MS2DB+: A Software for Determination of Disulfide Bonds Using Multi-Ion Analysis

William Murad and Rahul Singh*

***Abstract*—MS2DB+ is an open-source platform-independent web application for determining, in polynomial time, the disulfide linkages in proteins using tandem mass spectrometry (MS/MS) data. It utilizes an efficient approximation algorithm which allows the consideration of multiple ion-types (a, a_o, a^*, b, b_o, b^*, c, x, y, y_o, y^*, and z) in the analysis. Once putative disulfide bonds are identified, a graph optimization approach is used to obtain the most likely global disulfide connectivity pattern. Availability—http://haddock2.sfsu.edu/~ms2db/disulfidebond/**

***Index Terms*—Bioinformatics, disulfide bond, mass spectrometry, peptides, proteomics.**

I. Introduction

A DISULFIDE (S-S) bond is a covalent bond, obtained through the coupling of two thiol groups. Such bonds play an important role in protein folding and function. Methods for determining S-S bonds can be broadly grouped into two categories: those that use sequence-level information to predict S-S bonds, and those based on data from mass spectrometry (MS), crystallography or NMR. For a detailed review of these methods, we refer the reader to [1] and references therein. Of the above, MS-based approaches are especially interesting due to their high accuracy. Furthermore, unlike crystallography or NMR, MS-based disulfide bond determination can be carried out with relatively small quantities of the analytes. MS2DB+ is a dedicated software for determining S-S bonds using tandem mass spectrometry data. The unique features of MS2DB+ include:

- *Ability to account for multiple ion-types during analysis and matching*: The determination of S-S bonds requires matching theoretical spectra of ionized protein-digested peptide fragments with the MS/MS spectra. Current state-of-the-art methods [2]–[4] typically consider only b/y ions in their analysis. These ions predominate in collision-induced dissociation (*CID*), which is one of the commonly used fragmentation models. However, non $-b/y$ ions, such as a, a^*, a_o, b^*, b_o, c, x, y^*, y_o, and z ions can also be found, especially in other fragmentation methods such as electron-capture dissociation (*ECD*), electron-transfer dissociation (*ETD*), and electron-detachment dissociation (*EDD*). Furthermore, recent studies indicate that accounting for these other ion types can lead to improved specificity and accuracy in identifying S-S bonds even in cases where b/y ions predominate [5].
- *Algorithmically efficient search process*: The number of disulfide bonded configurations increases rapidly with the number of cysteines, the types of connectivity patterns (defined by inter/intra fragment bonds as well as the number of fragments that can be bonded), and the number of ion-types being considered. This implies that if we consider f fragment ion types, then up to f^k types of fragments may occur for a disulfide-bonded peptide structure consisting of k peptides. If the $i\text{th}$ fragment ion consists of r_i amino acid residues, then the size of the search space Σ consisting of all disulfide bonded ions can be as large as

$$|\Sigma| = O(f^k \times \Pi_{i=1}^{k} r_i). \tag{1}$$

  In MS2DB+, the efficient search of Σ involves two steps. In the *initial matching* step, the mass values of theoretically possible disulfide bonded peptide structures are compared with the precursor ion mass values from the MS-spectra. In this step, the experimental precursor ion mass values are used to filter the search space in a manner similar to that in [4]. In the subsequent *confirmatory* step, the theoretical spectra are compared with the (tandem) MS/MS spectra. The *confirmatory* step is especially important since a disulfide bonded peptide may not correspond to a precursor ion even if they have similar mass. In addition to the precursor ion mass-based filtering, MS2DB+ uses a fully polynomial-time approximation algorithm to truncate the search-space. This method is briefly described in the next section. It should also be noted that in MS2DB+, there are no implicit restrictions on the S-S connectivity patterns.
- *Global disulfide connectivity determination via an optimization strategy*: In its final step, MS2DB+ uses a graph-based optimization strategy to coalesce the S-S bonds identified in the above step into the most likely globally consistent connectivity pattern.

II. Algorithmic Underpinnings of MS2DB+

In MS2DB+, during the initial matching phase, precursor ions from the MS/MS spectra are compared with disulfide-bonded peptide fragments arising from the digested protein. In this initial phase, a two-stage filtering-and-search technique is employed: *first*, any theoretical S-S bonded combination which exceeds the precursor ion mass being matched is automatically discarded. *Second*, the method identifies, from among the remaining disulfide-bonded peptide fragments, those with mass close to the given experimental spectra. That is, we determine the pair (S, t), where S corresponds to the set of

Manuscript received September 12, 2011; accepted July 04, 2012. Date of publication October 19, 2012; date of current version May 29, 2013. *Asterisk indicates corresponding author.*

W. Murad is with the Department of Computer Science, San Francisco State University, San Francisco, CA 94132 USA (e-mail: whemurad@mail.sfsu.edu).

