Published in final edited form as:

Nat Biotechnol. 2008 December ; 26(12): 1336–1338. doi:10.1038/nbt1208-1336.

Beyond Edman Degradation: Automated *De novo* Protein Sequencing of Monoclonal Antibodies

Nuno Bandeira¹, Victoria Pham², Pavel Pevzner^{1,*}, David Arnott², and Jennie R Lill²

¹ Department of Computer Science and Engineering, University of California, San Diego, CA 92093.

² Protein Chemistry Department, Genentech Inc. 1 DNA Way, South San Francisco, CA 94080.

Abstract

De novo protein sequencing of monoclonal antibodies is required when the cDNA or the original cell line is not available, or when characterization of posttranslational modifications is needed to verify antibody integrity and effectiveness. We demonstrate that Comparative Shotgun Protein Sequencing (CSPS) based on tandem mass spectrometry can reduce the time required to sequence an antibody to 72 hours, a dramatic reduction as compared to the classical technique of Edman degradation. We therefore argue that CSPS has the potential to be a disruptive technology for all protein sequencing applications.

Antibodies have been exploited as indispensable reagents for biomedical research and as diagnostic and therapeutic agents^{1, 2}. The specificity and effector functions of antibodies are highly dependent on the amino acid sequence and the presence (or absence) of specific modifications³. Although DNA sequencing is routinely used in the initial characterization of monoclonal antibodies, subsequent mutations and modifications are typically recognized by analysis at the protein level. Pre-clinical antibodies may be derived from immunized hosts, commercial sources, gifts from collaborators, or from hybridomas that no longer secrete antibodies and for which the cDNA is not available. It is therefore critical to sequence the antibodies in order to monitor the integrity of the molecule, to troubleshoot performance in pre-clinical assays, to regenerate cDNA by reverse engineering, ultimately to perform quality control⁴. In addition, protein-level rearrangements (such as observed on IG4 antibodies) can only be revealed by protein level analysis.

Sequencing of (unknown) proteins (and antibodies in particular) remains a challenge. Since antibodies are not directly inscribed in the genome and are constantly created anew, tandem mass spectrometry (MS/MS) database search approaches are not applicable, making Edman degradation the only viable option⁵. This is a low-throughput and time-consuming approach since it is characterized by short peptide reads (limited to about 30 aa), requires proteolytic digestion, peptide fractionation, and peptide-by-peptide sequencing. Mass spectrometry can rapidly generate data that can be used either for cDNA primer design (followed by ≈2 weeks of additional experiments) or combined with Edman degradation to expedite sequencing⁶. While this hybrid MS/MS+Edman approach has been successfully applied to the *de novo* sequencing of antibodies⁴, the Achilles heel of MS/MS sequencing has been the interpretation of spectra (the accuracy of MS/MS sequencing algorithms remains low with only ~30% of spectra correctly reconstructed⁷⁻⁹). Bridging this MS/MS sequencing gap not only significantly reduces the total sequencing time but also considerably reduces the sequencing

* To Whom Correspondence should be addressed: Tel: 858 822 4365, ppevzner@cs.ucsd.edu.

costs and required expertise (i.e. no need for additional Edman+cDNA sequencing instrumentation).

We recently introduced Shotgun Protein Sequencing (SPS) based upon MS/MS spectra from overlapping peptides^{10, 11}. While we demonstrated the feasibility of assembling MS/MS spectra from snake venoms into long protein contigs, assembling spectra into intact proteins requires addressing additional experimental and computational challenges¹⁰. Key experimental challenges include optimizing protease cocktails for generating rich peptide ladders and adopting the new generation of highly accurate mass spectrometers for SPS. The key computational challenge is not unlike “comparative fragment assembly” in classical DNA sequencing when a known genome (e.g., human) is used as a template for assembling another genome (e.g., macaque)¹².

We developed Comparative SPS (CSPS) for assembling spectra into unknown proteins using known proteins as templates (<http://proteomics.bioproteomics.org/Software.html>). Using CSPS, we were able to *de novo* sequence unknown monoclonal antibodies in less than 72 h. We further demonstrate that CSPS identifies unexpected modifications that would frequently go unrecognized by either Edman sequencing or MS/MS database searching.

