HHS Public Access

Author manuscript

J Am Soc Mass Spectrom. Author manuscript; available in PMC 2018 May 01.

Published in final edited form as:

J Am Soc Mass Spectrom. 2017 May ; 28(5): 803–810. doi:10.1007/s13361-016-1580-0.

Automated Antibody *De Novo* Sequencing and Its Utility in Biopharmaceutical Discovery

K. Ilker Sen[1,*], **Wilfred Tang**[1], **Shruti Nayak**[2], **Yong Kil**[1], **Marshall Bern**[1], **Berk Ozoglu**[3], **Beatrix Ueberheide**[2], **Darryl Davis**[3], and **Chris Becker**[1]

[1]Protein Metrics Inc. 1622 San Carlos Ave, Suite C, San Carlos, CA 94070, USA

[2]New York University, Langone Medical Center, 430 East 29th street, 8th floor room 860, New York, NY 10016, USA

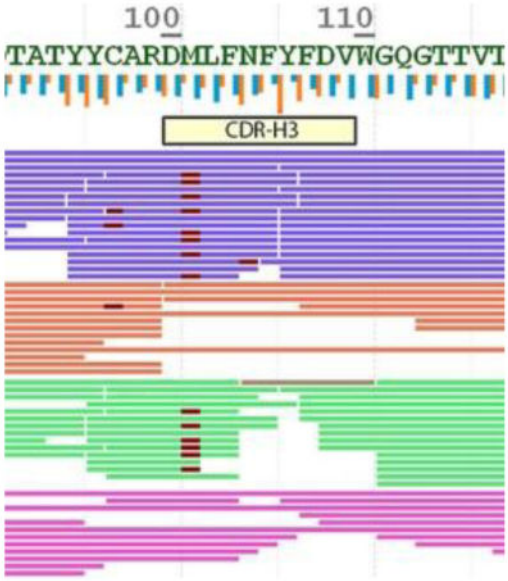
[3]Janssen Research & Development, LLC, 1400 McKean Road, Spring House, PA 19477, USA

Abstract

Applications of antibody *de novo* sequencing in the biopharmaceutical industry range from the discovery of new antibody drug candidates, to identifying reagents for research, and determining the primary structure of innovator products for biosimilar development. When murine, phage display, or patient derived monoclonal antibodies against a target of interest is available, but the cDNA or the original cell line is not, *de novo* protein sequencing is required to humanize and recombinantly express these antibodies, followed by *in vitro* and *in vivo* testing for functional validation. Availability of fully automated software tools for monoclonal antibody *de novo* sequencing enables efficient and routine analysis. Here, we present a novel method to automatically *de novo* sequence antibodies using mass spectrometry and the Supernovo™ software. The robustness of the algorithm is demonstrated through a series of stress tests.

Graphical abstract

Address reprint requests to: K. Ilker Sen, 1622 San Carlos Ave, Suite C, San Carlos, CA 94070, USA. +1 (860) 615-9871. sen@proteinmetrics.com.



Introduction

Edman degradation[1] and a combination of Edman degradation with mass spectrometry [2] have been used successfully in the past to determine complete protein sequences [3, 4]. Today, there are a number of algorithms and software packages available for purely mass spectrometry based *de novo* sequencing at the peptide level, such as PepNovo [5] pNovo [6], PEAKS [7], and Novor [8]. Typically, these algorithms predict a list of putative peptide sequences using a combination of MS/MS fragmentation data and the precursor ion mass for each peptide. A few software packages take this process one step further by combining the *de novo* sequenced peptide information with a proteomic database search, and thus allow mapping of identified peptides onto a homologous sequence [9–11]. Together with manual spectra interpretation and assembly of peptides into amino acid chains, *de novo* sequencing of proteins by mass spectrometrists is feasible [12], albeit resource intensive.

When it comes to antibodies, the diversity in protein sequences present a particular challenge to the peptide *de novo* sequencing and database matching strategy. Antibody sequences are fine-tuned to a particular antigen to maximize the protein-protein interactions. Given the exposure to many antigens in a lifetime, the number of unique antibody sequences in humans is estimated to be greater than ~10 billion [13]. This sequence diversity is created by the recombination events that produce the final heavy and light chain sequences of an antibody, as well as somatic hypermutations during affinity maturation, and isotype switching to modulate the effector function. As a result, each immune reaction produces unique sequences, which are fathomably not available in proteomic databases.

The strategy employed today to *de novo* sequence antibodies involves interpretation of spectra by experts, iterative searches facilitated by humans, and manual sequence assembly. An alternative strategy is using orthogonal methods, such as phage display, in conjunction with mass spectrometry to sequence antibodies [14]. However, these steps incur a heavy cost

on the time needed to complete an analysis, which in turn prevents antibody *de novo* sequencing from being applicable to high-throughput applications.

Here, we describe the methodology and a new algorithm, Supernovo™ (Protein Metrics Inc.), for automated *de novo* sequencing of monoclonal antibodies. We applied a combination of peptide *de novo* sequencing, proteomic database searching with wildcards, *in silico* recombination of antibody coding exons, and final sequence assembly from identified peptides to automatically sequence monoclonal antibodies using bottom-up proteomics. We discuss the novel and high-throughput applications in the biopharmaceutical industry, which are enabled by the hands-free nature of the sequence determination process.

