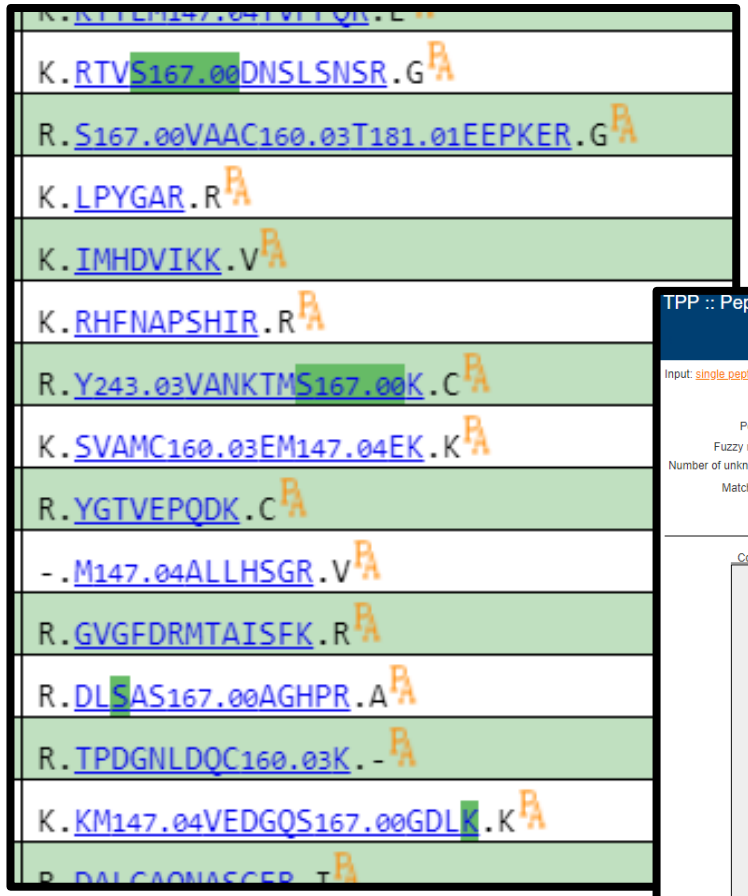
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

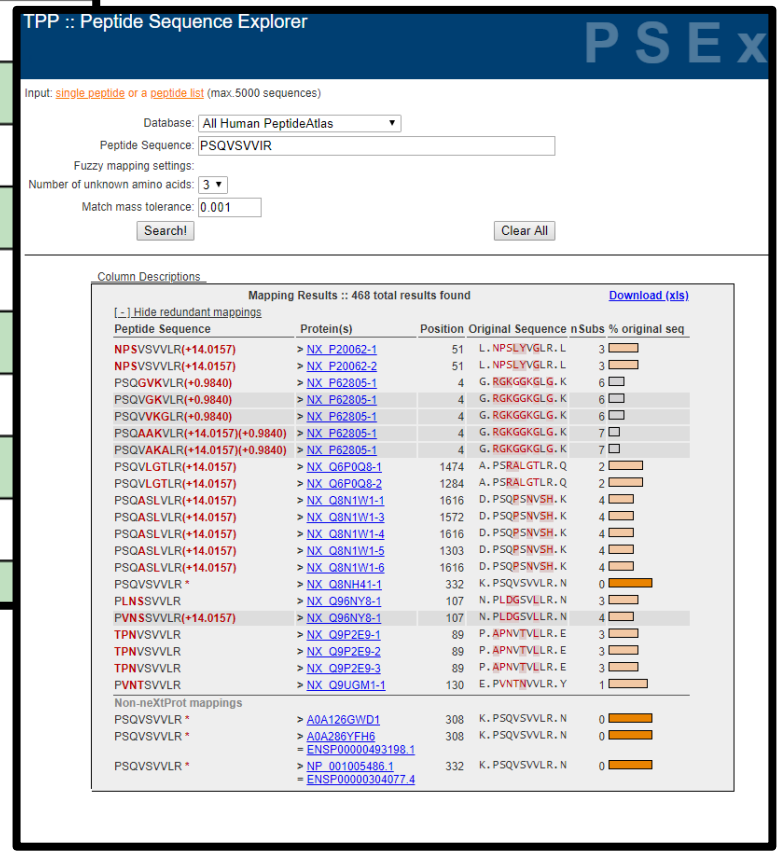
PEFF in TPP

- Search with Comet
- Results in *PepXML*
- Visualize, filter, sort, export…
- Map sequences
- Protein inference

In progress:

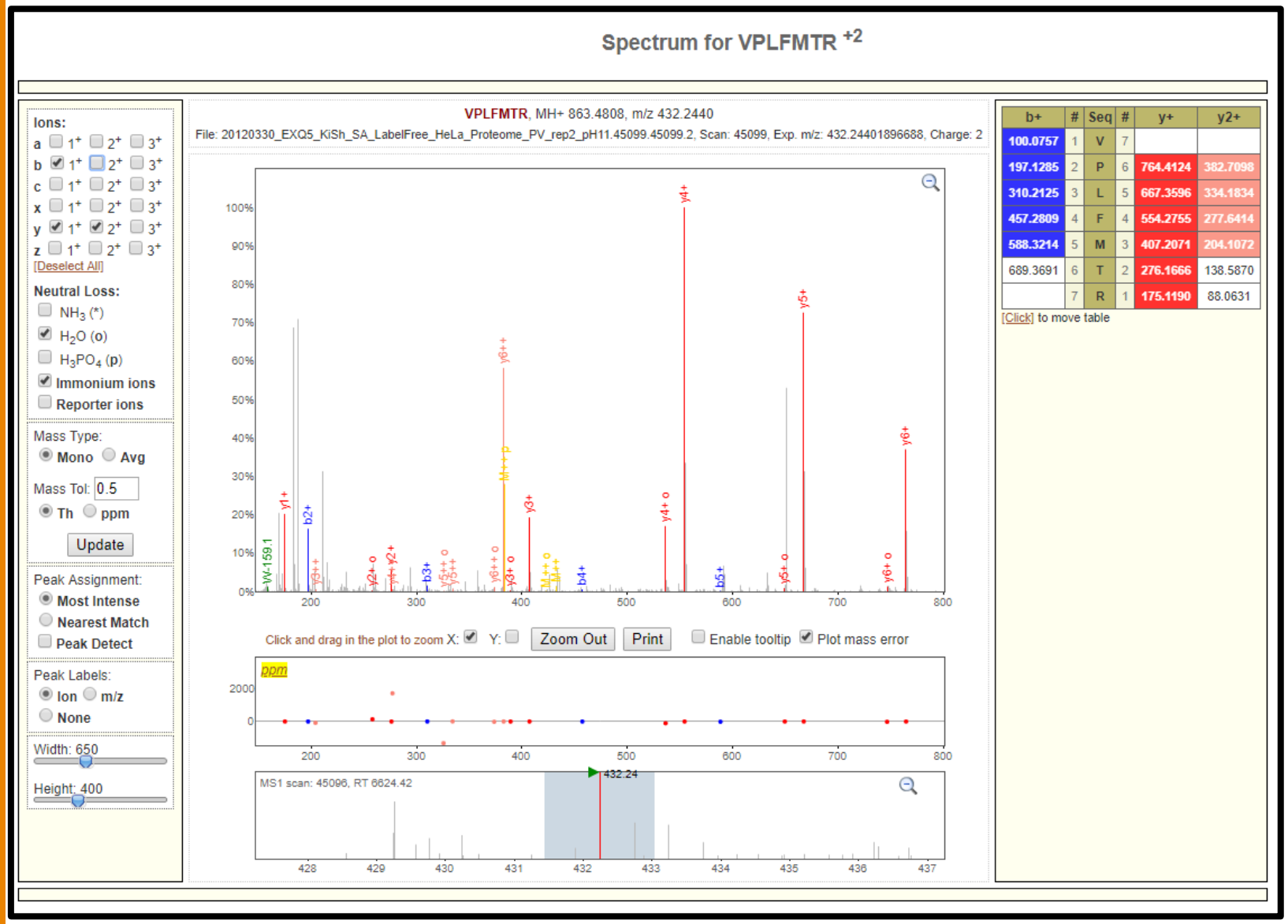
- Results validation (decoys…?)
- In/dels
- Signal / transit / etc
- Proteogenomics
- Make your own!
- Other…!?





Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PEFF in PeptideAtlas



b+	#	Seq	#	y+	y2+
100.0757	1	V	7		
197.1285	2	P	6	764.4124	382.7098
310.2125	3	L	5	667.3596	334.1834
457.2809	4	F	4	554.2755	277.6414
588.3214	5	M	3	407.2071	204.1072
689.3691	6	T	2	276.1666	138.5870
	7	R	1	175.1190	88.0631

USI:

Canonical Protein sequence: K.VPLFMSR.A

Observed Peptide sequence: K.VPLFMTR.A

Known variant dbSNP:rs1147990

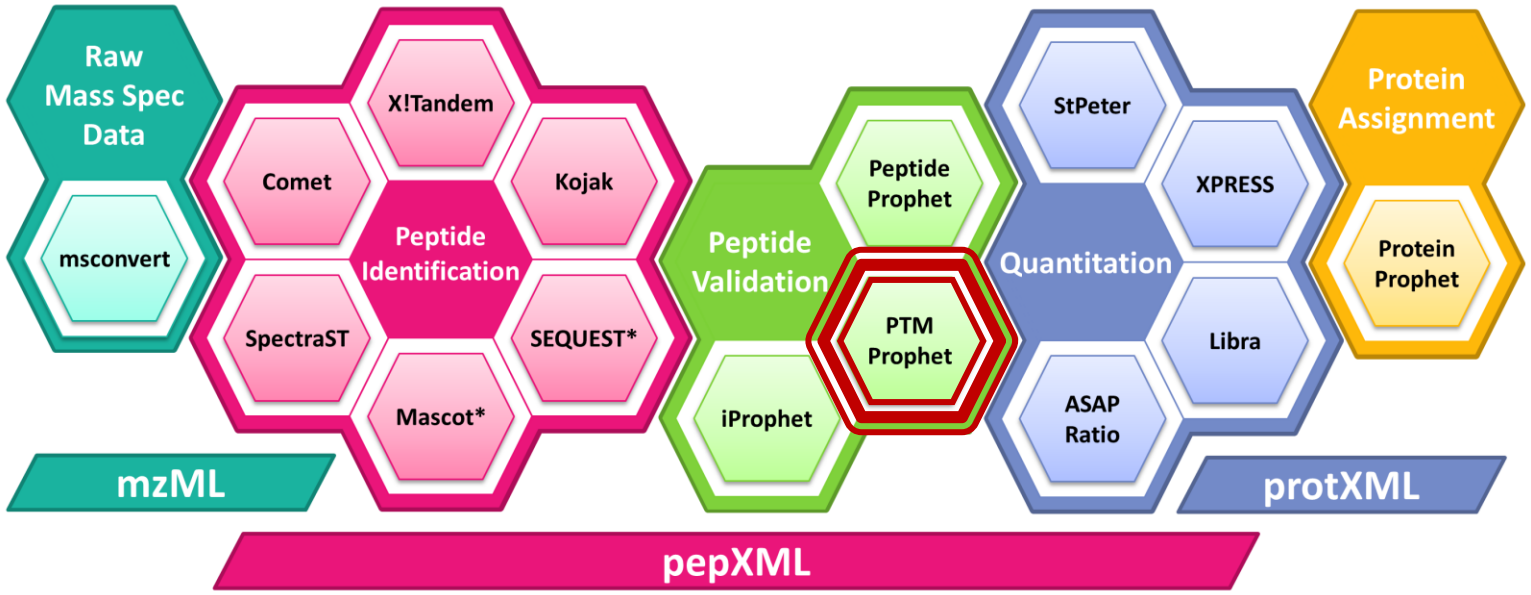
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Agenda

- Introduction to TPP and overview of the latest features in 5.2
- Proteogenomics & PEFF applications
- **Analyzing PTMs with PTMProphet**
- Cross-linking analysis with Kojak 2.0
- Deploying the TPP using Docker containers and cloud computing platforms
- Open Discussion and Feedback

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PTMProphet: Localizing PTMs



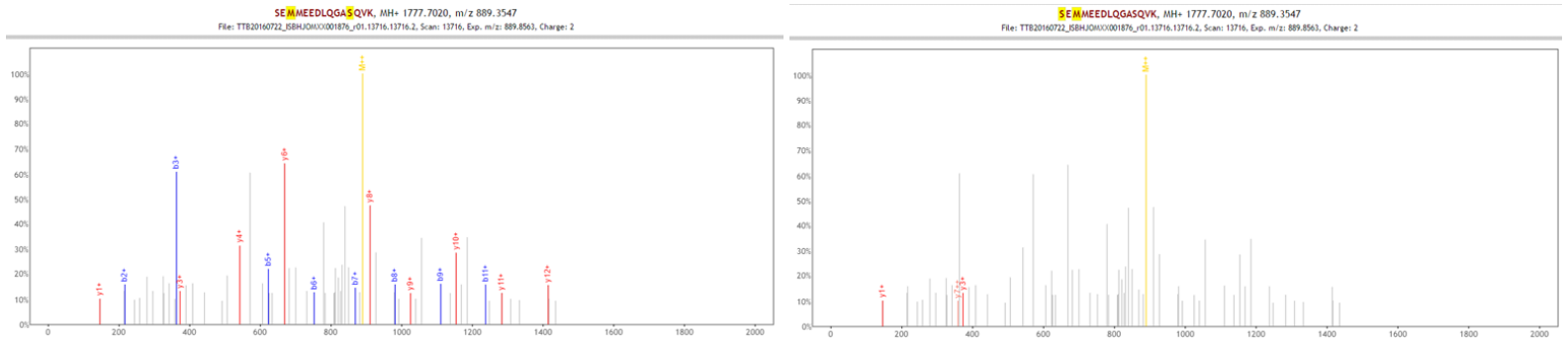
GOALS:

1. For each Modified PSM, Compute Accurate Probabilities of each potential modification site being modified
2. Compute **Information Content** thereby allowing comparing PSMs having different numbers of potential modification sites and number of modifications

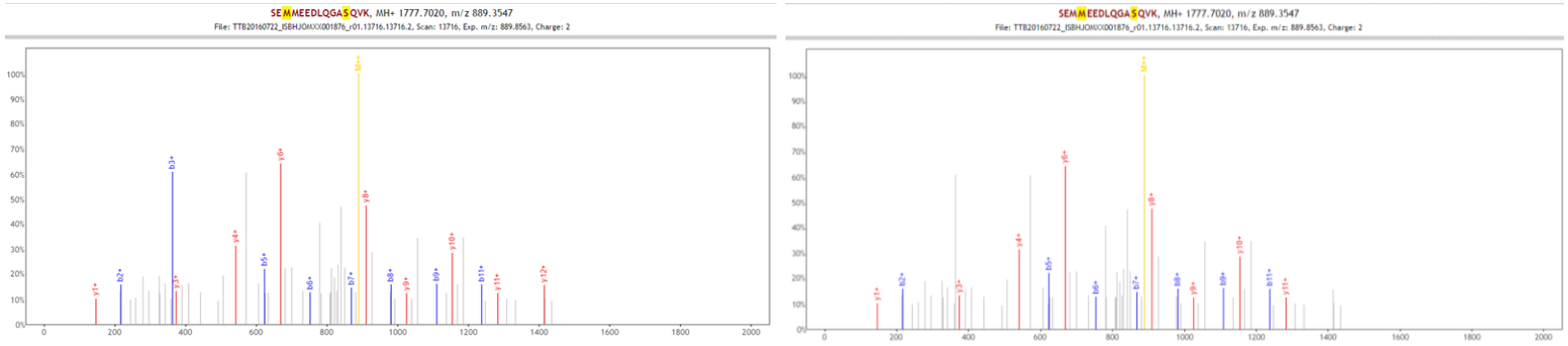
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Simple Peptide: One +16, One +80

SEMMEEDLQGASQVK vs SEMMEEDLQGASQVK



SEMMEEDLQGASQVK vs SEMMEEDLQGASQVK



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PTMProphet Algorithm

1. For each PSM, evaluate all possibilities of peptide modification
2. Sum matched peak intensities: ***Ψ(P)*** for peptide ***P***
3. Compute ***Ψ*** for each peptide possibility
4. For each potential PTM site ***s*** on the peptide, compute:

   $\mathbf{P^{mod}}$ **= argmax(** ∀ **P** *with site* ***s*** *modified* **| Ψ(P))**

   $\mathbf{P^{unmod}}$ = **argmax(** ∀ **P** *with site* ***s*** *unmodified* **| Ψ(P))**

5. Compute "common" matched peak intensity:

   $C(P^{mod},P^{unmod})$

6. Compute discretized observed maximum noncommon intensities:

   $$O^m = \frac{\Psi(P^{mod}) - C(P^{mod},P^{unmod})}{i} \text{ and } O^u = \frac{\Psi(P^{unmod}) - C(P^{mod},P^{unmod})}{i}$$

7. Compute observed maximum noncommon matched peaks:

   M^m and M^u

8. Compute probability for each potential PTM site
9. EM > 0 - apply expectation / maximization algorithm until probabilities remain constant
10. Normalize all probabilities by the number of modifications in the peptide
11. Record the output in pepXML

    S(0.000)EM(1.000)M(0.000)EEDLQGAS(1.000)QVK

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PROBLEM: Site Probabilities May Not Be Comparable

- Different numbers of potentially modified sites in different peptides

SEMMEEDLQGA**S**QVK (2 phospho sites)
SES**S**EEDLQGA**S**QVK (4 phospho sites)

- Different numbers of modifications in different peptides

^{p}**S**E**S**^{p}**S**EEDLQGA^{p}**S**QVK (3 phospho mods)
^{p}**S**ES**S**EEDLQGA**S**QVK (1 phospho mod)

