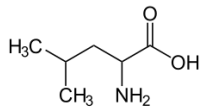
Valens™ with CLIP™

Confirmation of Leucine or Isoleucine Presence (CLIP™)

First published June 19, 2019

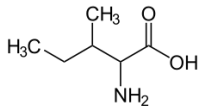
Low parts-per-million accuracy and faster mass spectrometers (MS) have made *de novo* sequencing of proteins practical in recent years. Antibodies are an important class of molecules that uniquely require *de novo* sequencing. Used in therapeutics, diagnostics, and research applications, monoclonal antibodies are biologics produced in cell expression systems. Over time, cell expression systems degrade or are lost, and there is a need to recover the antibody sequence from the protein. Given a correct antibody sequence and good expression, newly generated antibody material would have the same binding and specificity properties as the original antibody. Additionally, good reagent antibodies is a keystone for ensuring reproducible science. Immunoassays such as ELISAs are dependent on high-quality antibodies to detect analytes. Batch-to-batch variation in antibody reagents is a common source of inconsistent binding and confounding immunoassay results [1]. Knowing the antibody sequence is an important part of characterizing antibody reagents. To ensure accurate antibody characterization, *de novo* protein sequencing must be highly accurate with unambiguous assignment of all amino acid residues.

Leucine



(CH3)2CH-CH2-CH(NH2)-COOH

Isoleucine



C_2H_5-CH(CH_3)-CH(NH_2)-COOH

Average Mass: 113.2 Da

Monoisotopic Mass: 113.08406 Da

Distinguishing between isoleucine (Ile) and leucine (Leu) residues is troublesome in MS-based *de novo* sequencing as the side chains are constitutional isomers and have the same mass, 113.08406 Da. Two peptides differing in amino acid composition by a substitution of Ile with Leu have identical masses. Additionally, textbook proteomics CID or HCD fragmentation of the peptide bonds of the two peptides would yield the same peptide fragment masses. Due to the lack of discriminatory signal, many *de novo* sequencing tools report isoleucines or leucines in a peptide as an ambiguous Xle residue. However, Ile and Leu in antibodies can be distinguished using alternate proteomics strategies, specifically:

1. Enzyme specificity to leucine cleavage [2]
2. EThcD generates isoleucine and leucine diagnostic *w*-ions [3, 4]
3. V/J gene homology validation.

Enzyme specificity. MS/MS proteomics workflows use enzymes or chemicals to cleave proteins into smaller peptides since MS/MS instruments produce more interpretable fragmentation data from peptides than proteins. Each digestion enzyme has a propensity to cleave at specific residues, and the choice of enzyme to use depends on the application. In Valens ™, a sample comprised of a single antibody is digested by multiple enzymes to ensure unique peptides covering the entire antibody are generated. Of these enzymes, chymotrypsin tends to cleave after tyrosine (Tyr), tryptophan (Trp), leucine (Leu), and phenylalanine (Phe), while pepsin tends to be less specific and cleaves before or after glutamic acid (Glu), tyrosine (Tyr), leucine (Leu), phenylalanine (Phe), and alanine (Ala) [2]. Both enzymes are more prone to cleave around leucine than