*R. Singh is with the Department of Computer Science, San Francisco State University, San Francisco, CA 94132 USA (e-mail: rahul@sfsu.edu).

Digital Object Identifier 10.1109/TNB.2012.2212029

TABLE I
SEARCH SPACE REDUCTION ACHIEVED BY MS2DB+

Protein	Full Search (exponential)		Proposed Search (polynomial)		DMS decrease	FMS decrease
	DMS size	FMS size	DMS size	FMS size		
ST8Sia IV	2917	632207	970	17892	66.75%	97.17%
Beta-LG	3220	163833	2676	4673	16.89%	97.15%
FucT VII	1657	384463	1116	8771	32.65%	97.72%
C2GnT-I	2093	1466082	973	30338	53.51%	97.93%
Lysozyme	24973	4675068	2481	78198	90.07%	98.33%
B1,4-GalT	3142	1697663	1463	49647	53.44%	97.08%
FT III	4627	632455	2206	3933	52.32%	99.38%
B122A	11767	6035341	7320	119056	37.79%	98.03%
Average DMS and FMS decrease					**48.10%**	**97.94%**

TABLE II
MS2DB+: SENSITIVITY, SPECIFICITY, AND ACCURACY

Protein	*Sensitivity*	*Specificity*	*Accuracy*
ST8Sia IV	1.00	1.00	1.00
Beta-LG	0.50	1.00	0.95
FucT VII	1.00	1.00	1.00
C2GnT-I	0.75	1.00	0.98
Lysozyme	1.00	1.00	1.00
B1,4-GalT	1.00	1.00	1.00
FT III	0.50	1.00	0.94
B122A	0.67	1.00	0.93
Aldolase	X	1.00	1.00
Aspa	X	1.00	1.00
Average	**0.802**	**1.00**	**0.976**

disulfide-bonded peptide fragments (indexed by their mass) and t corresponds to the targeted mass value from the experimental spectra. This step can be thought of as the subset-sum problem for which we obtain near-optimal solutions in polynomial time using an approximation algorithm. Our approximation strategy trims as many elements as possible from the search space based on a trimming parameter ε. For the search space DMS (initial matching), this implies removing as many elements as possible to create the resultant trimmed set DMS^*, such that for every element DMS_i removed from DMS, there remains an element $DMS'_i \in DMS^*$ which is "close" in terms of its mass to the deleted element DMS_i as per (2):

$$\left(\frac{DMS_i}{1+\varepsilon}\right) \le DMS'_i \le DMS_i. \qquad (2)$$

Each match found in the initial matching step is validated in the confirmatory phase to eliminate a correspondence that may have occurred by chance. During this step, ions obtained from the fragmentation of the disulfide-bonded peptide fragments (forming the *FMS* search space) are compared with ions from the MS/MS stage. Again, a two-stage filtering-and-search process conceptually similar to the one described above is used. Next, an empirical match score is calculated based on the number and abundance of fragment ions for each match from the initial matching step. MS2DB+ also supports a probability based scoring model, adopted from [6], to evaluate the match between an experimental spectrum and a theoretical spectrum. Computation of the initial matches and the confirmatory matches provides a "local" (putative bond-level) view of the possible disulfide connectivity of a molecule. Next, the disulfide bonds are integrated to obtain a globally consistent view, since one cysteine can participate in at most one S-S bond. The global disulfide topology is found by computing the maximum weight matching in an undirected graph $G(V, E)$, where the set of vertices V corresponds to the set of cysteines and the set of edges E contains all the putative S-S bonds found after the confirmatory match stage. Fig. 1 presents an example for the glycosyltransferase Lysozyme demonstrating this optimization strategy. In this particular case, ten different putative bonds were initially determined [Fig. 1(a)] and the optimization strategy successfully identified the correct topology: C^{24}-C^{145}, C^{48}-C^{133} [Fig. 1(b)]. Details of the algorithm can be found in [5].