PTMProphet Information Content Stats:

H_t^{norm}: Multiple Modification & Site Normalized Shannon's Entropy

Quantifies the amount of information stored in the PTM site assignment for a peptide with ***s*** modification sites and ***m*** modifications, range: *[0, 1]*

M_t: Localized Modifications Estimate

Estimates the number of modifications confidently localized that can be used to directly compare PSMs containing ***m*** modification, range: *[0, m]*

I_t: Normalized Per-Modification Information Content

Estimates the per-modification localization certainty that can be used to directly compare PSMs with different number of modifications, range: *[0, 1]*

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PTMProphet Site Probabilities Not Directly Comparable

- For example, site probability 0.75 implies different conclusions in different PSMs
  - If s=4 and m=3 and
    S(0.750)ES(0.750)S(0.750)EEDLQGAS(0.750)QVK

    then, $\boldsymbol{H_t^{norm}} = \boldsymbol{1.0}\ (H_t = 3.0\ )$
    $\boldsymbol{I_t} = \boldsymbol{0.0}$
    $\boldsymbol{M_t} = \boldsymbol{0.0}$

    Maximum Uncertainty

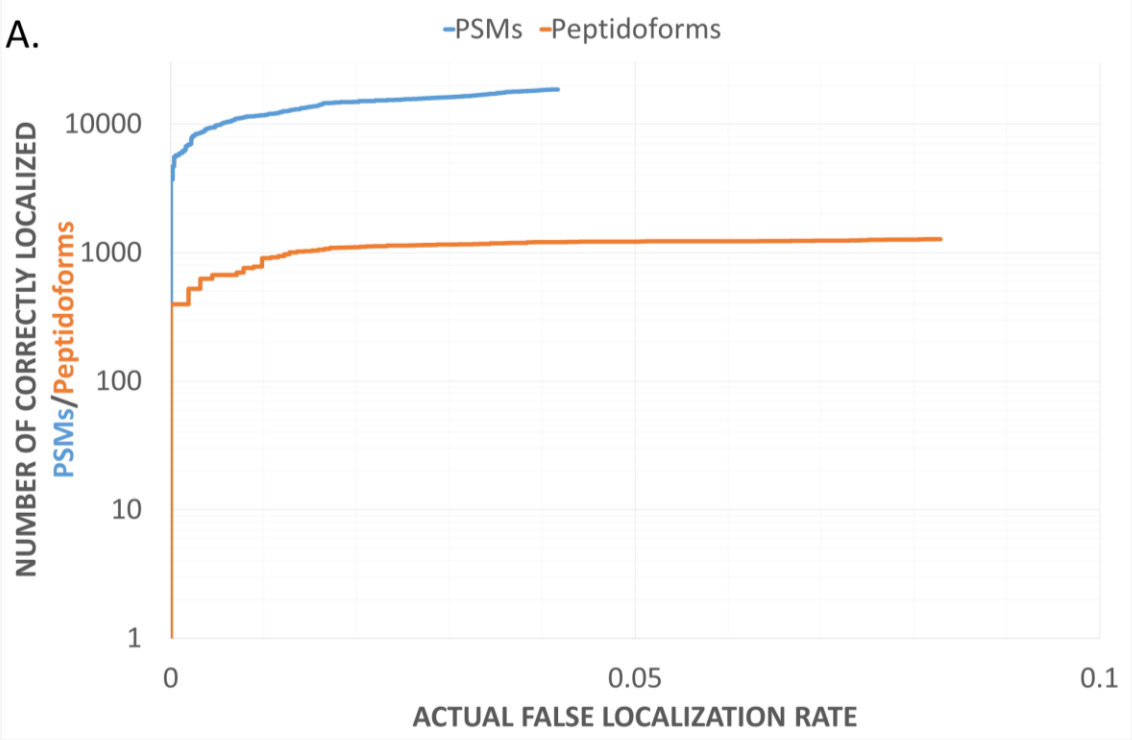
  - If s=4 and m=1 and
    S(0.000)ES(0.750)S(0.250)EEDLQGAS(0.000)QVK

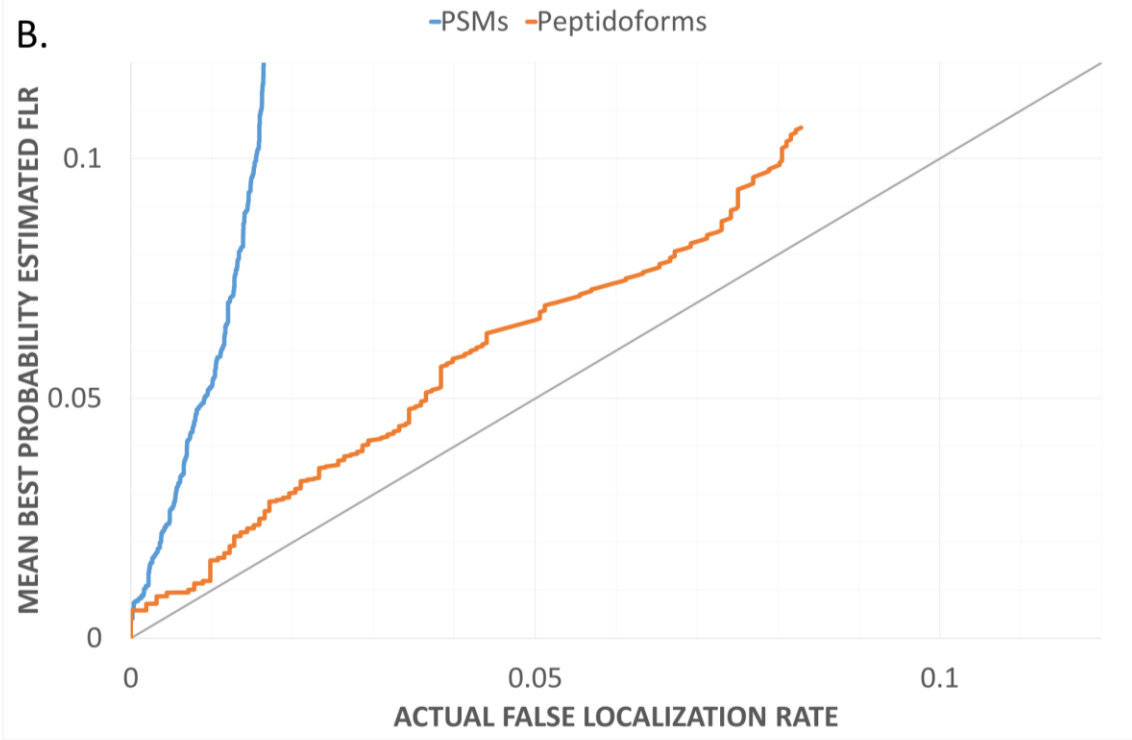
    then, $\boldsymbol{H_t^{norm}} = \boldsymbol{0.41}\ (H_t = 0.41\ )$
    $\boldsymbol{I_t} = \boldsymbol{0.59}$
    $\boldsymbol{M_t} = \boldsymbol{0.59}$

    Mid-range Uncertainty

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Conservative FLR Estimates





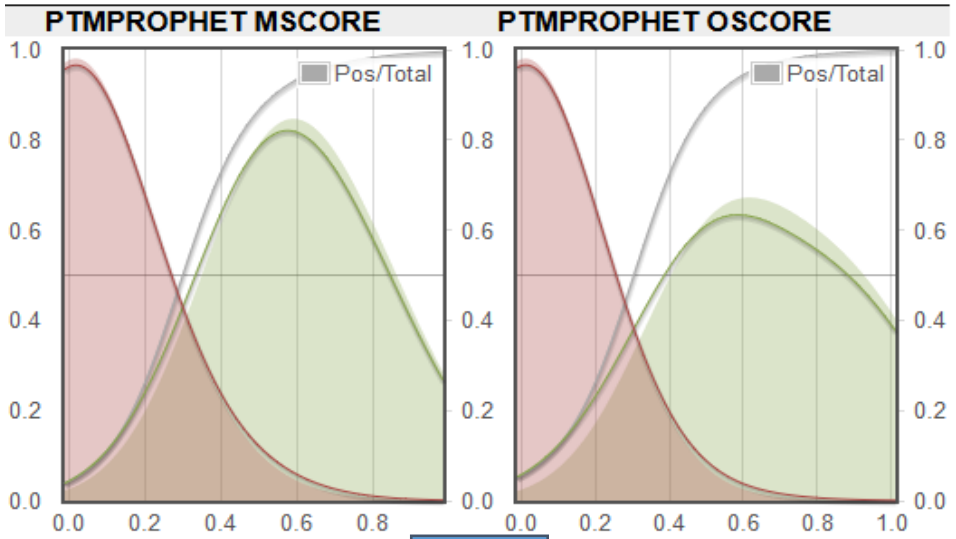
B_t: <u>Mean Best Probability Statistic</u>