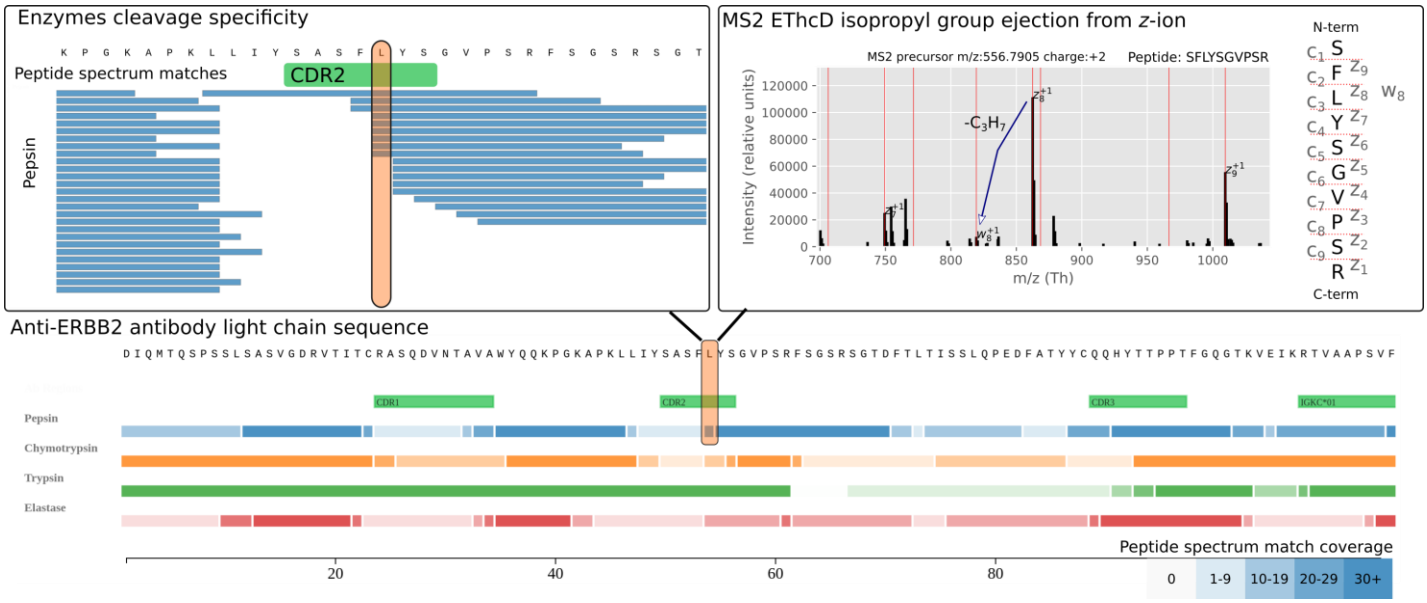


Figure 1: Evidence for Leu at light chain for CDR2 of anti-ERBB2 antibody. Pepsin shows propensity to cleave at the site, and the Leu is also supported by *w*-ion evidence. The leucine is also supported by the nucleotide codon from corresponding IgK V gene sequence.

isoleucine (see Figure 2). The lack of peptides terminating at an unknown Xle site would be evidence suggesting the Xle is an Ile, whereas observing peptides terminating at the Xle site would suggest the site is a Leu.

