The MS2DB++ Webserver: Disulfide Bond Determination Through Evidence Combination

William Murad and Rahul Singh*, *Member, IEEE*

***Abstract*—MS2DB++ is a web server for computationally determining disulfide connectivity in proteins by combining evidence from multiple methods. The constituent methods implemented as part of the MS2DB++ webserver include a mass spectrometry-based method and two protein sequence-based predictive methods. The software also allows users to incorporate results from up to two other external methods of choice. The results from all these methods are combined using Dempster-Shafer theory through the use of four different formulations for evidence combination. In practice, MS2DB++ can be especially helpful in obtaining the disulfide topology in cases where no single method performs consistently across a set of molecules due to complexity of the bonding topology, specificities of the fragmentation pattern, or limitations of computational models.**

Availability—**http://haddock2.sfsu.edu/~ms2db/ms2db++**

***Index Terms*—Biology computing, decision support systems, disulfide bonds, mass spectrometry, prediction methods, , , proteomics.**

I. Introduction

At THE CURRENT state of the art, multiple classes of computational methods exist for determining disulfide (S-S) bonds [1]. Yet no single class of methods is known to perform accurately across a diverse set of molecules. For instance, computational techniques that are based on analyzing data from mass spectrometry (MS) can err, if 1) there is insufficient fragmentation of precursor ions due to the presence of cross-linked or circular disulfide bonds, 2) the S-S bonds have similar reduction rates under conditions of partial reduction, or 3) if the molecule being analyzed has multiple S-S bonds or large number of cysteines under non-reduction conditions. Sequence-based predictors of S-S connectivity too have weaknesses, since sequence-based features often turn out to be insufficient to predict complex structural features such as the S-S topology.

Different methods however also bring to bear their own advantages; mass spectrometry and crystallography-based methods tend to be highly accurate. Sequence-based predictors, though less accurate, do not require expensive instrumentation and can be run using only the protein sequence. Intriguingly, the results provided by different methods can concur and/or complement each other. For instance, sequence-based predictors may provide a more accurate estimation, than MS-based methods for S-S bonds in regions of a peptide that are poorly fragmented in a mass spectrometer. Conversely, in parts of the same peptide that are sufficiently fragmented, MS-based methods tend to be more accurate.

The aforementioned scenario highlights the need for methods that can rigorously combine evidence from one or more classes of techniques keeping in mind that the results may concur or conflict in parts. This paper describes MS2DB++, an open-source platform-independent web server which implements this idea. In it, results from different methods are combined using the theoretical framework of Dempster-Shafer theory (DST). Evidence combination allows MS2SDB++ to achieve high sensitivity, specificity and accuracy, even under conditions when some of the constituent methods fail to correctly determine specific S-S bonds. The interested reader is referred to [2] for a detailed exposition of the theoretical and algorithmic foundations on which MS2DB++ has been built.

II. Constituent Methods

MS2DB++ provides three different techniques for S-S bond determination: a tandem mass spectrometry-based approach called MS2DB+ [3], [4], a support vector machine (SVM)-based predictor based on [5], and a cysteines separation profiles-based (CSP)-based predictor derived from [6] which uses the separation between oxidized cysteine residues to predict S-S connectivity. The CSP-based approach uses the representation proposed in [7] and the observation that S-S bonding patterns can be used to discriminate structural similarity. Additionally, MS2DB++ allows users to enter disulfide connectivity results from two other external methods of choice, allowing thereby, combination of evidence from up to five distinct methods.

Of the three constituent methods, MS2DB+ identifies, in polynomial time, the disulfide linkages in proteins using tandem mass spectrometry (MS/MS) data. It uses an efficient approximation algorithm which allows the consideration of 12 different ion types in the analysis of disulfide bonding patterns. The SVM-based predictor is trained using 521 features which are generated using two windows of size 13 centered on the termini of every putative cysteine pair to capture the respective local environments. Every residue in the window is encoded by a 20-element bit-vector, where each bit is set to one if the corresponding amino acid is present. Additionally, the distance between pairs of cysteines is used as a feature. In the CSP-based

Manuscript received December 10, 2012; accepted August 07, 2013. Date of current version January 07, 2014. This work was supported by the NSF under grant IIS-0644418 (CAREER). *Asterisk indicates corresponding author.*

W. Murad is with the Department of Computer Science, San Francisco State University, San Francisco, CA 94132 USA.

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNB.2013.2289391

TABLE I
EVIDENCE COMBINATION RULES IN MS2DB++

Rule	Mathematical Definition
Dempster	$\sum_{B\cap C=A;A\neq\emptyset} m_1(B)\times m_2(C)/\sum_{B\cap C\neq\emptyset} m_1(B)\times m_2(C)$
Yager	$\sum_{B\cap C=A} m_1(B)\times m_2(C)$
Campos	$(X\times\sum_{B\cap C=A;A\neq\emptyset} m_1(B)\times m_2(C))/1+\log(X)$; $X=1/\sum_{B\cap C\neq\emptyset} m_1(B)\times m_2(C)$
Shafer	$[(1-\alpha_1)m_1(A)+(1-\alpha_2)m_2(A)]/2$

method, the unknown S-S connectivity of the input protein is predicted to be the same as that of a database protein having the most similar cysteine separation profile. Both the SVM and the CSP-based predictors were trained using a database of 439 proteins. This dataset was obtained after filtering the UniProt SP43 dataset, using the filtering techniques described in [2].