Methods

Sample Preparation

Monoclonal antibody samples were diluted to 1 mg/mL concentration using ultrapure phosphate buffered saline solution pH 7.4 (VWR) and denatured by adding 360ul of 8M Gunadine HCl, 4mM EDTA solution (Sigma-Aldrich). 25mM DTT was added and the sample was incubated at 37°C for 1 hour to reduce disulfide bonds. Reduced antibody was then alkylated with 60 mM iodoacetamide or iodoaceticacid (Sigma-Aldrich) at room temperature for 30 minutes. At the end of alkylation, 37 mM DTT was added to quench unreacted iodoacetamide. Sample was immediately buffer exchanged into 50 mM Tris, 1mM $CaCl_2$, pH 8.0 using Zeba spin desalting columns (Thermo Scientific). Proteolytic digestion was carried out by addition of enzyme (sequencing grade Trypsin, Promega/Chymotrypsin, Roche/Lys-C, Promega/Pepsin, Promega) at a 1:20 enzyme to protein ratio. The trypsin and Chymotrypsin digest samples were incubated at room temperature and the Lys-C digest sample was incubated at 37°C over night with gentle agitation. All the samples were acidified to 1% final TFA after enzymatic digestion. For the pepsin digest, sample was acidified to pH 2 and the digestion was allowed to proceed at 37°C for 1 hour. Pepsin digest was inactivated by heating the sample at 95°C for 15 minutes. The antibody digests were desalted as described [15]. Briefly, 30µl slurry of R2 20µm Poros beads (Life TechnologiesCorporation) was added to each sample and the samples were incubated at 4°C with agitation for 2 hours. The beads were loaded onto the empty TopTip (Glygen) using a microcentrifuge for 30 sec at 1500 RPM. The sample vials were rinsed three times with 0.1% TFA and each rinse was added to the corresponding Toptip followed by microcentrifugation. Extracted poros beads were further rinsed with 0.5% acetic acid. Peptides were eluted by addition of 40% acetonitrile in 0.5% acetic acid followed by the addition of 80% acetonitrile in 0.5% acetic acid. The organic solvent was removed using a SpeedVac concentrator and the samples were reconstituted in 0.5% acetic acid.

Liquid Chromatography/Mass Spectrometry

Agilent AdvanceBio peptide mapping column (1.0 × 150 mm, 2.7 µm particle size) coupled to an Agilent 1290 UPLC was kept at 65 °C and equilibrated with 0.1% formic acid. 2 µg of the digested sample was loaded on to the column, followed by 5 min wash using 0.1% formic acid, and 35 minute gradient of 0–40% acetonitrile, followed by 10 min gradient of 40–90% acetonitrile in 0.1% formic acid. Eluents were analyzed on an Orbitrap Fusion mass

spectrometer (Thermo Scientific), with a top 5 DDA method and 5 second dynamic exclusion to ensure multiple MS2 triggers per elution peak. Both precursors and fragments were measured in the orbitrap using 60,000 and 17,500 resolution (at 400 m/z) respectively. In an alternative method, an aliquot of each sample was loaded onto the EASY spray 50cm C18 analytical column (<2μm bead size) using the auto sampler of an EASY-nLC 1000 HPLC (ThermoFisher) in solvent A (2% acetonitrile, 0.5% acetic acid). The peptides were gradient eluted directly into the Orbitrap Elite mass spectrometer (Thermo Scientific) using a 30minute gradient from 2% to 25% solvent B (90% acetonitrile, 0.5% acetic acid), followed by 10 minutes from 25% to 35% solvent B and 10 minutes from 40% to 100% solvent B. The Orbitrap Elite mass spectrometer acquired high resolution full MS spectra with a resolution of 60,000 (at *m/z* 400) using an AGC target of 1e6 with maximum ion time of 200 ms and scan range of 00 to 1500 *m/z*. Following each full MS, a data dependent scan function was used to obtain five HCD and ETD spectra at a resolution of 15000 (1 microscan). The AGC target was set to 5e4 using a maximum ion time of 100 ms, an isolation window of 2 *m/z,* fixed first mass of 150 *m/z,* and dynamic exclusion of 15 seconds. Normalized Collision Energy (NCE) was set to 30 for HCD fragmentation and ETD was performed with a reaction time of 100ms using supplemental activation at NCE of 35.

Analysis

Data analysis is fully automated through the Supernovo software. Raw data files from any mass spectrometer and MS/MS fragmentation using collision induced dissociation (CID), high energy collision dissociation (HCD), and electron transfer dissociation (ETD) fragmentation methods are supported. An overview of how the software works is shown in Figure 1. Typically, the only inputs given to the software are the raw data files for multiple enzymes, the alkylating agent if any, and the mass tolerance settings for precursor and fragment ions. Supernovo then automatically constructs a template antibody sequence, and proceeds to iterative *de novo* sequencing until a final sequence is converged. These steps are described in more detail below.