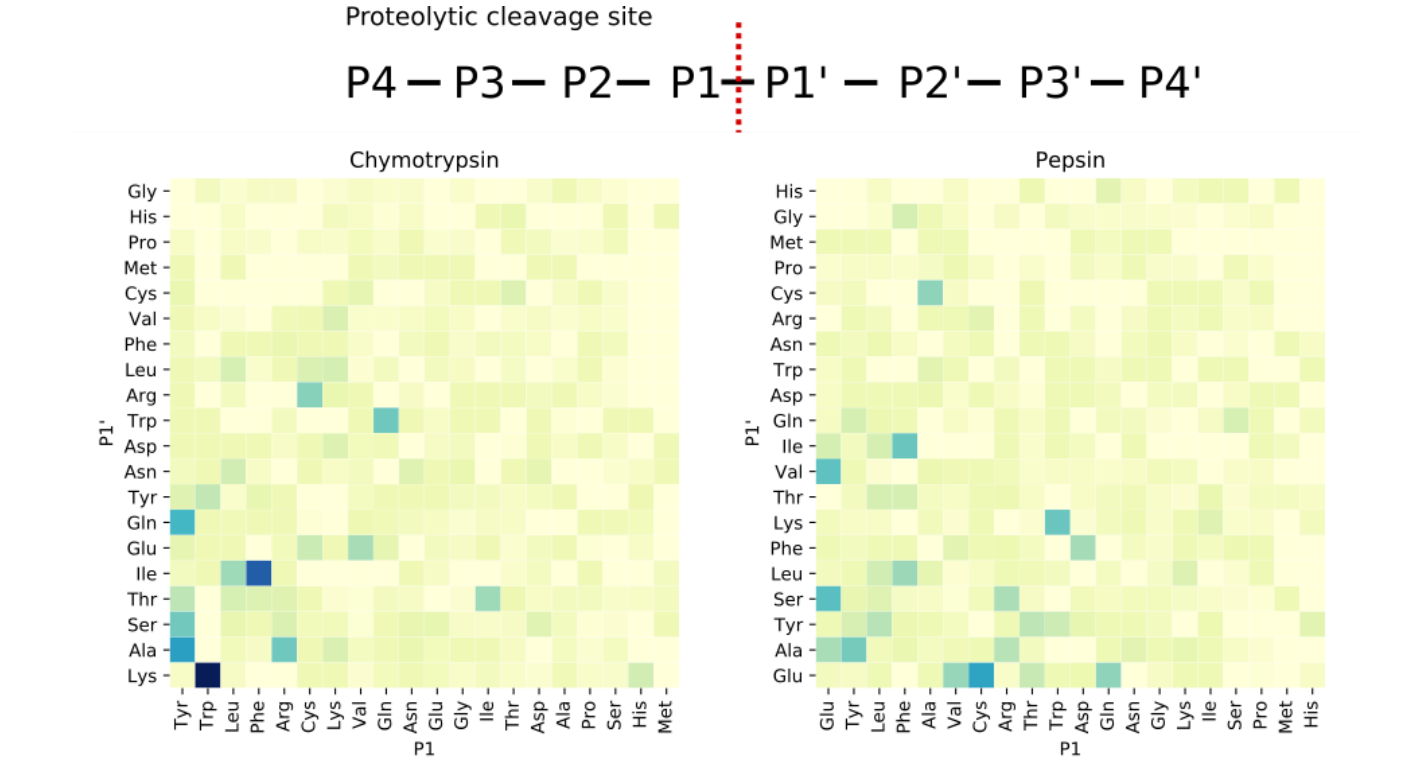


Figure 2: Chymotrypsin and pepsin enzyme cleavage specificity normalized to possible antibody cleavages. Using Schechter and Berger nomenclature, P1 is the residue position on the protein substrate prior to a cleavage site and $P1^0$ is the residue position after the cleavage site. Residues are ordered such that the most frequent P1 cleavage is on the left and most frequent $P1^0$ cleavage is on the bottom.

EThcD fragmentation of Ile and Leu side chains. While fragmentation typically occurs along the peptide backbone, fragmentation of residue side chains is also possible. Ile and Leu residues share the same elemental composition but have different bonding patterns (see above figure). Fragmentation of Ile and Leu side chains result in ions with distinct mass. Xiao et al. [3] as well as others showed side-chain fragmentation can be produced using an MS^n strategy with ETD fragmentation for MS2 fragmentation of peptides and HCD fragmentation to generate

Enzyme	anti-*β*Gal HC coverage	anti-*β*Gal HC spectrum depth	anti-ERBB2 HC coverage	anti-ERBB2 HC spectrum
trypsin	0.968	50.7	0.949	39.8
chymotrypsin	0.736	15.2	0.869	36.0
elastase	0.560	15.7	0.659	17.8
pepsin	0.738	8.5	0.800	15.3

MS3 for targeted *z*-ions. The *z*-ions where the unknown Xle is at $N - C\alpha$ bond breakage results in Ile and Leu characteristic *w*-ions. As shown in Figure 1, Leu side chain fragmentation results in an ejection of an isopropyl group (loss of C_3H_7 from the *z*-ion), whereas the Ile fragmentation results in an ejection of an ethyl group (loss of C_2H_5 from the *z*-ion). Additionally, Zhokhov et al. [4] demonstrated EThcD fragmentation of short peptides can produce characteristic *w*-ions using MS2 alone. In their analysis, only peptides composed of less than eight amino acids produced *w*-ions.