We obtained two monoclonal antibodies that had been raised against the B and T cell Lymphocyte Attenuator molecule (BTLA): a first-generation antibody (aBTLA) and a mutated version of the original species (mt-aBTLA). Antibodies were raised in mice against human BTLA and were selected for their ability to attenuate T cell responses *in vitro* to protect against graft versus host disease. The antibodies were separately digested with Lys-C, Glu-C, Asp-N, chymotrypsin, pepsin and trypsin and the resulting peptide mixtures were analyzed with LTQ-FTMS and LTQ-Orbitrap instruments (see Supplement).

de novo MS/MS sequencing is complicated by incomplete peptide fragmentation and unexplained peaks. SPS addresses this difficulty with a three-stage alignment→assembly→consensus approach¹⁰ similar to DNA fragment assembly (see Fig.1): The alignment stage identifies *spectral alignments* from overlapping peptides¹¹, the assembly stage combines the spectral alignments into *spectral contigs*, and the consensus stage determines the sequences of spectral contigs resulting in *protein contigs*. Unlike DNA assembly that aligns (unambiguous) DNA reads, SPS aligns un- interpreted spectra of modified and unmodified peptides alike, each of which could be interpreted in thousands of different ways (possible *de novo* peptide reconstructions).

Comparative SPS (CSPS) complements SPS by using homologous sequences from known proteins (e.g., known antibodies) as templates to assemble unknown proteins. CSPS first constructs a set of homologous proteins by matching the SPS contigs against the protein database (all NCBI-nr rat/mouse sequences were used here) and further scores each protein by the overall alignment score of all contigs matched to this protein (see Supplement for algorithmic details). All proteins with scores above the threshold are selected and the theoretical spectra of these proteins are constructed. For our purposes, the theoretical spectrum of a protein is the set of all possible b-ions representing an “idealized” top-down spectrum of the protein. The resulting “long” theoretical spectra of the selected proteins are further assembled with real spectra/contigs using the Shotgun Protein Sequencing tool¹¹. The theoretical protein spectra serve as the ‘glue’ connecting SPS contigs that map to at least one common mass on the same theoretical spectrum (ClustalW alignments are used to map multiple homologous proteins to the same reference protein; see Supplement for details). Sets of contigs matched to the same protein but without common masses on the protein spectrum are still ordered but not glued into the same CSPS contig. After application of SPS, a consensus sequence is again derived using

only the mass differences determined from the overlapped spectra (i.e. homology glues contigs but does not directly influence the resulting protein sequence).

The CSPS assembly of aBTLA heavy chain contigs is illustrated in Figure 2. To validate our approach, all resulting CSPS contigs were compared with the aBTLA sequence obtained by manual Edman degradation sequencing (and an MS/MS database search of the constant regions).

SPS resulted in 63 contigs covering 95% of the aBTLA heavy chain (not counting contigs from proteases and contaminants); grouped by CSPS into 3 long contiguous regions (CSPS contigs) of lengths 288, 40 and 92 aa using two homologous proteins (see Supplement). Comparison of the CSPS contigs with the Edman degradation data revealed that the three sequence gaps not covered by CSPS contigs had no coverage by MS/MS spectra. Thus, these gaps were caused by particularities of the sequence that hinder MS/MS analysis rather than by shortcomings of the CSPS algorithm. For example, the [(N)STFRSV(S)] gap contains the NXT motif indicative of glycosylation. Indeed Asn297 is typically glycosylated on the heavy chain of antibodies and this impedes the identification of these fragments. In addition to this area the first three N-terminal amino are missing since the N-terminal peptides were either too short (< 6 aa) or too long (> 18 aa) for MS/MS identifications. CSPS results on the mutant BTLA antibody (mt-aBTLA) were similar with 97% sequence coverage with 3 contigs of lengths 292, 40 and 97. In addition, this sequence clearly illustrates the ability of CSPS to predict multiple mutations and modifications – 25 out of 28 (89%) mass offsets from the closest homologous protein correctly matched the target sequence (see Supplement).

It turned out that the sequence gaps were identical to the corresponding regions in the homologous proteins. When combined with the resulting match to the mass of the intact protein, these identical homologies could be used to connect the long contigs into a contiguous sequence (this step should be taken with caution since multiple mutations may result in compensatory offsets of total mass zero). Even without this final step, sequencing the aBTLA light chain resulted in 2 contigs (34 and 179 aa) covering 97% of the sequence. Similarly, the mt-aBTLA light chain resulted in a single contig of length 217 covering 99% of the target sequence.

Distinguishing Leucine from Isoleucine is an enduring challenge in MS/MS-based peptide sequencing; their identical atomic composition makes these impossible to distinguish without specially designed MS/MS protocols¹³. Thus, distinguishing Leucine from Isoleucine and sequencing regions not covered by MS/MS spectra represent the directions where CSPS and Edman degradation may complement each other. We remark that while I/L assignments should be done with caution, our homology-based inference correctly identified most of these residues.