Database construction—Monoclonal antibody sequences are constructed *in vivo* by germ line genetic rearrangement of the variable (V), diversity (D), joining (J), and constant (C) regions (Figure 2), followed by somatic hypermutations on the assembled sequence [16]. The diversity in antibody sequences is a result of combinatorial and junctional diversity of the V(D)JC recombination, which is mediated by recombination activating genes 1 and 2 (RAG-1 and RAG-2) enzymes and terminal deoxynucleotidyl transferase (TdT) [17]. Further diversity is introduced to the assembled sequence via an isotype switching mechanism and hypermutations on the final sequence during a B-cell's maturation. A majority of the diversity and the mutations are focused on the complementary determining regions (CDRs), although framework differences to the germline are not uncommon. We replicated part of the biological V(D)JC recombination process *in silico* in order to determine a starting template sequence for iterative *de novo* sequencing based on mass spectrometry data.

Individual amino acid sequences for V- D- J- and C-segments of various species were obtained from the international ImMunoGeneTics information system (IMGT®) database [18]. We consider C-regions to be the spliced full CH1-H-CH2-CH3-CHS sequence for IgGs, and we allow flexibility for isotype switching. Supernovo performs an initial database search using the Byonic engine [19], and identifies the most likely germline candidates for V- and C- regions independently (Figure 1). Since the J-segment is short and flanked by highly variable D-region in the heavy chain, Supernovo extends each of the J-region sequences in the database with the selected V- and C- regions. Utilizing Byonic's wildcard search [19], Supernovo thus identifies the most likely J-region, and subsequently constructs the first full template sequence comprosing V-, J-, and C- segment. Antibody combinatorial diversity is thus replicated *in silico* considering all possible V-J-C combinations. Because the D-region is only found in heavy chain CDR3 and typically subjected to extensive somatic hypermutation and sequence extensions at the D/J junction, we omitted this region from in-silico assembly.

De novo sequencing—Once a template has been chosen, Supernovo iteratively improves the sequence by de novo sequence candidate generation using constrained de novo sequencing [20] and with repeated wildcard searches using Byonic [19, 21], thus augmenting and automating the manual process described in [12]. A wildcard modification has the mass of the precursor ion minus the mass of a candidate peptide, that is, the missing or extra mass required to match the precursor mass. The wildcard modification is applied in turn to each residue in the candidate peptide. Candidate sequences for the next Byonic search are generated by replacing wildcard mass deltas (denoted by curly braces {.}) with new sequence. For example, …VG{+14.0162}S… might become …VAS…, …LGS…, …VGT…, and other possibilities. Candidate sequences are also supplied by a de novo candidate generator [20] in order to identify, at least approximately, spectra without high-scoring wildcard matches. Once a new candidate sequence is chosen, the process is repeated until the last 2 iterations converge to the same protein sequence.

Supernovo's iteration is entirely data-driven and applies no biological bias nor special knowledge of conserved antibody sequence. Any substitution or insertion/deletion is possible at any point in the antibody sequence. Thus even if a wrong initial template for the V- or J- regions was chosen, the correct sequence will be determined via the iterative algorithm. The process resembles DNA assembly, but with the additional challenges of short reads (peptides with typical lengths of 5 – 30) and incomplete/uneven coverage.

Inspection and corrections—Supernovo outputs a final sequence once iterations have converged, and thus allows the scientist to inspect the results using a viewer. Inspection is partitioned to several stages: At the high level, amino acid residues that are poorly supported by MS/MS data are highlighted on the sequence (Figure 3). When the low confidence sequences are at the constant region where germline deviations are unlikely to happen, we typically accepted the determined the sequence as it is. At a medium level, Supernovo shows metrics for fragmentation and digestion summary per amino acid. These correspond to the summary of the accumulated evidence for each cleavage between amino acid residues, at both MS1 (digestion) and MS2 (fragmentation) levels. Finally, a detailed inspection is

performed at the CDR regions by examining the MS1 and MS2 spectra, intensity of the peptide (XIC), mass errors on precursor and fragments, as well as the predicted and observed retention time of the peptides.

Results

Monoclonal antibodies' *de novo* sequencing

We tested our *de novo* sequencing algorithm on bottom-up mass spectrometry data obtained on several monoclonal antibodies of publicly disclosed sequences: U.S. National Institute of Standards and Technology's reference monoclonal antibody 8671 (NIST mAb), which was recently made available to the public; a therapeutic antibody, Bevacizumab (Avastin); a stable-isotope labeled monoclonal antibody, SILuMab; Waters Inc. mass check standard (Waters mAb); and five commercially available monoclonal antibodies typically used in immunology research, SP34-2, UCHT1, OKT3, FN-18, and 29E.2A3.

For the NIST mAb, four raw data files corresponding to four enzymatic digestions by Trypsin, Lys-C, Pepsin, and Chymotrypsin were imported into Supernovo. Without any additional input, Supernovo was able to determine to correct sequence of the antibody using default settings. Figure 3 shows the sequence coverage in the heavy chain V-region including all three CDRs. The red and yellow highlighted segments of the sequence correspond to low and medium confidence regions based on ion fragmentation covering these residues. Variable modification such as oxidation, deamidation, and N-linked glycans are shown as highlights on individual peptides. All PTMs are quantified and may be inspected in the viewer.