V and J reference gene validation. Lastly, most antibody sequences are derived naturally and are likely to use the same nucleotide sequence as the original V and J gene sequences. Predicted Ile and Leu can be confirmed by checking the encoding codon for antibody labeled V and J gene segments. Of note, antibodies contain complementarity-determining regions (CDRs) that are highly prone to mutation. Ile and Leu validation by checking the translated nucleotides from genes for CDR1 and CDR2 are suspect. CDR3s of antibodies are uniquely derived with no corresponding gene segments to validate against.

Dataset

To illustrate Confirmation of Leu and Ile Presence (CLIP ™), MS/MS data was generated from two monoclonal antibodies, anti-ERBB2 (Absolute Antibody Ab00103) and anti-*β*Gal (Abterra Bio) antibody. Each antibody heavy and light chain was digested by four enzymes: chymotrypsin, trypsin, pepsin, and elastase. Digested peptides were analyzed by a ThermoFisher ™ Orbitrap Fusion ™ Lumos ™ Tribrid ™ in doublet HCD and EThcD mode. *De novo* protein sequencing was performed using Abterra Biosciences' Valens™ algorithm, and CLIP™ was used to call Ile and Leu residues.

Results

Different enzymes had different efficiency for generating peptides, but the aggregate of all enzymes resulted in complete sequence coverage. For the heavy chain (HC) of both antibodies, aggregating across all enzymes was required for 100% sequence coverage (see Table 1). Depth of coverage is defined as the average number of amino acid residues from peptide spectrum matches covering a sequence. For the anti-*β*Gal heavy chain, the average depth of coverage across all enzymes was 90.1 and was even higher for the anti-ERBB2 heavy chain at 108.8. A combination of good depth of coverage and fraction of monoclonal antibody with coverage is necessary for reliable *de novo* sequencing and identifying Xle sites.

Table 1: Peptide spectrum match coverage of heavy chains from anti-*β*Gal and anti-ERBB2 monoclonal antibodies. Fraction of heavy chain sequence covered is 56-97% depending on the enzyme. Anti-ERBB2 coverage was higher than anti-*β*Gal.

CLIP™ applies Bayesian inference that incorporates enzyme specificity evidence and characteristic *z* and *w*-ion evidence to produce a log-likelihood score for an unknown Xle site being a leucine or isoleucine. Of the 62 sites on the anti-ERBB2 antibody, the accuracy of correctly calling isoleucine or leucine at a log-likelihood threshold of 1.0 was 89%. Only, 9 sites did not have a log-likelihood score surpassing the threshold, 86% of sequences were called. The anti-*β*Gal antibody had 67 sites, with similar 83% accuracy and 88% call rate at the same threshold.

Prediction errors are typically the result of isoleucines and leucines appearing too close together in sequence. Peptides containing both isoleucines and leucines can produce decoy *w*-ions through radical site migration and cause erroneous calls [4]. Fortunately, isoleucines and leucines are not commonly found adjacent in CDRs.

Complete and confident CLIP™ is achieved by validating predictions with V and J gene sequences. If a questionable prediction is made in a CDR, the traditional approach of targeted MS/MS runs of antibody peptide digests using MS2 EThcD and MS3 HCD can be used for Ile and Leu determination.

Conclusion

Although isoleucine and leucine residues share the same mass, distinguishing the residues accurately and efficiently is possible from *de novo* protein sequencing using mass spectrometry.

References

[1] M. Baker. Reproducibility crisis: Blame it on the antibodies. *Nature*, 521(7552):274–276, May 2015.

[2] B. Keil. *Specificity of proteolysis*. Springer Science & Business Media, 2012.

[3] Y. Xiao, M. M. Vecchi, and D. Wen. Distinguishing between Leucine and Isoleucine by Integrated LC-MS Analysis Using an Orbitrap Fusion Mass Spectrometer. *Anal. Chem.*, 88(21):10757–10766, 11 2016.

[4] S. S. Zhokhov, S. V. Kovalyov, T. Y. Samgina, and A. T. Lebedev. An EThcD-Based Method for Discrimination of Leucine and Isoleucine Residues in Tryptic Peptides. *J. Am. Soc. Mass Spectrom.*, 28(8):1600–1611, 08 2017.