CSPS turns the inconvenience of sample-handling modifications into valuable redundancy that can be used to help decode the protein sequence^{10, 11}. Figure 1 illustrates that protein contigs assemble both modified and unmodified versions of overlapping peptides and reveal many modifications. This analysis revealed unexpected modifications such as the conversion of cysteine residues to dehydroalanine (DHA)¹⁴, possibly caused by the elevated temperatures during the reduction step. It also revealed unexpected modifications C+209 and C+223 that cannot be readily explained as chemical adducts and may represent *in vivo* modifications that eluded the conventional analysis (see Supplement). Such unexpected modifications may be critically important for efficacy and safety of antibodies.

CSPS opens up many possibilities for sequence discovery in the biotechnology industry compared to traditional methods. Replacing Edman degradation with CSPS significantly increases the resulting coverage from the same amounts of material (95-99% sequence coverage vs $\approx$ 10% for Edman sequencing), greatly speeds up the analytical protocol and allows one to automatically discover post-translational modifications. Thus CSPS opens a possibility

to correlate unexpected modifications with changes in antibody efficiency while simultaneously tracking mutations. Also, CSPS is already faster than the cDNA sequencing route commonly employed in many laboratories.

While we demonstrated that CSPS can automatically sequence antibodies, further efforts are needed to improve its efficiency, reliability, and robustness. An important direction is to assign reliability scores to amino acids, further optimize the protease cocktails and incorporate complementary MS/MS fragmentation such as ETD¹⁵.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Dan Eaton, Jill Calamine-Fenaux, Richard Vandlen, Oleg Borisov and Bao-Jen Shyong for help with various aspects of this manuscript. This project was supported by NIH grant NIGMS 1-R01-RR16522.

Appendix A. Antibody variability

Appendix B. Mass spectrometry data acquisition and algorithmic details

Separation of the heavy chain and light chain by SDS-PAGE

Antibodies were reduced in DTT (10 mM in 2 × sample buffer: Tris.HCl (0.5 M, pH 8.0), 20% sodium dodecyl sulfate (SDS), 0.5% bromophenyl blue & 26% glycerol) at 95 °C for 5-10 min. Sample was alkylated using iodoacetamide (IAA, 20 mM in MQ water) at RT for 20 min. The HC and LC were separated on a pre-cast 4-20% Tris glycine gel for 2 h at 120 V. Protein bands were visualized with Coomassie blue stain R250 (0.05% Coomassie blue R250, 10% acetic acid) and de-stained with 10% acetic acid solution. The gel was rinsed thoroughly with MQ water. Bands corresponding to the HC and LC were excised and further de-stained with 50% acetonitrile in 50 mM NH₄HCO₃ solution (50 µL) at RT for 30 min. The solution was removed and gel pieces were dehydrated in acetonitrile (50 µL) at RT for 10 min. The gel pieces were dried in the speedvac to complete dryness.

Proteolytic digestions and mass spectrometry

For tryptic digestion the HC and LC gel pieces were re-hydrated in 25 µL of NH₄HCO₃ buffer (25 mM) containing 0.03 mg/mL of trypsin. Samples were chilled on ice for 1 h, excess trypsin solution was removed, digestion buffer NH₄HCO₃ (25 mM, 25 µL) was added and trypsin digestion was performed at 37 °C overnight. In-gel chymotrypsin, Asp-N, Glu-C and pepsin digestions were performed as described above in the appropriate digestion buffers. Chymotrypsin digestion was performed in 100 mM NH₄HCO₃ at 37 °C for 3 h and overnight. Asp-N and Glu-C digestions were performed in sodium phosphate (50 mM, pH 8.0) at 37 °C overnight. Pepsin digestion was performed in 0.1% TFA at 37 °C for 30 min and 3 h. For in-solution Lys-C and Asp-N digestions, the antibody was denatured in 4 M urea at 95 °C for 10 min and reduced in 10 mM DTT at 60 °C for 1h. Alkylation was performed in the presence of 20 mM IAA at RT for 40 min. Excess detergent and reagents were removed by ultrafiltration (Microcon; 10 KDa MWCO). Lys-C, Asp-N digestions were performed in 0.1 M NH₄HCO₃ and sodium phosphate (50 mM, pH 8.0) respectively at 37 °C overnight. After digestion, peptides were extracted from the gel using 50% acetonitrile/0.1% TFA followed by 100% acetonitrile. Extracts were combined and dried down in a speedvac to ~ 10 µL. Peptide mixtures were analyzed either in an LTQ-FTMS instrument or an LTQ-Orbitrap mass spectrometer.

