Kojak 2.0: New Features For the Analysis of Cross-linked Proteins

Michael R. Hoopmann[1], Alex Zelter[2], Michael Riffle[2], Jimmy K. Eng[2], Trisha N. Davis[2], and Robert L. Moritz[1]
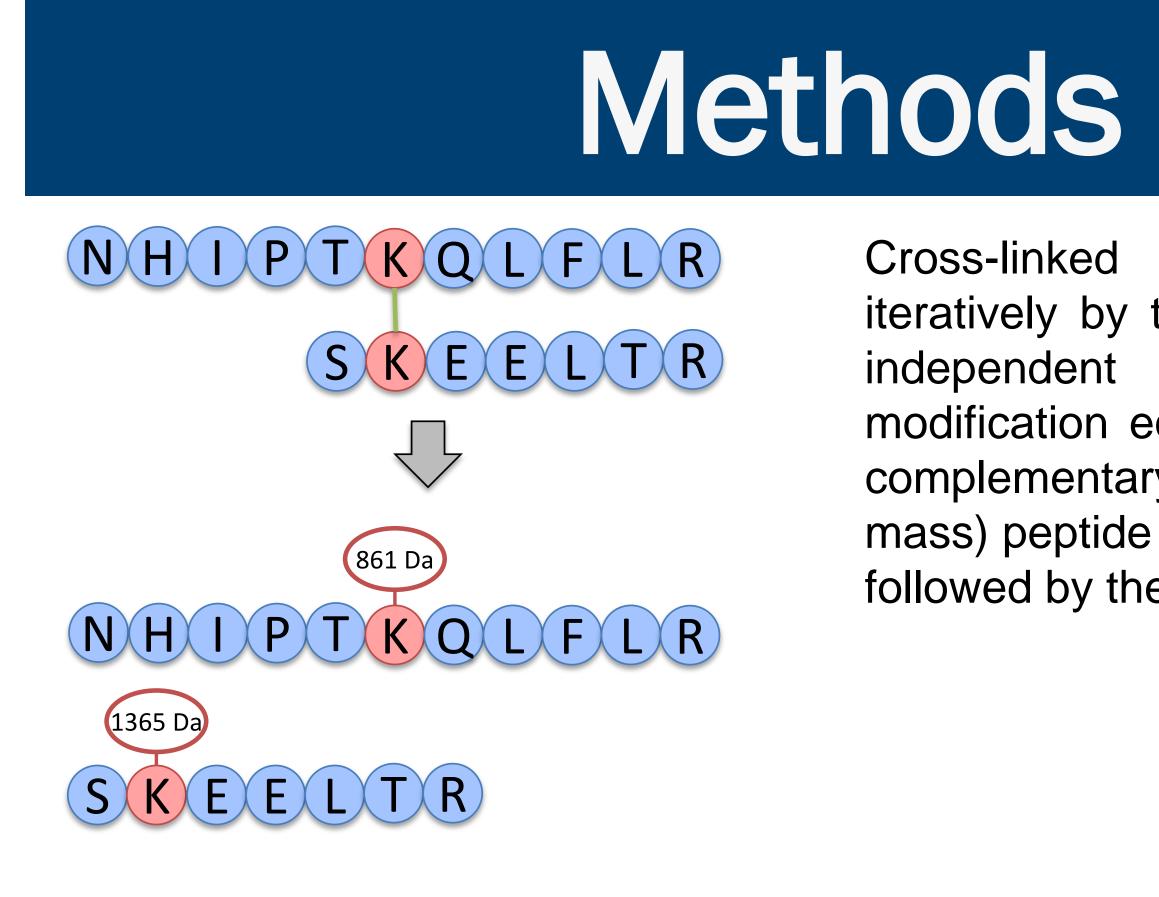
[1]Institute for Systems Biology, Seattle WA
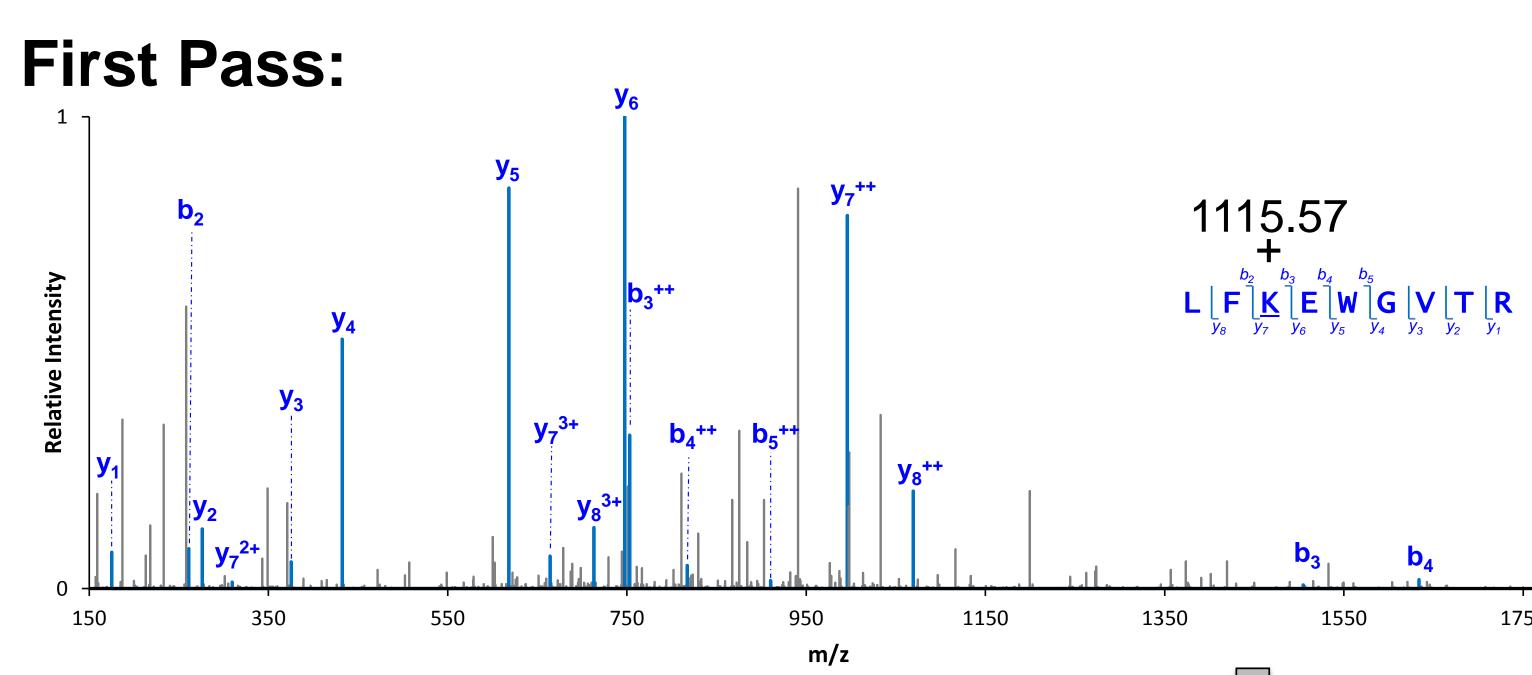[2]University of Washington, Seattle WA