The sequencing results of other mAbs are summarized in Table 1. All proteins except for the Waters mAb converged to 100% correct V region sequences for both the heavy and light chains. In Waters mAb, Supernovo identified two deviations from the published sequence, M49G/G50M and S68T/T70S, which were previously observed [22]. In addition, one Asn residue in the sequence was identified instead of the expected Asp, which was easily noticed since all asparagines in all peptides showed deamidation (Asn[+1] = Asp). Finally, Supernovo showed a GA motif instead of the published Q, which is isobaric. See supplemental data for all sequence coverages and comparisons for the Waters mAb.

Stress testing Supernovo

On the data set acquired on the NIST Monoclonal Antibody Reference Material, the entirety of heavy and light chain sequences deduced by Supernovo were correct, with the exception of isoleucine/leucine ambiguity. To evaluate the robustness of the algorithm more rigorously, we conducted two series of *in silico* stress tests.

In the first stress test, the analysis was artificially challenged by randomly discarding increasing number of spectra from the data set prior to submission to Supernovo. Figure 4 shows the number of sequence mistakes Supernovo makes versus percentage of spectra that were discarded. For up to ~90% of spectra discarded, where the data set submitted to Supernovo contained ~1,000 spectra, Supernovo's answer is either completely correct or manageable. Manageable mistakes are defined as unexpected insertions into the framework

scaffold, or unconserved constant region mutations, both of which can be spotted and corrected easily. As the amount of spectra discarded increases above 90%, the number of mistakes in Supernovo's answer predictably increases. Since the main cause is the lack of fragment ion information covering a sequence segment, these mistakes were marked as "low confidence" at the amino acid level by the software. The results of this test suggest that even with sparse information, Supernovo is able to determine and assemble the correct sequence, and highlight low confidence regions accurately.

In the second stress test, we introduced a randomly mutated starting antibody sequence template to the software, thus skipping the *in silico* V(D)JC recombination step. Under normal circumstances, the initial antibody scaffold is constructed by piecing together V, J, and C segments from germline database, and typically has high sequence similarity to the final sequence in at least the constant regions. In this test, the initial antibody scaffold was randomly mutated to differ from the correct sequence. Figure 5-top panel shows the number of mistakes Supernovo makes as the initial scaffold differs more and more from the final answer. For up to 20–30% deviation of initial vs. final, Supernovo's sequence mistakes are relatively low and manageable. At higher number of sequence deviation, Supernovo's mistakes have some interesting characteristics: (1) The sequence errors that are marked as high confidence, i.e. strongly supported by fragmentation data, are assembly errors, i.e. correct sequence peptides placed in the wrong location in the protein. (2) Mistakes tend to occur repeatedly at the similar "problematic spots" along the sequence Figure 5-bottom. These spots concentrate around the antibody hinge region, N-linked glycosylation site, and protein N-terminus. It is somewhat expected that *de novo* sequencing these regions would be problematic without a highly homologous template, because fragmentation data supporting the sequences is typically sparse. These results suggest that Supernovo is capable of performing *de novo* sequencing successfully on a wide diversity of antibodies, including antibodies derived from animal species where the immunoglobulin germline segments are not well-known.

Discussion

Advances in mass spectrometers in terms of resolution and mass accuracy, as well as improvements to electrospray ionization and fragmentation techniques, enabled scientists to perform protein *de novo* sequencing using a combination of software and manual interpretation. A variety of tools that are capable of *de novo* sequencing individual spectra, and thus providing peptide level information, significantly simplified the process. Consequently, the combination of LC/MS and bioinformatics surpassed Edman degradation in terms of ease of experimentation and analysis. The throughput remained a challenge however, particularly in the biopharmaceutical industry where time constraints are a major concern. With the availability of a completely automated and robust tool, *de novo* sequencing could now be employed as a routine application suitable for an industrial setting.

Practicalities and Challenges

De novo antibody sequencing using Supernovo is automated and hands free. Since the accuracy of the results is highly dependent on the laboratory techniques and instrumentation,

the determined sequence needs to be inspected by the analyst. We triaged the inspection process to maximize the throughput of the sequence verification. In the first pass, we focused on whether the sequence has low confidence regions, and whether these are due to post translational modifications (PTMs) or exact mass substitutions such as residue order, GA/Q, GG/N, N[+1] vs. D. We were able to correct these sequences, validate the correction, and prepare reports with ease. In the second tier, we observed low confidence CDR regions, unconserved insertions or mutations, or intact mass mismatch to the theoretical mass of the determined sequence. These were often due to sequence segments that produced too long or too short peptides through enzymatic digestion, or lack of fragmentation information in the constant region. Using germline homology, we could typically complete the sequencing with some analysis effort and without the need to acquire more experimental data. In the last tier, we classified any data set as inadequate, if they showed constant regions with less than 95% sequence homology to germline. Keeping time constraints in mind, we applied tier two data acquisition on these samples using different enzymes or using Fab enrichment. The triaging strategy combined with the hands-free operation of the software allowed antibody *de novo* sequencing to be performed and completed in a high throughput fashion.