Mass spectral acquisition

Peptide mixtures from Asp-N, chymotrypsin, pepsin and trypsin digests were analyzed on the nano LC.2D HPLC system (Eksigent, Dublin, CA, USA) coupled to the LTQ-FTMS (Thermo Fisher, San Jose, CA, USA). Peptide mixtures were loaded onto the pre-column for 8 min at 2.5 μ L/min in solvent A (0.1% formic acid in water). Peptide separation was performed on a Pico Tip column (15 cm, O.D.=360, I.D.=75, tip= 15 $\pm$ 1 μ m, New Objective, Woburn, MA, USA) packed with reverse phase C18 material (Magic C18, 2 $\text{\AA}$, 5 μ m, Michrom Bioresources, Auburn, CA, USA) using a 60-min or 90-min gradient of 2 to -0% B (0.1% formic acid in acetonitrile) at a flow rate of 250 nL/min. Peptides eluting from RP-HPLC were introduced into the FTMS *via* a nanospray source (2.5 kV). Data dependent acquisition was performed whereby the full MS scan was acquired in the FTMS and the 5 most abundant ions were selected for MS/MS using 25% relative collision energy and analyzed in the LTQ. Or b) a NanoAcquity UPLC system (Waters, Dublin, CA) where peptides were loaded onto a pre-column (5 μ m Symmetry $\text{\textcircled{R}}$ C18, 180 $\times$ 20 mm) and separated using an analytical column (1.7 μ m BEH-130 C18 column 100 $\times$ 100 mm, Waters, Dublin, CA) with a flow rate of 1 μ L a minute and a gradient of 2% Solvent B to 90% Solvent B (where Solvent A is Water + 0.1% Formic acid and Solvent B is 100% Acetonitrile + 0.1% Formic Acid) applied over 40 min with a total analysis time of 55 min. Peptides were eluted directly into a nanospray ionization source with a spray voltage of 2 kV and were analyzed using an LTQ XL-Orbitrap mass spectrometer (ThermoFisher, San Jose, CA). Precursor ions were analyzed in the Orbitrap at 60,000 resolution. MS/MS was performed in the Orbitrap at 15,000 resolution with the instrument operated in data dependent mode whereby the top 10 most abundant ions were subjected for fragmentation.

Spectral Clustering and Pre-processing

MS/MS peak lists were extracted from the RAW files and converted to mzXML format using ReAdW (<http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW>). The resulting MS/MS spectra were clustered using MS-Clustering (<http://proteomics.bioproteomics.org/>) to increase the signal-to-noise ratio and further processed for precursor charge determination, parent mass correction and replacement of raw peak intensities with likelihood scores as previously described⁷.

Comparative Shotgun Protein Sequencing (CSPS)

As previously described^{10, 11}, the output of Shotgun Protein Sequencing (SPS) is a set of protein contigs, each of which represented as a set of (mass,score) pairs. Differences between consecutive masses reveal amino acid masses (or grouped amino acid masses) and scores indicate the relative confidence on each mass by aggregating likelihood scores from the corresponding masses on all overlapped spectra.

The CSPS algorithm can be described as a series of consecutive stages, starting from a set S of SPS contigs and a database D of possible homologous protein sequences:

1. Determination of candidate homologous proteins; each contig in S generates a set of sequence tags of length 8, allowing for at most one missing mass; a protein in D is considered a candidate homologous protein if it matches at least one sequence tag. The subset of all candidate homologous proteins is then reduced using a common maximum- parsimony approach to retain the minimal number of proteins covering all the contig/protein matches (briefly, we select the protein P matched to most non-selected contigs, mark all contigs matched to P as selected and iterate until all contigs are selected). The final set of candidate homologous proteins is denoted here as DH .