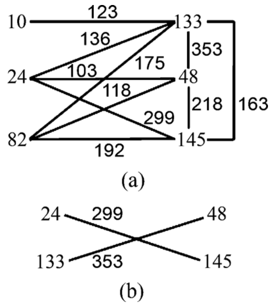

Fig. 1. Bond topology determined for Lysozyme. (a) Ten different putative S-S bridges were identified prior to the global optimization step. (b) Final (global) connectivity obtained after optimization. The edge weights correspond to the match scores.

In Table I, the effectiveness of the approximation algorithm (in terms of trimming the search space) is summarized. Ten different ion types (a, b, b^*, b_o, c, x, y, y^*, y_o, and z) were considered to create Table I. In Table II, we present different metrics describing the performance of the method on a set of ten proteins having varying disulfide bond topologies. Comparative analysis conducted in [5] indicate that results from MS2DB+ were comparable or better than those obtained with the two gold standards in the area, MassMatrix [2] and MS2Links [3]. MS2DB+ also outperformed all predictive methods tested in [5].

III. Implementation

MS2DB+ is coded in PHP and runs on an Apache web server. The source code is publicly available at the MS2DB+ website. The software accepts most of the commonly used MS/MS data formats, including Sequest DTA, mzXML, mzData and mzML. Two modes are supported for analysis: *standard analysis* (completely automatic) and *advanced analysis* (user may customize the thresholds and parameters). Both these modes require three inputs: a) MS/MS file(s) containing experimental data; b) the protein sequence in FASTA format; and c) the protease used to digest the protein sample. The customizable parameters (*for the advanced analysis*) include: initial and confirmatory matching thresholds, MS/MS abundance threshold, which contributes to suppressing MS-data noise, and the trimming parameters ε and δ used to trim the theoretical search spaces.

The user may also: a) choose different combinations of multiple ion types to be considered during the spectral matching (based on knowledge of the dissociation method); b) select the maximum number of missing cleavage sites during protein's digestion; and c) specify the protein region where a disulfide bond is not expected to occur. Once the data has been entered and processed, MS2DB+ presents the global consistent disulfide connectivity in a user-friendly graph and lists the S-S bonds found along with their empirical and statistical scores. For each disulfide bond found, MS2DB+ also presents insights in the matching process, such as the confirmatory matches encountered (containing their mass, charge state, and ion combination details). The fragment matches are color coded by their abundance to aid in analysis.

MS2DB+ also provides relevant intermediary data pertinent to the analysis. This includes: 1) information on all initial matches, including the MS/MS file involved in the match, the precursor ion mass and charge state, the peptide sequences, and the cysteines present; 2) the confirmatory match score for each one of the initial matches; and 3) a list of all disulfide bonds and their respective match score, *pp* score, and *pp*2 score, as determined after the global optimization step. This information is provided in XML format at the bottom of the page in which the disulfide connectivity results are presented.

Acknowledgment

The authors thank T.Y. Yen, R. Varadarajan, and S. Bhattacharyya for the data and feedback. This work was supported by the NSF under Grant IIS-0644418 (CAREER).

References

[1] R. Singh, "A review of algorithmic techniques for disulfide-bond determination," *Briefings Funct. Genomics Proteomics*, vol. 7, pp. 157–172, 2008.

[2] H. Xu, L. Zhang, and M. A. Freitas, "Identification and characterization of disulfide bonds in proteins and peptides from tandem MS data by use of the MassMatrix MS/MS search engine," *J. Proteome Res.*, vol. 7, pp. 13–144, 2008.

[3] E. T. Yu, A. Hawkins, I. D. Kuntz, L. A. Rahn, A. Rothfuss, K. Sale, M. M. Young, C. L. Yang, C. M. Pancerella, and D. Fabris, "The collaboratory for MS3D: A new cyber infrastructure for the structural elucidation of biological macromolecules and their assemblies using mass spectrometry-based approaches," *J. Proteome Res.*, vol. 7, no. 11, pp. 4848–4857, 2008.

[4] S. Choi, L. Jeong, S. Na, H. Lee, H. Y. Kim, K. J. Lee, and E. Paek, "New algorithm for the identification of intact disulfide linkages based on fragmentation characteristics in tandem mass spectra," *J. Proteome Res.*, vol. 9, no. 1, pp. 625–35, 2010.

[5] W. Murad, R. Singh, and T. Y. Yen, "An efficient algorithmic approach for mass spectrometry-based disulfide connectivity determination using multi-ion analysis," *BMC Bioinformatics*, vol. 12, no. Suppl. 1, p. S12, 2011.

[6] H. Xu and M. A. Freitas, "A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data," *BMC Bioinformatics*, vol. 8, pp. 133–142, 2007.