Institute for Systems Biology

Overview

- **Kojak solves cross-linked MS/MS spectra.**

- **^{15}N-labeling resolves homodimer interactions.**

- **New features improve scoring, file support, and compatibility with major pipelines.**

Introduction

Shotgun MS analysis of chemically cross-linked proteins is a versatile tool in proteomics. Data analysis of cross-linked proteins has unique challenges for which specialized algorithms are required. Kojak was initially released in 2015 and designed to perform database searching on MS/MS spectra of cross-linked peptides. Designed to be computationally efficient, Kojak is highly customizable and allows for analysis with many different cross-linkers on both small and large datasets. Its simple interface, combined with adherence to open data standards, enabled Kojak's use with diverse experimental conditions and allowed integration into analytical pipelines. Development of the algorithm continues to build upon these core features. Here we present Kojak version 2.0, a major update to the original Kojak algorithm.
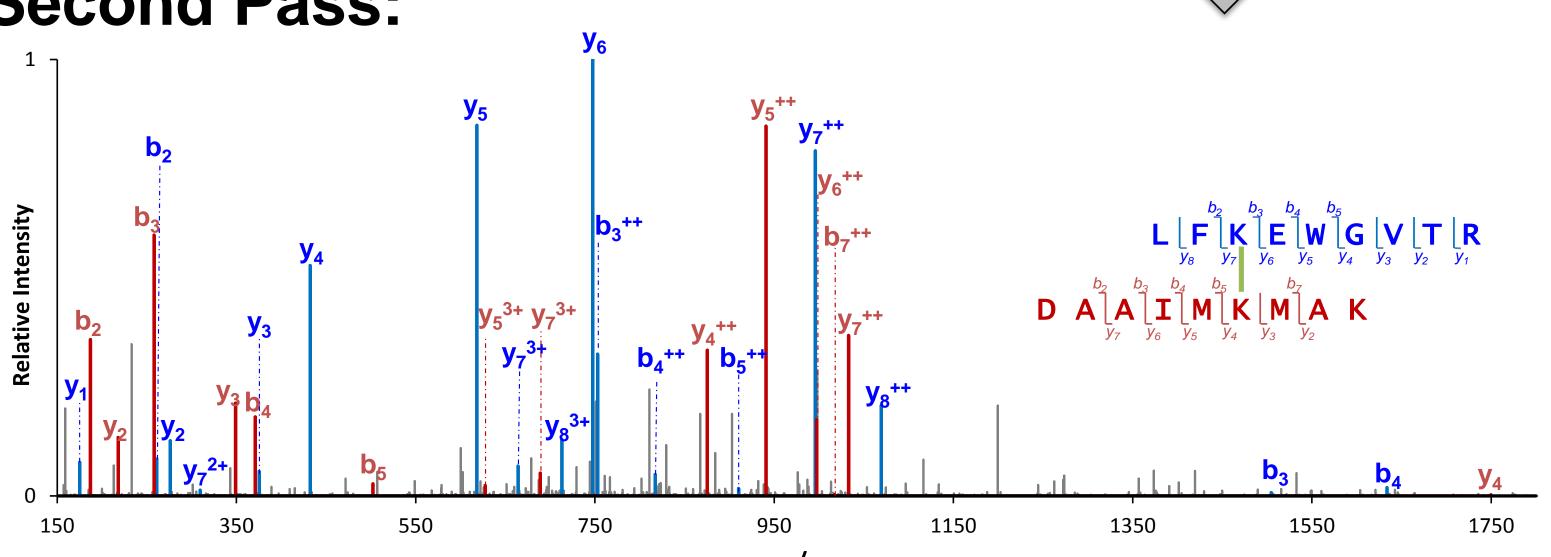
Methods

Cross-linked peptides are solved iteratively by treating the peptides as independent molecules with a modification equal to the mass of the complementary peptide. The larger (by mass) peptide sequence is solved first, followed by the smaller peptide.

First Pass:

In the first pass, a user-defined subset of the best scoring "large" peptides is found. In the second pass, only "small" peptides complementary to these top scorers are considered, for efficiency. Precursor mass is used in both stages to determine candidate peptides and their linked counterpart masses.
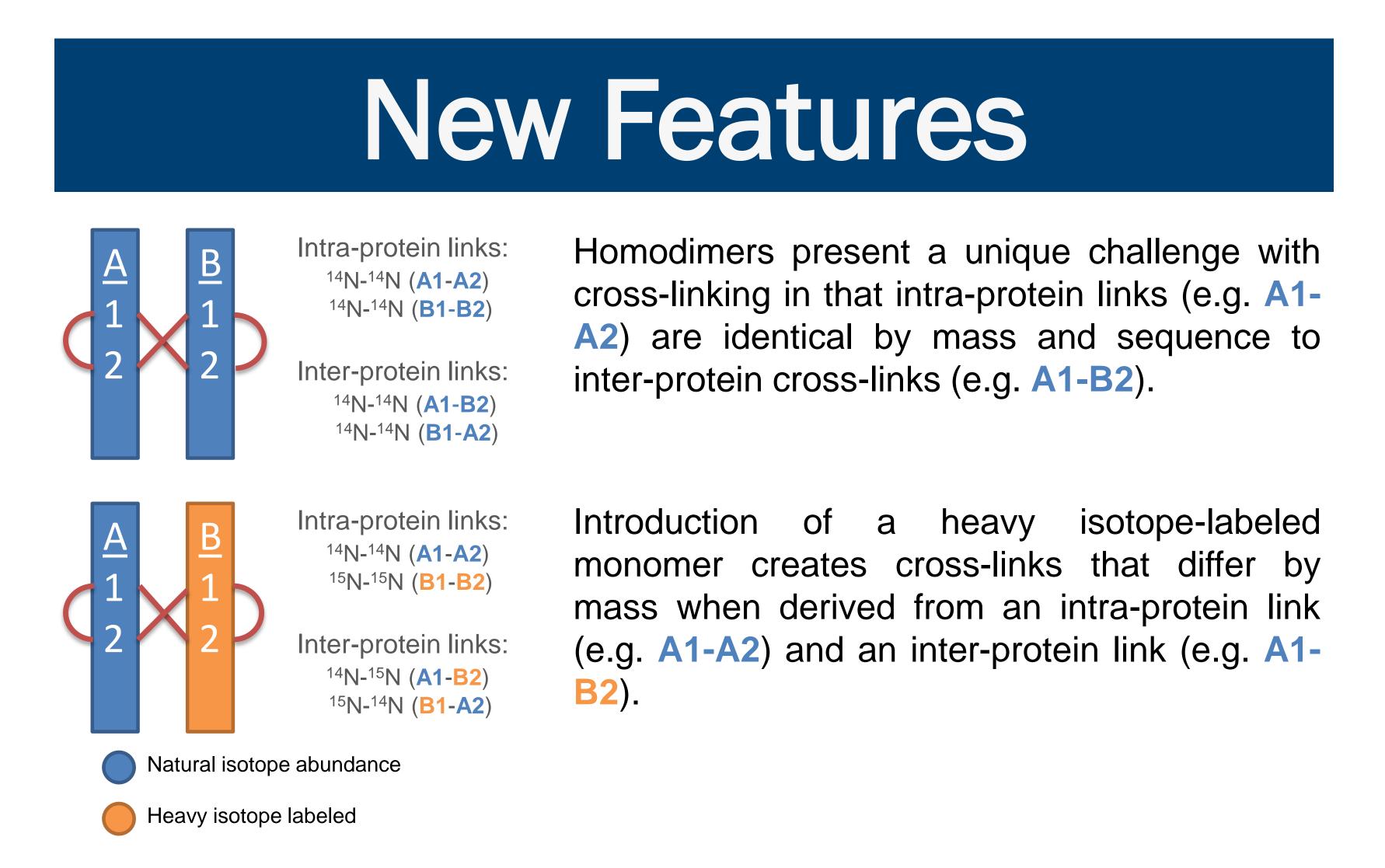
Rank	Score	Peptide
1	2.0994	LFKEWGVTR
2	1.1350	GVKLPNVTEFFR
3	0.9000	HGRGGTKGLGSGIGR
4	0.8600	HGEVILVMEGKR
5	0.8450	DAGTKLMGFSGR

Second Pass:

New Features

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**)

Homodimers present a unique challenge with cross-linking in that intra-protein links (e.g. **A1-A2**) are identical by mass and sequence to inter-protein cross-links (e.g. **A1-B2**).

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**)

Introduction of a heavy isotope-labeled monomer creates cross-links that differ by mass when derived from an intra-protein link (e.g. **A1-A2**) and an inter-protein link (e.g. **A1-B2**).