2. Alignment of contigs to protein sequences; this step uses previously described spectral alignment algorithms^{SR1,SR2,10,11}; all contigs are aligned to all sequences in DH, regardless of the tag matches in step 1. At this stage, contig/protein alignments are scored by adding the SPS contig mass scores for all matched masses; while it would be straightforward to further score putative mutations (mass offsets determined by the alignment algorithm) using existing scoring matrices such as BLOSUM, we abstained from doing so because a) the alignments used to derive these matrices may not necessarily reflect the particular recombination/hyper-mutation context of antibody sequences^{SR3} and b) SPS contigs' mass scores provide a data-derived unbiased way to estimate the confidence of contig/protein matches that is usually unavailable for protein/protein alignments. Each contig is assigned to the protein resulting in the highest-scoring contig/protein alignment and the alignment score of each protein is simply the sum of alignment scores of its assigned contigs; the highest-scoring protein is referred to as the reference protein R.
3. Alignment of the reference protein to other homologous proteins; since antibody variability may lead to contigs from the same protein being assigned to different partially-homologous antibody fragments, we further used ClustalW to map contig/protein alignments back to the same reference protein. Pairwise ClustalW^{SR4} alignments were computed between the reference protein and all other proteins in DH and deemed significant if the alignment score was not less than 250 (using default values for all ClustalW's parameters). Significant ClustalW alignments were used to transfer contig/protein alignments to the reference protein R: an SPS contig mass m aligned to amino acid i in protein P becomes aligned to amino acid j in protein R iff i was aligned to j by a ClustalW alignment.
4. Gluing overlapped contigs and determining extended contig sequences; contig/protein alignments to the reference protein R are used to glue contig masses aligned to the same amino acid positions in R. As previously described^{10, 11}, contigs can be represented as a path in an A-Bruijn graph where gluing matched masses corresponds to merging graph vertices and matching in/out edges. Extended contig sequences are simply defined as the highest-scoring path in each resulting connected component.

The following example from the mutant aBTLA light chain illustrates how CSPS uses matches to a reference protein to glue contigs whose overlap is too small for reliable detection of the contig/contig overlap:

Table B1

CSPS coverage of positions 61-96 in the mutant aBTLA light chain; the correct protein sequence (unavailable to the algorithm) is shown at the top followed by the found reference protein sequence (gi|42543442) used to locate the SPS contigs C152 and C48 whose sequences were glued at the masses highlighted in red. While the matching contig masses are insufficient to robustly call the contigs overlap based solely on experimental data, their significant alignment to neighboring locations on the same homologous protein serves as the scaffold supporting their connection into a longer extended contig. Sequence variations are underlined – CSPS adequately corrected the T/S variation between the target/reference proteins and additionally suggests that position 76 may be an Asparagine (N) that is sometimes deamidated to form an Aspartic Acid (D).

Target	E S G V P D R F T G S G S G T D F T L T I <u>S</u> S V Q A E D L A V Y Y C K Q
Ref	E S G V P D R F T G S G S G T D F T L T I <u>T</u> S V Q A E D L A V Y Y C K Q
C152	E S G V P D R F T G S G S G T D F T L
C48	N F T L T I <u>S</u> S V Q A E D L A V Y Y C K Q

References

1. Ferrara N, Hillan KJ, Gerber HP, Novotny W. *Nature reviews* 2004;3:391–400.
2. Reichert JM, Valge-Archer VE. *Nature reviews* 2007;6:349–356.
3. Gilbert, SF. *Developmental Biology*. Edn. 8th. Sinauer Associates, Inc.; Sunderland, MA: 2006.
4. Pham V, et al. *Analytical biochemistry* 2006;352:77–86. [PubMed: 16545334]
5. Pham V, Tropea J, Wong S, Quach J, Henzel WJ. *Analytical chemistry* 2003;75:875–882. [PubMed: 12622379]
6. Gatlin CL, Eng JK, Cross ST, Detter JC, Yates JR 3rd. *Analytical chemistry* 2000;72:757–763. [PubMed: 10701260]
7. Frank A, Pevzner P. *Analytical chemistry* 2005;77:964–973. [PubMed: 15858974]
8. Ma B, et al. *Rapid Commun Mass Spectrom* 2003;17:2337–2342. [PubMed: 14558135]
9. Mo L, Dutta D, Wan Y, Chen T. *Analytical chemistry* 2007;79:4870–4878. [PubMed: 17550227]
10. Bandeira N, Clauser KR, Pevzner PA. *Mol Cell Proteomics* 2007;6:1123–1134. [PubMed: 17446555]
11. Bandeira N, Tsur D, Frank A, Pevzner PA. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104:6140–6145. [PubMed: 17404225]
12. Pop M, Phillippy A, Delcher AL, Salzberg SL. *Briefings in bioinformatics* 2004;5:237–248. [PubMed: 15383210]
13. Armirotti A, Millo E, Damonte G. *Journal of the American Society for Mass Spectrometry* 2007;18:57–63. [PubMed: 17010643]
14. Bar-Or R, Rael LT, Bar-Or D. *Rapid Commun Mass Spectrom* 2008;22:711–716. [PubMed: 18265430]
15. Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. *Proceedings of the National Academy of Sciences of the United States of America* 2004;101:9528–9533. [PubMed: 15210983]