In antibody *de novo* sequencing, the challenging section is often the heavy chain CDR-3 region, which is typically the most diversified region of an antibody sequence, owing to hypermutation and the sequence extension mechanism during V(D)J recombination mediated by terminal deoxynucleotidyl transferase (TdT). If the antigen binding is facilitated through hydrophobic interactions, a common scenario, heavy chain CDR-3 may contain an abundance of residues that produce short peptides when cleaved by semi-specific enzymes such as chymotrypsin, pepsin, or elastase. The same hydrophobic patch may not contain charged residues favored by the most commonly used specific enzymes, which in turn would result in long peptides which do not fragment well with CID/HCD. Thus, due to a lack of fragmentation data, the first pass of routine proteolytic digestion, LC/MS, and analysis may not converge to a high-confident sequence for these antibodies. The solution is simply using alternative specific enzymes for proteolysis, or limited time non-specific enzyme digestion to prevent over-cleavage, and repeating LC/MS and data analysis.

An exciting development in the mass spectrometry field relevant to *de novo* sequencing is the simple and routine use of alternative fragmentation methods such as ETD and UVPD. When the fragmentation of long peptides is no longer a challenge, unequivocal data will be available to the software for automated *de novo* sequencing.

The final challenge in antibody *de novo* sequencing is to remove the ambiguity of isobaric residues leucine and isoleucine. Innovative experimental methodology to remove this ambiguity is continuously being developed [23]. The accuracy of these method in Ile/Leu assignment is expected to surpass the current paradigm of using germline homology and chymotryptic digestion specificity to determine an assignment with ~80% confidence.

Applications in Biopharmaceutial Discovery

One of the most common uses of antibody *de novo* sequencing is to recover information that was lost. Bogdanoff *et al.* [12] recently sequenced an antibody with a therapeutic potential whose cell line was not available. Since DNA based sequencing was not possible, protein *de*

novo sequencing was the only option to revive the material. Both academic and industry core labs are often asked for this kind of application, although the demand is probably not on a regular basis.

A disease-specific utility was reported by Bergen et al. [24], where the authors have isolated a monoclonal antibody from a multiple myeloma patient's serum and determined the unique antibody's light chain CDR sequence. This clonotypic peptide may then be used to monitor minimal residual disease after treatment of the cancer, using a blood sample and without the need of a bone marrow extraction, which is an invasive method currently used in the clinic.

A more contemporary application of antibody *de novo* sequencing is determining the sequence and sequence variants of an innovator biologic, so that a biosimilar may be developed. With $67 billion worth of biopharmaceutical product patents expiring by 2020, there is an increasing market for biosimilar production, where *de novo* sequencing is used as a first step to determine and verify the innovator amino acid sequence.

Additionally, two other practical applications of antibody *de novo* sequencing are attractive to innovator biotech/biopharma companies at the discovery stage: 1) isotype selection and 2) bispecific proof of concepts. In the former, the main goal is to investigate the role of the isotype on the agonistic potential of surrogate antibodies. Typically, biopharmaceutical companies initiate a hybridoma or phage display campaign to raise antibodies against a target human protein (antigen) of interest. Monoclonal antibodies against the target's mouse or rat surrogate, however, may already be commercially available. Using *de novo* sequencing, the team may sequence the commercial anti-mouseTarget antibody, and recombinantly generate chimeric versions with various human Fc variants. These chimeric antibodies would be tested *in vivo* and *in vitro* in transgenic mice, which allows assessing the contribution of isotype selection to the overall functionality of the antibody. The transgenic mice expresses human FcγR, and thus this strategy informs on the data-driven Fc choice for the therapeutic candidate. Without *de novo* sequencing, this approach would only be feasible via obtaining the original hybridoma cell line of the commercial antibody, which is not always possible or feasible, or initiating a new campaign for generation of a surrogate antibody. Both incur significant increases in time and cost compared to mass spectrometry based *de novo* sequencing.

The second discovery application is particularly attractive to the Immunology and Immuno-Oncology fields. A monoclonal antibody is a dimer with both "arms" binding to the same target. Bispecific antibodies, on the other hand, can bind to more than one target, and thus have sparked an interest in the industry due to a multitude of novel target engagement possibilities. Choosing the combination of targets that would have the desired efficacy may be challenging [25]. In some cases, bispecific antibodies showed improved efficacy over combination therapies [26, 27], although negative cases have also been reported [28]. Thus, a careful selection and screening of target combinations is required. *De novo* antibody sequencing could be used for this purpose to generate a series of bispecific proof-of-concept molecules for the determination of optimal target combinations. Commercially available and functionally validated antibodies against a variety of antigens may be *de novo* sequenced, which facilitates the generation of a panel of recombinant bispecific antibodies that engages

two targets at once. *In vivo* and *in vitro* testing in primary murine and human assays allows then for the determination of optimal combination of targets that are synergistic in a bispecific format, prior to initiation of a therapeutic campaign, and therefore derisking the proposal. Without *de novo* sequencing this approach for all intents and purposes would not be feasible.