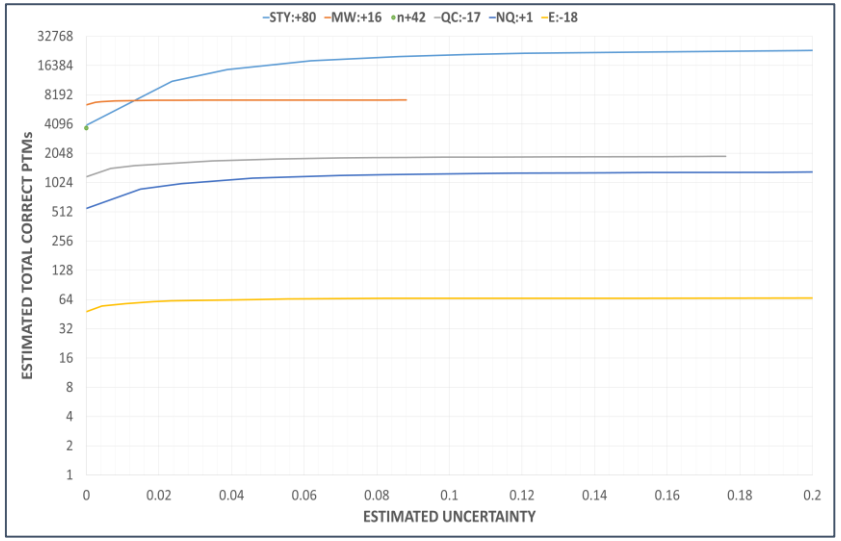
Easy to compute and works well in practice, should be considered in the context of Information Content, range: *[0, 1]*

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

PTMProphet: Wednesday Poster 400

$$Mscore = \frac{M^m}{M^m + M^u} \qquad Oscore = \frac{O^m}{O^m + O^u}$$





$$\frac{Pr(Mscore, Oscore|Modified)Pr(Modified)}{Pr(Mscore, Oscore)}$$

$$H_t^{norm} = -\frac{1}{m}\sum_{i=1}^{s} p_i \log_{s/m} p_i$$

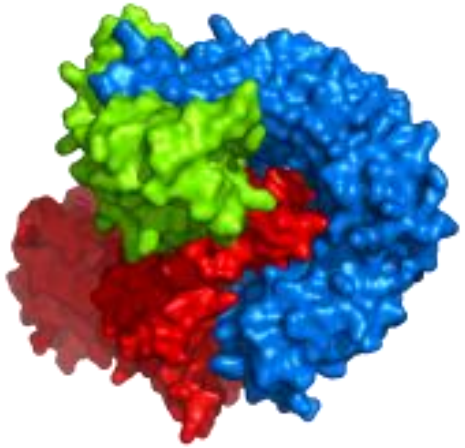
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Agenda

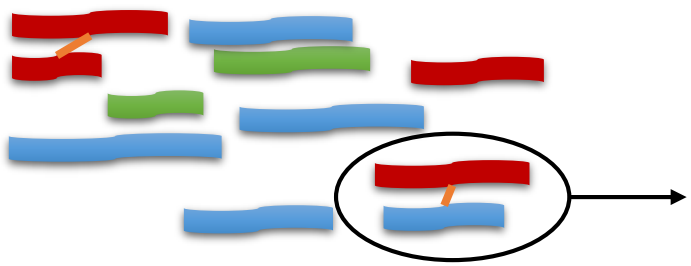
- Introduction to TPP and overview of the latest features in 5.2
- Proteogenomics & PEFF applications
- Analyzing PTMs with PTMProphet
- **Cross-linking analysis with Kojak 2.0**
- Deploying the TPP using Docker containers and cloud computing platforms
- Open Discussion and Feedback

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Kojak: Identifying protein interactions through cross-linking

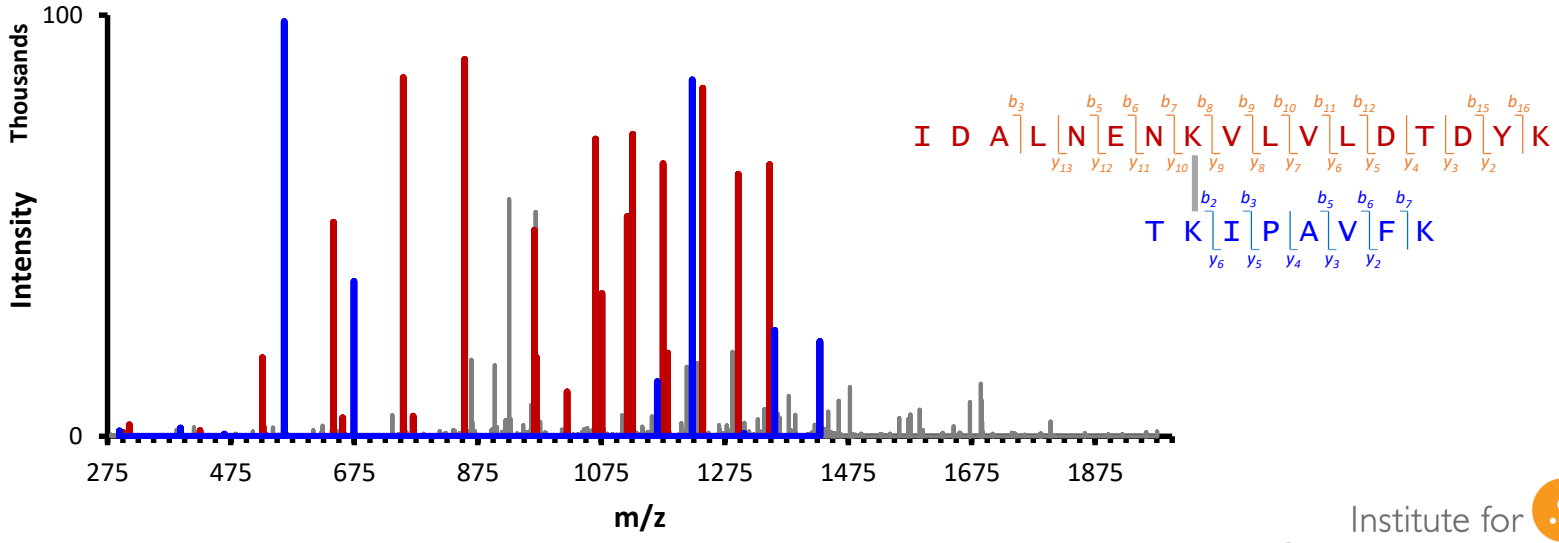


Enzymatic Digestion



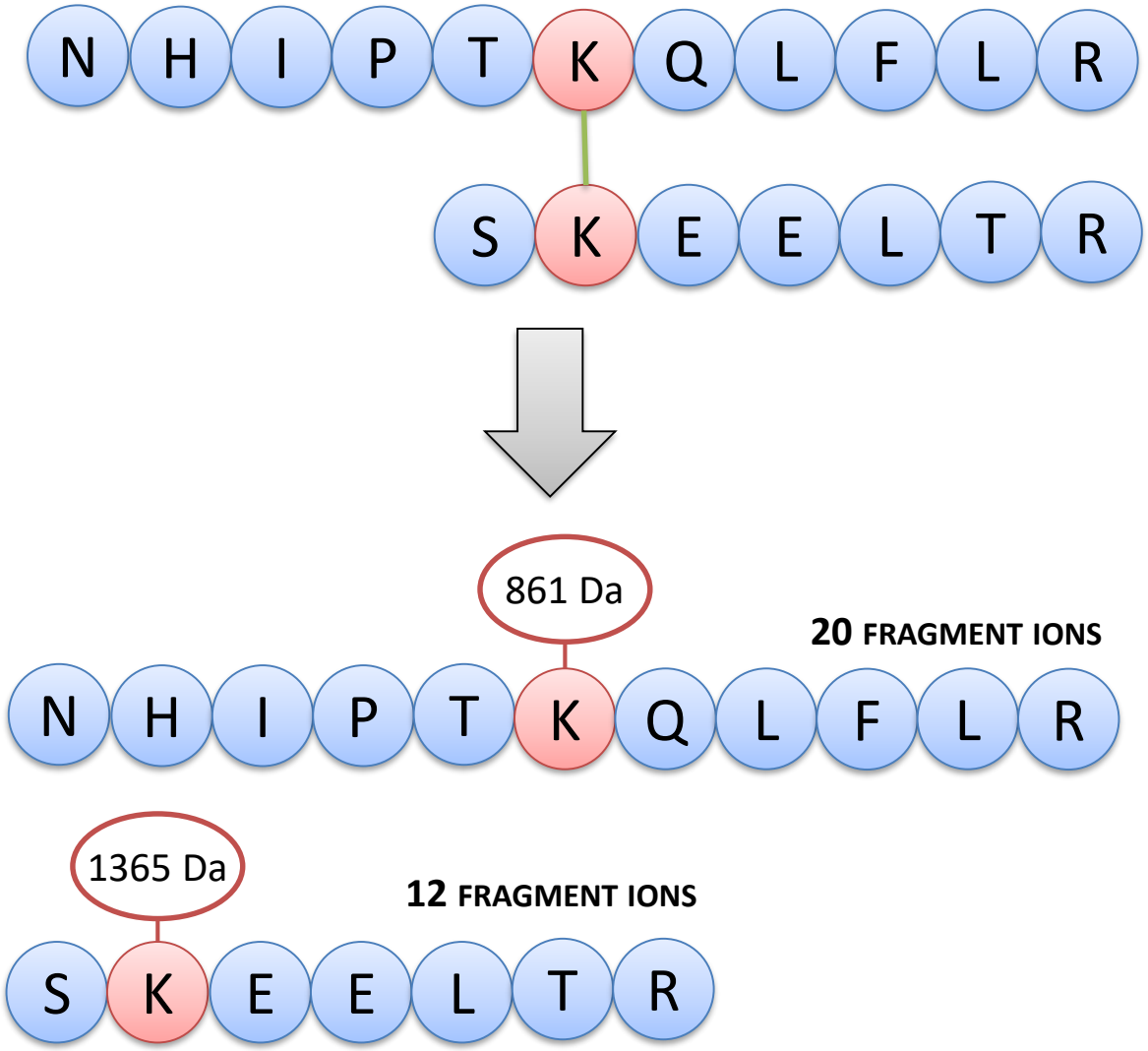
Kojak performs sequence identification of MS/MS spectra from two linked peptides.