Additional references

- SR1. Pevzner PA, Mulyukov Z, Dancik V, Tang CL. *Genome Res* 2001;11:290–9. [PubMed: 11157792]
- SR2. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. *Nat Biotechnol* 2005;23:1562–7. [PubMed: 16311586]
- SR3. Di Noia JM, Neuberger MS. *Annu Rev Biochem* 2007;76:1–22. [PubMed: 17328676]
- SR4. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. *Nucleic Acids Res* 2003;31:3497–500. [PubMed: 12824352]

Figure 1.

Protein contigs resulting from Shotgun Protein Sequencing of aBTLA tandem mass spectra. **Left:** Protein contig resulting from 24 spectra from the aBTLA heavy chain. Each spectrum is shown superimposed with a sequence of arrows indicating its sequence of recovered masses; modified variants of the consensus sequence are indicated by red arrows (6 different modifications on 7 spectra). **Right:** Protein contig illustrating the variety of discovered modifications on or next to Cys-146 (aBTLA heavy chain); modified residues are indicated with red labels next to the corresponding edge: C+57 (the expected mass, spectra 1,7, not highlighted), C+71 (acrylamide adduct, spectra 2,3,8,11), C-34 (DeHydroAlanine (DHA), spectra 4,5), DHA+16 (oxidation, spectrum 6), M+16 (oxidation, spectra 1,3), either C+209 or DHA+243 (spectrum 9), either C+223 or DHA+257 (spectrum 10).

Figure 2.

Comparative Protein Sequencing. The heavy chain contigs matched to two different proteins (gi|148540420 and gi|148686583) homologous to different regions of the aBTLA heavy chain – 9 SPS-contigs matched gi|148540420, 47 SPS-contigs matched gi|148686583 and 8 contigs matched both (see Figure 2). The protein regions matched by the latter were confirmed by a corroborating CLUSTALW alignment of gi|148540420/gi|148686583. **a)** Homology-derived order of 36 aBTLA heavy chain protein contigs. Each protein contig is represented as a colored dash on the circumference; the color gradient and black arrow at the top indicate the arbitrary contig order (since the contig order is unknown beforehand). Colored arrows indicate the contiguous contig order induced by the homologous proteins and the dashed colored arrows indicate protein contigs in the correct order but separated by sequence gaps. The recovered contig order is indicated by indices next to each contig; negative indices $-j$ indicate that the j -th contig resulted in a reversed sequence (i.e. inferred from a sequence of y-ions rather than b-ions). **b)** Linear rendering of the homology-induced contig order illustrated in part a). **c)** The complete aBTLA heavy chain sequence recovered by our approach; highlighted sections were covered by protein contigs (95% coverage) and the missing amino acids were obtained from the homologous protein sequences.

Figure A.

Center: Structure of a typical immunoglobulin (antibody) protein. Two identical heavy chains and two identical light chains are connected by disulfide linkages. The antigen-binding site is composed of the variable regions of the heavy and light chains, whereas the effector site of the antibody is determined by the amino acid sequence of the heavy chain constant region. Bottom: Rearrangement of the light chain genes during B lymphocyte differentiation. While the developing B cell is still maturing in the bone marrow, one of the 300 or more V gene segments combines with one of the 5 J gene segments and moves closer to the constant (C) gene segment. Top: Rearrangement of the heavy chain genes. A heavy chain gene contains three segments (V, D, and J) that come together to form the variable region, as well as a constant region. Reproduced with permission from Sinauer Associate Inc³.

In vertebrates, antibodies are composed of a pair of heavy chains (HC) (~52-70 KDa each) and light chains (LC) (~25 KDa each) which are linked together *via* disulphide bonds (Fig.1)³. This staggering diversity is facilitated during the early phase of B lymphocyte development by integration of the immunoglobulin variable (v), diversity (d) and joining (j) gene segments³

and is primarily manifested in the hyper-variable regions of the heavy and light chains which constitute the complementarity determining regions (CDRs). The CDRs are mounted on a scaffold of framework residues that vary to a lesser extent among antibodies. The remainder of the antibody framework is termed the constant region and follows a more conserved sequence, although variation may also occur here due to mutations at the genetic level or post translational modifications at the protein level³.