Conclusions

Today's liquid chromatography and mass spectrometry instrumentation and laboratory techniques are routine and adequate enough to perform high throughput bottom-up proteomics suitable for *de novo* sequencing. The challenge, and thus the bottle neck, remains to be bioinformatics. Utilizing *in silico* V(D)JC recombination and a novel algorithm, Supernovo addresses this challenge by providing hands free operation, robust results as demonstrated by stress tests, and metrics and visualizations to help the scientist validate the results, enabling routine and high throughput *de novo* sequencing of monoclonal antibodies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Edman P. A method for the determination of amino acid sequence in peptides. Arch Biochem. 1949; 22:475. [PubMed: 18134557]
2. Shemyakin MM, Ovchinnikov YA, Vinogradova EI, Kiryushkin AA, Feigina MY, Aldanova NA, et al. The rational use of mass spectrometry for amino acid sequence determination in peptides and extension of the possibilities of the method. FEBS Lett. 1970; 7:8–12. [PubMed: 11947417]
3. Pham V, Tropea J, Wong S, Quach J, Henzel WJ. High-throughput protein sequencing. Anal Chem. 2003; 75:875–882. [PubMed: 12622379]
4. Pham V, Henzel WJ, Arnott D, Hymowitz S, Sandoval WN, Truong BT, et al. De novo proteomic sequencing of a monoclonal antibody raised against OX40 ligand. Anal Biochem. 2006; 352:77–86. [PubMed: 16545334]
5. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal Chem. 2005; 77:964–973. [PubMed: 15858974]
6. Chi H, Sun RX, Yang B, Song CQ, Wang LH, Liu C, et al. pNovo: de novo peptide sequencing and identification using HCD spectra. J Proteome Res. 2010; 9:2713–2724. [PubMed: 20329752]
7. Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom. 2003; 17:2337–2342. [PubMed: 14558135]
8. Ma B. Novor: real-time peptide de novo sequencing software. J Am Soc Mass Spectrom. 2015; 26:1885–1894. [PubMed: 26122521]
9. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics. 2012; 11:M111 010587.
10. Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR. Automated de novo protein sequencing of monoclonal antibodies. Nat Biotechnol. 2008; 26:1336–1338. [PubMed: 19060866]
11. Castellana NE, Pham V, Arnott D, Lill JR, Bafna V. Template proteogenomics: sequencing whole proteins using an imperfect database. Mol Cell Proteomics. 2010; 9:1260–1270. [PubMed: 20164058]

12. Bogdanoff WA, Morgenstern D, Bern M, Ueberheide BM, Sanchez-Fauquier A, DuBois RM. De Novo Sequencing and Resurrection of a Human Astrovirus-Neutralizing Antibody. ACS Infect Dis. 2016; 2:313–321. [PubMed: 27213181]
13. Fanning LJ, Connor AM, Wu GE. Development of the immunoglobulin repertoire. Clin Immunol Immunopathol. 1996; 79:1–14. [PubMed: 8612345]
14. Rickert KW, Grinberg L, Woods RM, Wilson S, Bowen MA, Baca M. Combining phage display with de novo protein sequencing for reverse engineering of monoclonal antibodies. MAbs. 2016; 8:501–512. [PubMed: 26852694]
15. Cristea IM, Williams R, Chait BT, Rout MP. Fluorescent proteins as proteomic probes. Mol Cell Proteomics. 2005; 4:1933–1941. [PubMed: 16155292]
16. Ferrier, P. V(D)J recombination. Landes Bioscience; New York, N.Y: 2009.
17. Jones JM, Gellert M. The taming of a transposon: V(D)J recombination and the immune system. Immunol Rev. 2004; 200:233–248. [PubMed: 15242409]
18. Lefranc MP, Giudicelli V, Busin C, Bodmer J, Muller W, Bontrop R, et al. IMGT, the International ImMunoGeneTics database. Nucleic Acids Res. 1998; 26:297–303. [PubMed: 9399859]
19. Bern M, Kil YJ, Becker C. Byonic: advanced peptide and protein identification software. Curr Protoc Bioinformatics. 2012; Chapter 13(Unit13):20.
20. Bhatia S, Kil YJ, Ueberheide B, Chait BT, Tayo L, Cruz L, et al. Constrained de novo sequencing of conotoxins. J Proteome Res. 2012; 11:4191–4200. [PubMed: 22709442]
21. Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem. 2007; 79:1393–1400. [PubMed: 17243770]
22. Hogan, J. Efficiency Gains with Sequence Variant Analysis by Mass Spectrometry. 2014. http://sciex.com/efficiency-gains-with-sequence-variant-analysis-by-mass-spectrometry
23. Poston CN, Higgs RE, You J, Gelfanova V, Hale JE, Knierman MD, et al. A quantitative tool to distinguish isobaric leucine and isoleucine residues for mass spectrometry-based de novo monoclonal antibody sequencing. J Am Soc Mass Spectrom. 2014; 25:1228–1236. [PubMed: 24845350]
24. Bergen HR 3rd, Dasari S, Dispenzieri A, Mills JR, Ramirez-Alvarado M, Tschumper RC, et al. Clonotypic Light Chain Peptides Identified for Monitoring Minimal Residual Disease in Multiple Myeloma without Bone Marrow Aspiration. Clin Chem. 2016; 62:243–251. [PubMed: 26430073]
25. Kontermann RE. Dual targeting strategies with bispecific antibodies. MAbs. 2012; 4:182–197. [PubMed: 22453100]
26. Dong J, Sereno A, Aivazian D, Langley E, Miller BR, Snyder WB, et al. A stable IgG-like bispecific antibody targeting the epidermal growth factor receptor and the type I insulin-like growth factor receptor demonstrates superior anti-tumor activity. MAbs. 2011; 3:273–288. [PubMed: 21393993]
27. Junttila TT, Li J, Johnston J, Hristopoulos M, Clark R, Ellerman D, et al. Antitumor efficacy of a bispecific antibody that targets HER2 and activates T cells. Cancer Res. 2014; 74:5561–5571. [PubMed: 25228655]
28. Weinblatt M, Schiff M, Goldman A, Kremer J, Luggen M, Li T, et al. Selective costimulation modulation using abatacept in patients with active rheumatoid arthritis while receiving etanercept: a randomised clinical trial. Ann Rheum Dis. 2007; 66:228–234. [PubMed: 16935912]