- Can be used to infer protein structure.
- Can be used to infer protein-protein interactions.



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

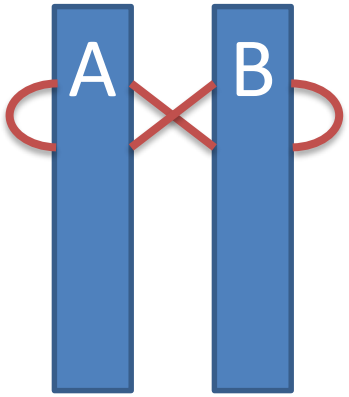
XL Fragmentation Properties



- Consider a cross-linked peptide to be two peptides with modification masses.
- Match peptide fragment ions separately, starting with the largest (by mass) peptide.
- Largest peptide likely has highest potential score.

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Dealing With Homodimers



Intra-protein links:
^{14}N-^{14}N (A1-A2)
^{14}N-^{14}N (B1-B2)

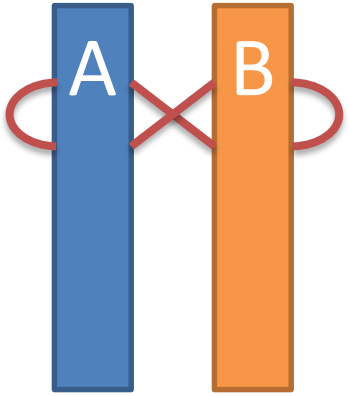
Inter-protein links:
^{14}N-^{14}N (A1-B2)
^{14}N-^{14}N (B1-A2)

New Kojak Parameter:
15N_filter = 15n

>ProteinX
MAAAAVVAATVPAQSMGADGASSVHWFRKGL
RSAVVKVKGSRSGQRLFKEWGVTRDAAIMKM
AKHLLPSLEELGFPTEGLGPAVWQGGE

Total Mass: 729.47 Da

b^+		AA		y^+
88.039	1	S		
159.076	2	A	6	643.450
258.144	3	V	5	572.413
357.213	4	V	4	473.344
485.308	5	K	3	374.276
584.376	6	V	2	246.181
		K	1	147.112



Intra-protein links:
^{14}N-^{14}N (A1-A2)
^{15}N-^{15}N (B1-B2)

Inter-protein links:
^{14}N-^{15}N (A1-B2)
^{15}N-^{14}N (B1-A2)

>15n-ProteinX
MAAAAVVAATVPAQSMGADGASSVHWFRKGL
RSAVVKVKGSRSGQRLFKEWGVTRDAAIMKM
AKHLLPSLEELGFPTEGLGPAVWQGGE

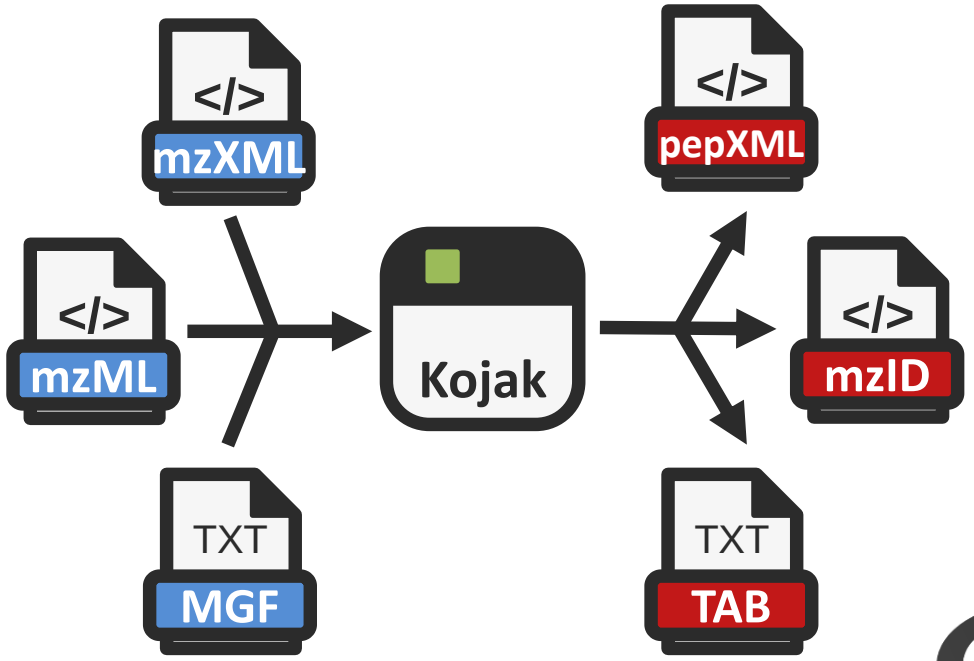
Total Mass: 738.45 Da

b^+		AA		y^+
89.036	1	S		
161.070	2	A	6	651.426
261.135	3	V	5	579.392
361.201	4	V	4	479.326
491.290	5	K	3	379.261
591.355	6	V	2	249.172
		K	1	149.106

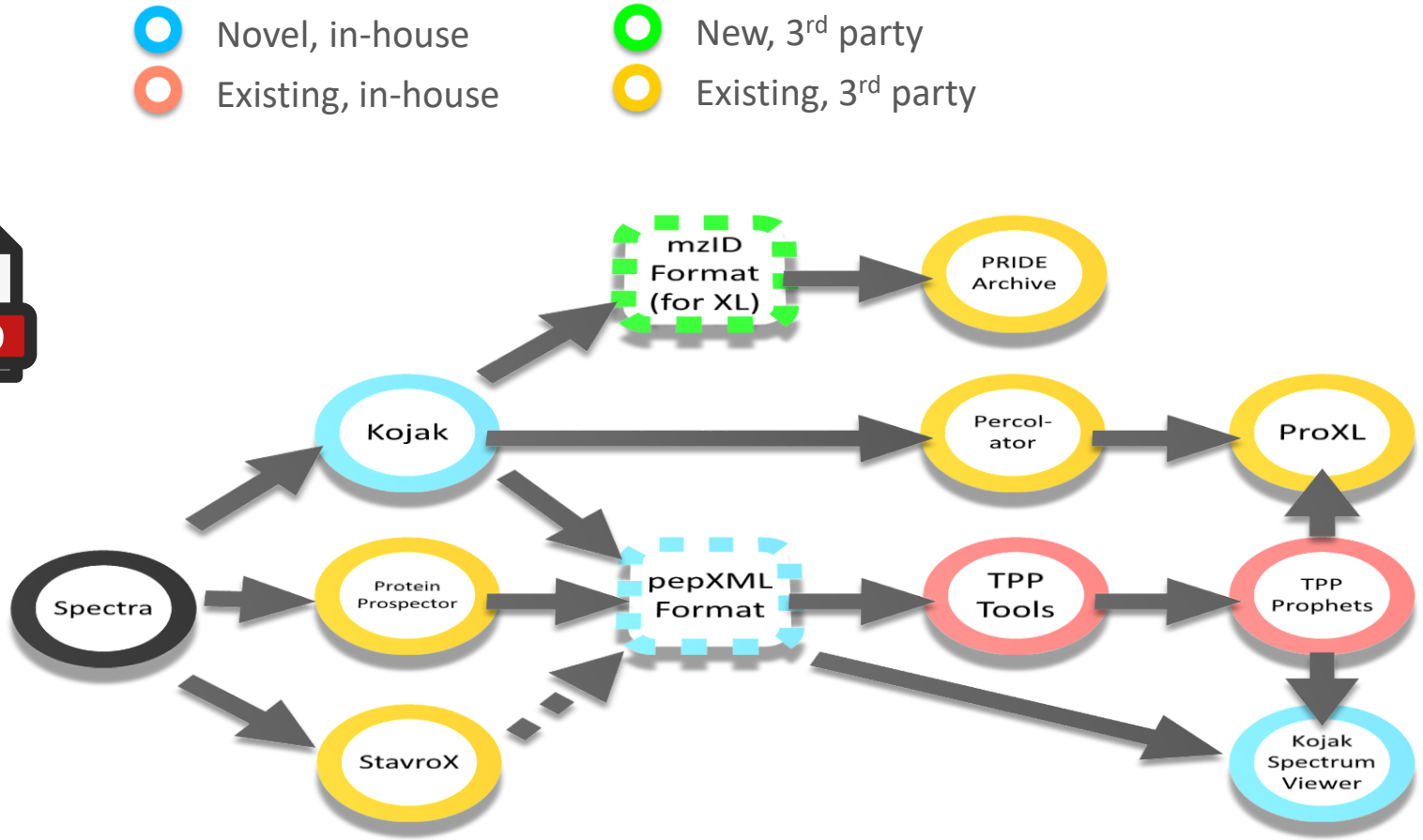
Natural isotope abundance

Heavy isotope labeled

More Supported File Formats & The Extended Kojak 2.0 Pipeline



- No conversion required.
- Uses PSI open standards.