- Natural isotope abundance
- Heavy isotope labeled

New Kojak Parameter: 15N_filter = **15n**

```
>ProteinX
MAAAAVVAATVPAQSMGADGASSVHWFRKGLR
SAVVKVKGSRSGQRLFKEWGVTRDAAIMKMAK
HLLPSLEELGFPTEGLGPAVWQGGE
```
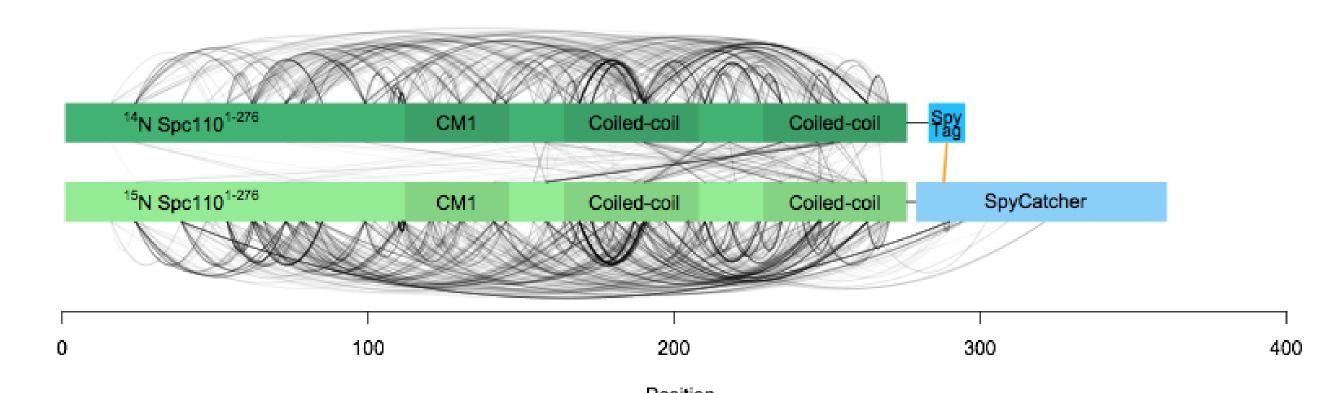
Total Mass: 729.47 Da

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

```
>15n-ProteinX
MAAAAVVAATVPAQSMGADGASSVHWFRKGLR
SAVVKVKGSRSGQRLFKEWGVTRDAAIMKMAK
HLLPSLEELGFPTEGLGPAVWQGGE
```

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

A user-defined label in the FASTA sequence identifier indicates computing ^{15}N peptide and fragment ion masses (above). Inter-protein links are then distinguishable from intra-protein links (below).
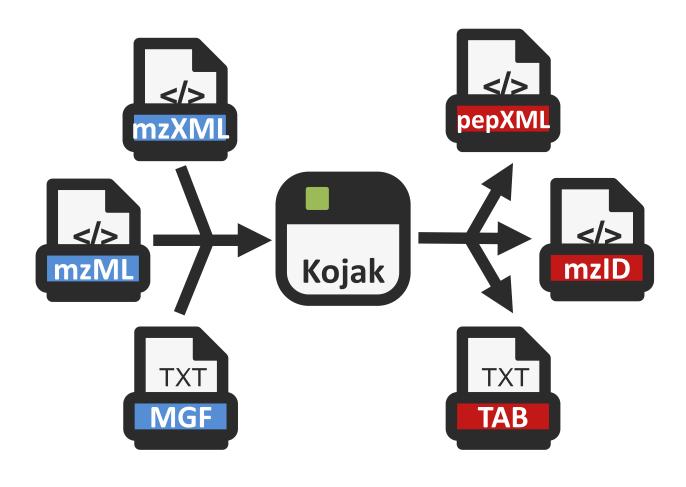
New Features

Peptides	Kojak Score	E-value
	3.150	6.32e-07
	1.225	3.27e-03
Total:	4.375	2.09e-09

Because Kojak Score (based on Xcorr[1]) is correlated with peptide length, expectation values are computed from the linear least squares regression on the log transform of the cumulative distribution function of the score histogram. Since the peptides are searched separately, each will have its own score histogram (and context), as well as the final score when compared to all searched sequences including non-linked peptides. On-the-fly decoy searches are used to supplement the score histograms when below a count threshold. Each score should be used when evaluating the cross-linked result, with emphasis on the lowest scoring peptide.[2]

Expanded File & Pipeline Support

Kojak supports additional open file standards for both input and output. This facilitates integration into most MS pipelines, including the Trans-Proteomic Pipeline[3] and ProXL.[4]

References:
1 Eng et al., *Proteomics*, **2013**, 13(1):22-4.
2 Trnka et al., *Mol Cell Proteomics*, **2014**, 13(2):420-34.
3 Deutsch et al., *Proteomics Clin Appl*, **2015**, 9(7-8):745-54.
4 Riffle et al., *J Proteome Res*, **2016**, 15(8):2863-70.

Support & Information

http://kojak-ms.org