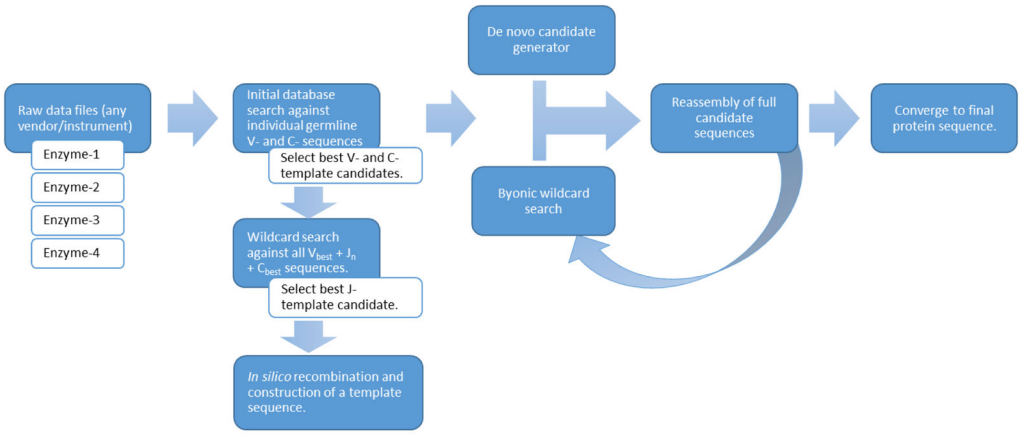


Figure 1.
Supernovo's *de novo* sequencing workflow.

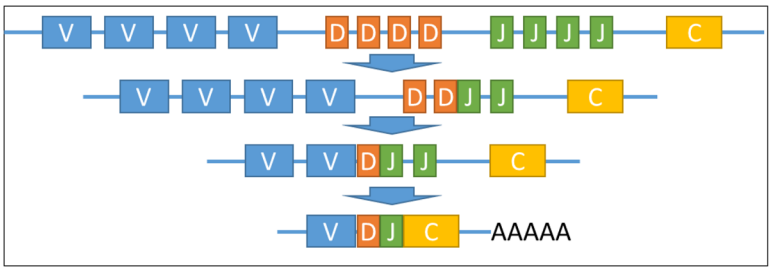


Figure 2.

A simplified view of the Variable (V), Diversity (D), Joining (J), and Constant (C) regions' recombination to produce a full heavy chain sequence. Light chain mechanism is similar, except the D-region does not apply.

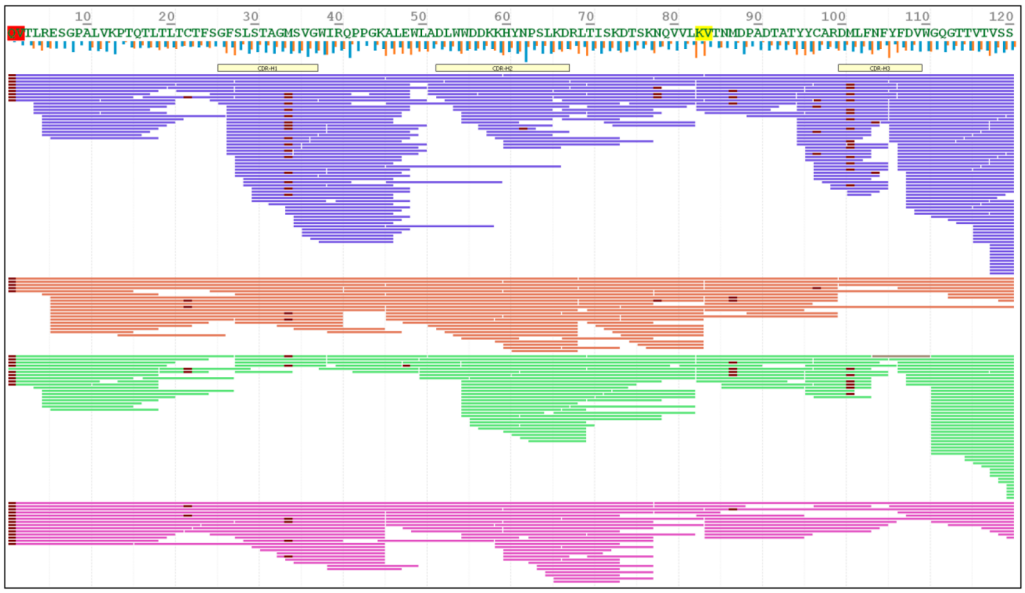


Figure 3.