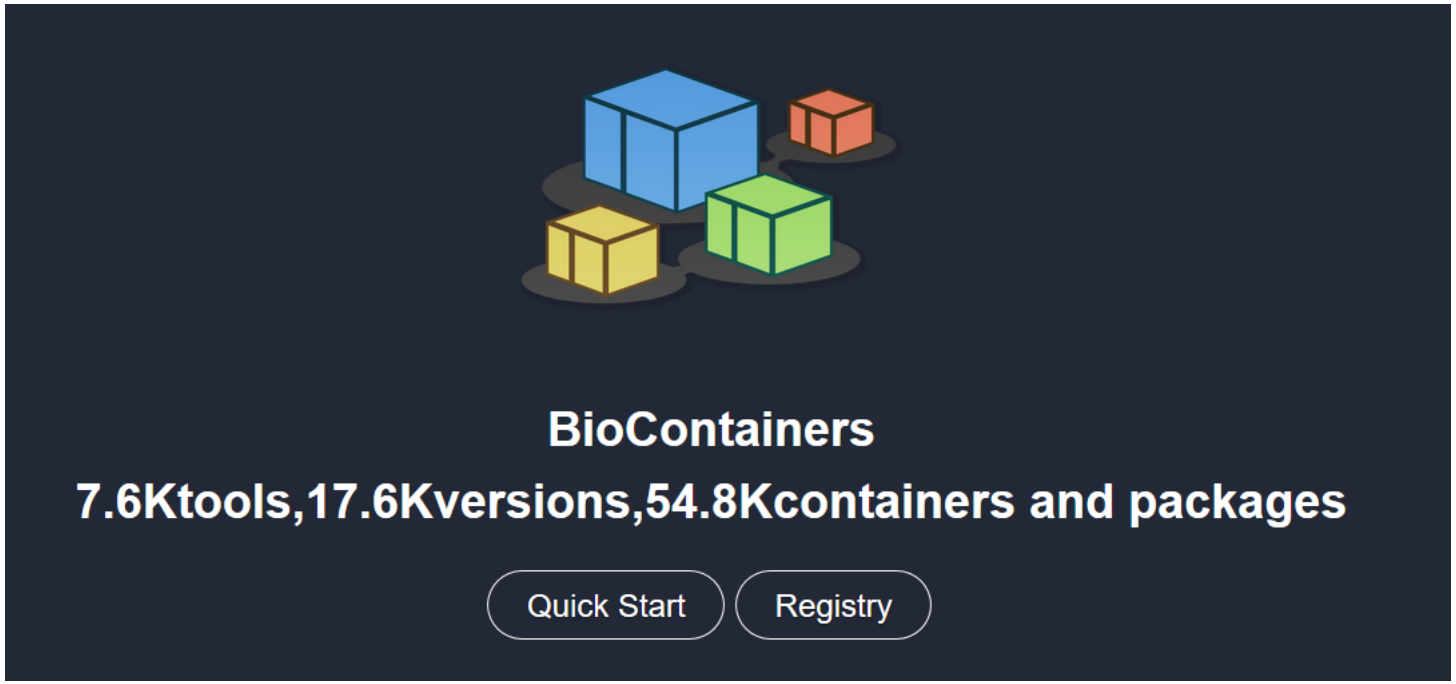


Agenda

- Introduction to TPP and overview of the latest features in 5.2
- Proteogenomics & PEFF applications
- Analyzing PTMs with PTMProphet
- Cross-linking analysis with Kojak 2.0
- **Deploying the TPP using Docker containers and cloud computing platforms**
- Open Discussion and Feedback

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

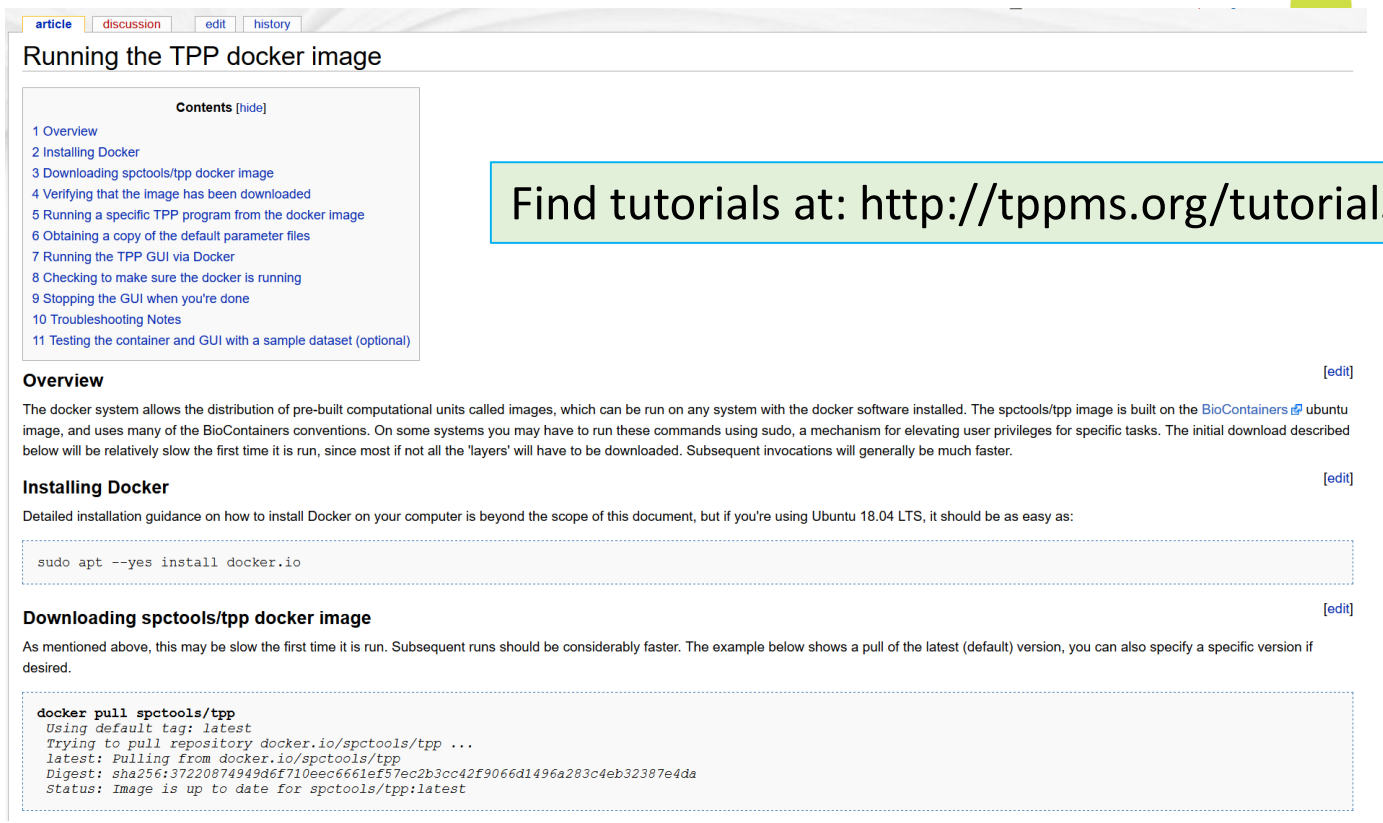
Docker Container Platform



toppred/1.10-4-deb	Add debian package toppred/1.10-4-deb/Dockerfile	a year ago
tpmcalculator/0.0.1	remove CMD	11 days ago
tpp	Updated for TPP 5.2 release version (#321)	a month ago
transdecoder/3.0.1dfsg-1-deb	Add debian package transdecoder/3.0.1dfsg-1-deb/Dockerfile	a year ago
transrate-tools/1.0.0-1b1-deb	Add debian package transrate-tools/1.0.0-1b1-deb/Dockerfile	a year ago
transtermhp/2.09-3b1-deb	Add debian package transtermhp/2.09-3b1-deb/Dockerfile	a year ago
tree-ppuzzle/5.2-8b3-deb	Add debian package tree-ppuzzle/5.2-8b3-deb/Dockerfile	a year ago

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Running TPP on Docker



article | discussion | edit | history

Running the TPP docker image

Contents [hide]

Find tutorials at: http://tppms.org/tutorial

Overview [edit]

The docker system allows the distribution of pre-built computational units called images, which can be run on any system with the docker software installed. The spctools/tpp image is built on the BioContainers ubuntu image, and uses many of the BioContainers conventions. On some systems you may have to run these commands using sudo, a mechanism for elevating user privileges for specific tasks. The initial download described below will be relatively slow the first time it is run, since most if not all the 'layers' will have to be downloaded. Subsequent invocations will generally be much faster.

Installing Docker [edit]

Detailed installation guidance on how to install Docker on your computer is beyond the scope of this document, but if you're using Ubuntu 18.04 LTS, it should be as easy as:

```
sudo apt --yes install docker.io
```

Downloading spctools/tpp docker image [edit]

As mentioned above, this may be slow the first time it is run. Subsequent runs should be considerably faster. The example below shows a pull of the latest (default) version, you can also specify a specific version if desired.

```
docker pull spctools/tpp
 Using default tag: latest
 Trying to pull repository docker.io/spctools/tpp ...
 latest: Pulling from docker.io/spctools/tpp
 Digest: sha256:37220874949d6f710eec6661ef57ec2b3cc42f9066d1496a283c4eb32387e4da
 Status: Image is up to date for spctools/tpp:latest
```

% apt --yes install docker.io

% docker pull spctools/tpp

% docker run -v /tmp/tppdata:/data spctools/tpp comet -p

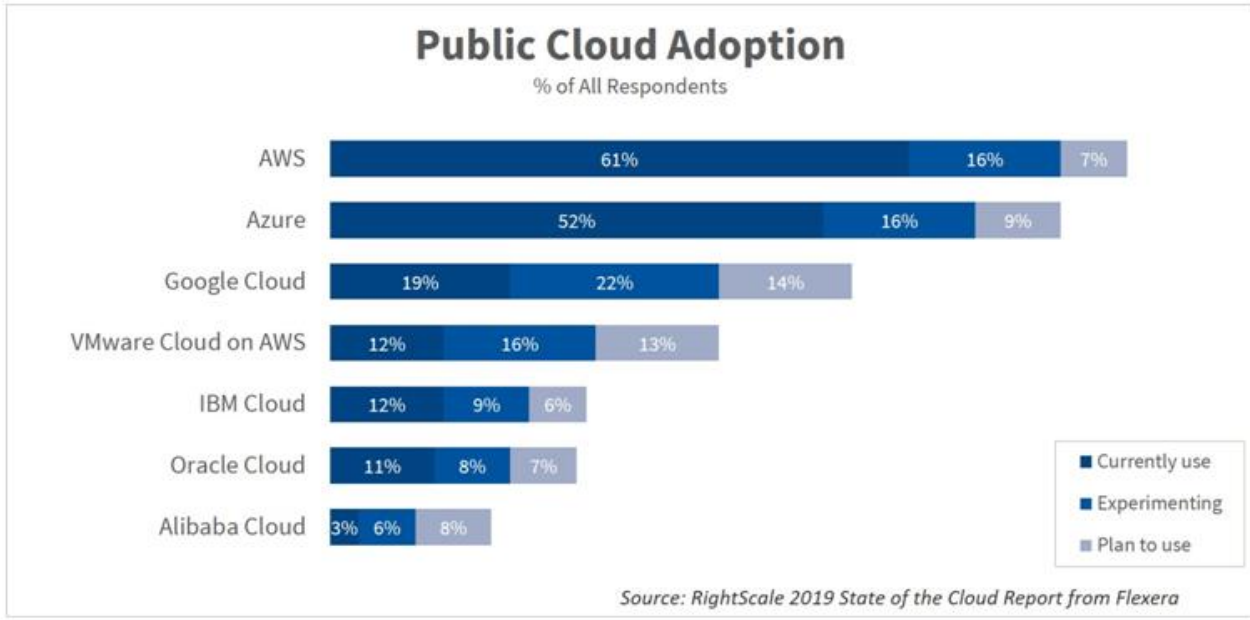
% docker run -dit --user=root -p 10401:10401 -v /tmp/tppdata:/data spctools/tpp apache2ctl -DFOREGROUND

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Cloud Computing Platforms

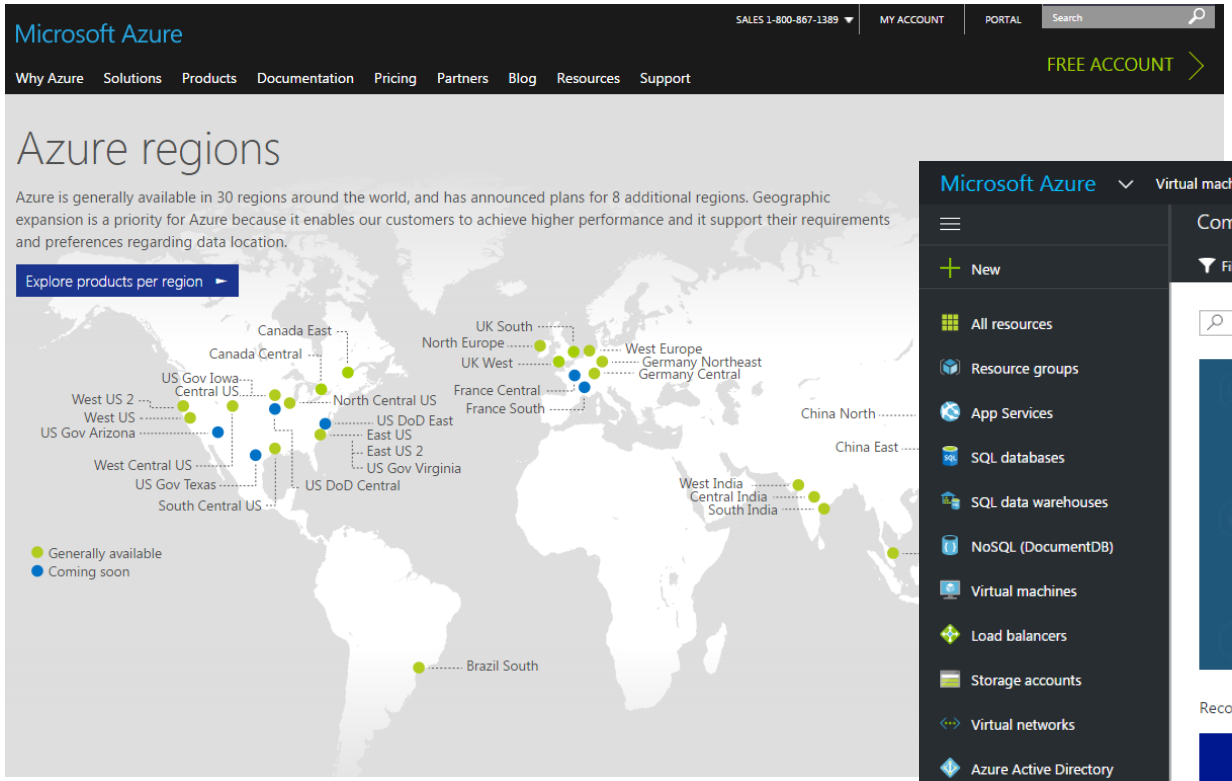
Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