Sequence coverage view of NIST mAb heavy chain, digested by pepsin (purple), trypsin (red), chymotrypsin (green), and LysC (magenta) enzymes. Fragmentation based confidence level is highlighted in the amino acid letters as red (low) and yellow (medium). Vertical blue bars represent fragmentation summary, and orange bars show digestion summary metrics.

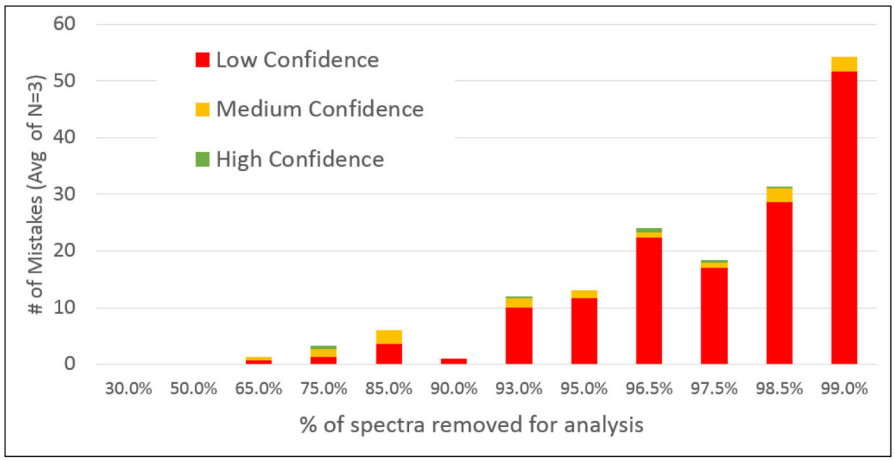


Figure 4.

The number of amino acid assignment mistakes made by Supernovo as a percentage of total number of available MS/MS spectra are randomly discarded.

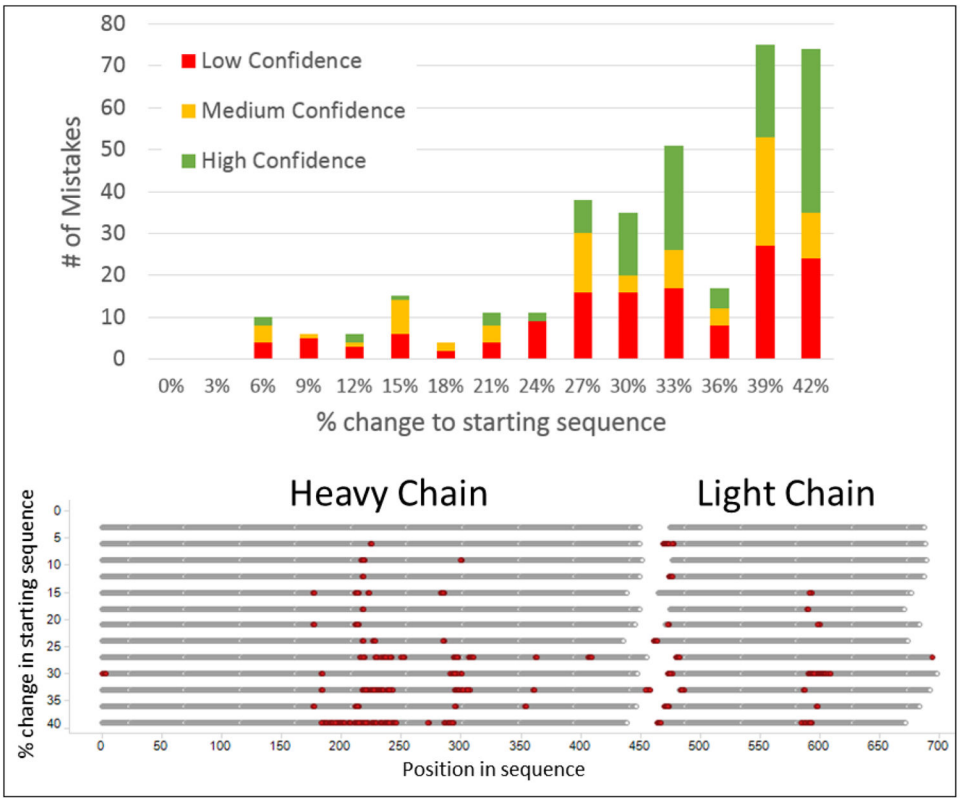


Figure 5.

Number of sequencing mistakes made by Supernovo as a function of random variance in the original template sequence feed (top) and the location of such mistakes in the protein sequence (bottom).

Table 1

Supernovo results on known sequence antibodies.

Sample	Species	Number of mistakes*	Nature of mistakes	Sequence Reference
NIST mAb	Humanized	0		NIST: 8671
Bevacizumab	Humanized	0		Genbank: ACW37587.1
SILuMab	Human/SILAC	0		Sigma-Aldrich: MSQC3
Waters mAb	Mouse	1 or 2	N[+1]→D, Q→GA(?)	Waters: 186006552
SP34-2	Mouse	0		GenBank: AFQ73617.1
UCHT1	Mouse	0		GenBank: AAE23188.1
OKT3	Mouse	0		GenBank: AAE02215.1
FN-18	Mouse	0		GenBank: AAB71638.1
29E.2A3	Mouse	0		GenBank: AGY23748.1

*sum of light chain and heavy chain V-regions.