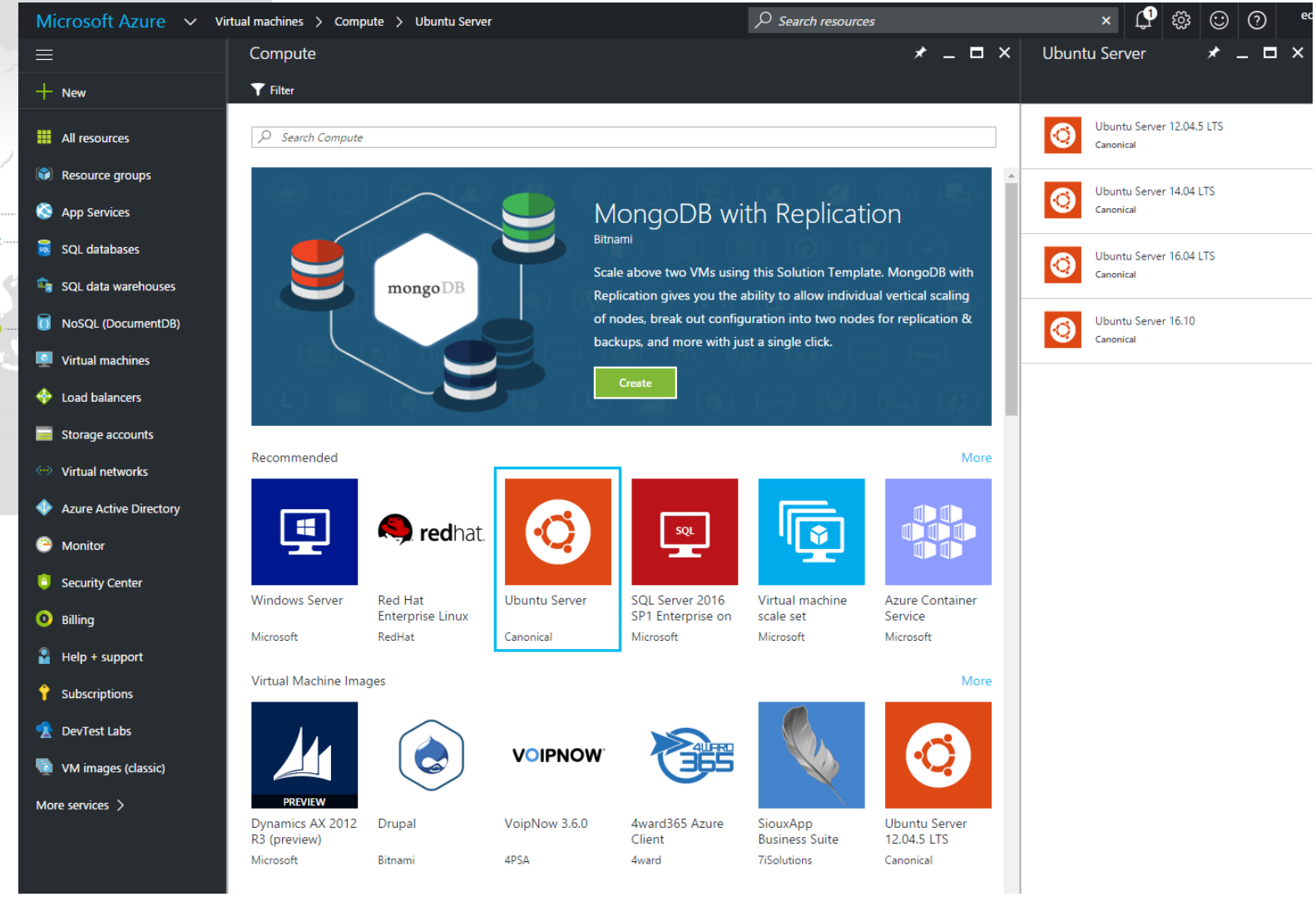
Source: Gartner (May 2018)



Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Microsoft Azure





TPP on the Cloud

Find tutorials at: http://tppms.org/tutorials

article | discussion | edit | history

TPP 5.2.0: Installing on Ubuntu 18.04 LTS

Note that this recipe will likely not be exactly applicable to different versions of TPP or different distributions of Linux, but perhaps can be a useful head start. For other versions of TPP, please see the README and INSTALL_LINUX files found in source code distribution for instructions on how to build TPP for the Linux platform.

Contents [hide]

Before we begin

This recipe describes installing TPP release 5.2.0.

First some general notes

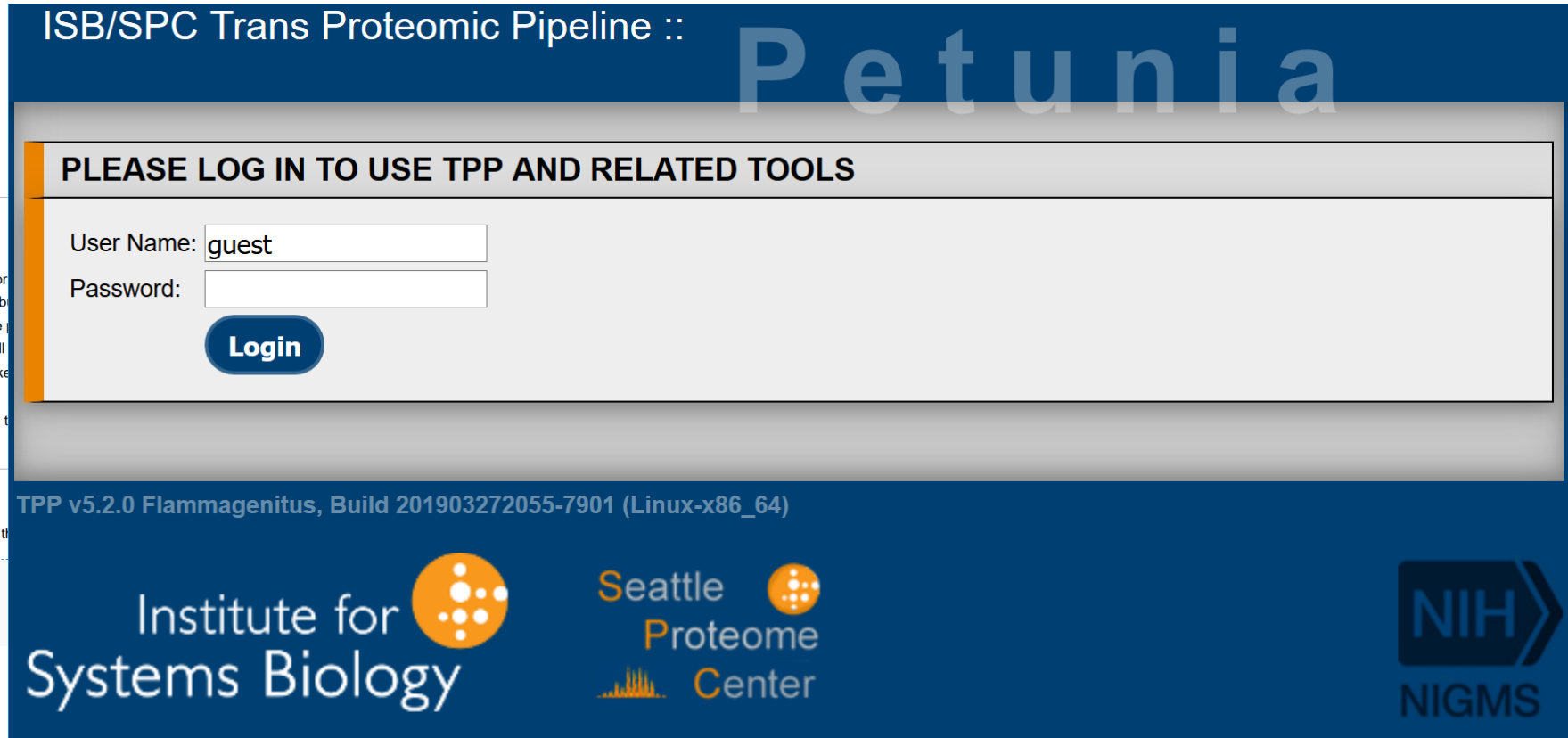
- We assume a fresh installation of Ubuntu 18.04 LTS as a starting point (LTS stands for
- This recipe was tested on a fresh Azure cloud computing platform instance running Ub
- We assume that there is an ordinary user account named 'tpp' that will be used as the
- This installation assumes you are installing to `/usr/local/tpp`. If you want to install
- This installation assumes you will make your data directory `/data`. If you want to make
- This installation assumes you are using Apache 2.4 as your web server
- The gray blocks are commands to be typed in at the shell directly. You should be able t

Installation

Installing prerequisite packages

Use the ubuntu package manager to upgrade any outdated packages and then install all t

```
sudo apt update
sudo apt --yes upgrade
sudo apt --yes install subversion
sudo apt --yes install build-essential
sudo apt --yes install perl
```



TPP on Amazon Web Services

TPP Web Application (TWA)

- Simple web based launcher to start petunia on a Amazon server
- Starts up an pre-configured TPP instance
- Doesn't require any software installation and is inexpensive to run
- Great tool for just trying out TPP
- Can be used when memory and better CPU is needed for an analysis

TPP Amazon command line tools (amztpp)

- Advanced command line toolset
- Launches parallel searches of files across multiple nodes
- Currently supports X!Tandem, OMSSA, MyriMatch, InsPect
- Manage all aspects of cloud computing including data transfer, scheduling, and instances
- Great for quickly and inexpensively processing large amounts of data

Direct Cloud support in TPP's User Interface, Petunia

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Discussion

- Are you currently using Docker to run proteomics tools?
- Are you currently using cloud computing?
- Using the TPP on the cloud?
- Experiences?
- Is there an aspect that we should add?
- Are you writing cloud computing funds into your grant budgets?

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Agenda

- Introduction to TPP and overview of the latest features in 5.2
- Proteogenomics & PEFF applications
- Analyzing PTMs with PTMProphet
- Cross-linking analysis with Kojak 2.0
- Deploying the TPP using Docker containers and cloud computing platforms
- **Open Discussion and Feedback**

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Group presentations at ASMS

For more info:

- Kojak 2.0 (MP 044)
- PEFF in TPP (MP 417)
- PTMProphet (WP 400)

Booth #118

http://www.tppms.org/asms/

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.