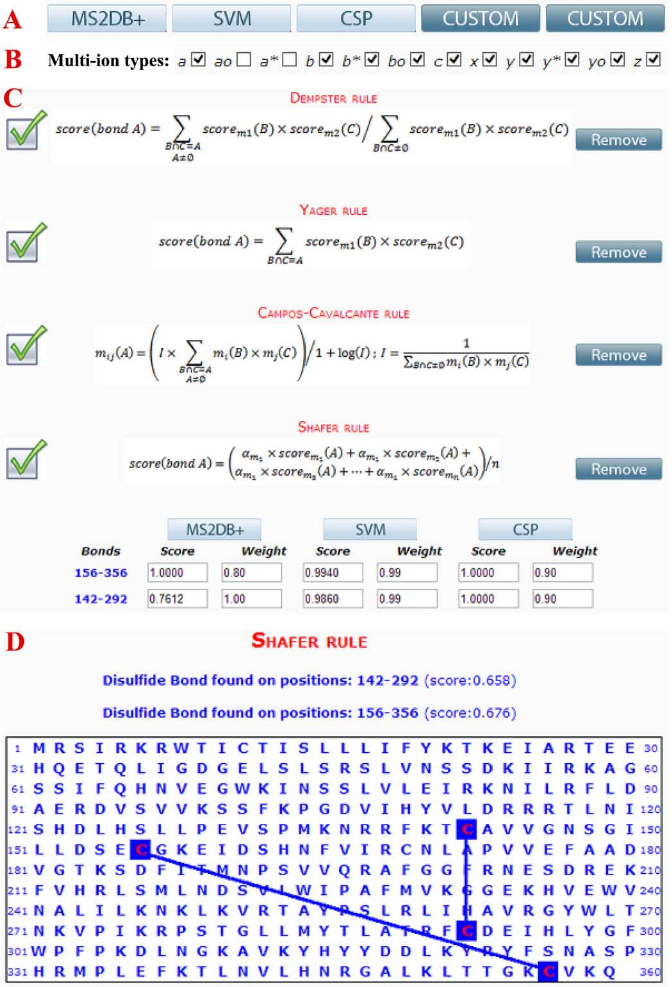
III. EVIDENCE COMBINATION IN MS2DB++

Dempster-Shafer theory (DST) is a mathematical theory of evidence which allows combination of information from multiple sources under conditions of epistemic uncertainty. In the following, we briefly describe our use of DST as an underpinning of MS2DB++, before describing the software itself.

Given the problem of finding S-S bonds for a protein P, in MS2DB++, a frame of discernment θ is used to exhaustively list all possible S-S bonds (called primitive hypotheses in DST) in P. The basic belief assignment function m assigns to each element $A \in 2^{\theta}$ a number in the range [0 1], corresponding to the measure of belief in the decision. For a given decision A, a belief interval I_A is defined using two evidential functions from DST: *belief* (lower bound of I_A) and *plausibility* (upper bound of I_A). The belief of A is the measure of how much the information given by a source supports a specific element to be the correct answer while the plausibility of A measures how much the information from a source does not contradict a hypothesis. Thus I_A represents the uncertainty associated with the decision A.

Different rules for evidence combination have been proposed in DST. In MS2DB++, four such rules are used: the Dempster rule, the Yager rule, the Campos rule, and the Shafer rule. These rules combine belief functions through their basic probability assignments and are defined in terms of combining two belief functions $m_1(\cdot)$ and $m_2(\cdot)$ in Table I, where α denotes a discount coefficient. Their extension to combining multiple belief functions is straightforward.

Of the combination rules proposed, the Dempster rule leads to normalized conjunctive pooling of evidence. The Yager rule, on the other hand, does not normalize out the conflicting evidence. The Campos rule de-rates the beliefs based on the conflicting evidences and assigns the remaining belief to the environment rather than to a common hypothesis. Finally, the Shafer rule applies a discounting function to each specific belief and then combines them by averaging. The process of evidence discounting is data driven and is done as follows: for the mass spectrometry-based approach, evidence is weighted based on precursor ion mass and abundance of the confirmatory matches.



Multi-ion types: a ☑ ao ☐ a* ☐ b ☑ b* ☑ bo ☑ c ☑ x ☑ y ☑ y* ☑ yo ☑ z ☑

☑ $$score(bond\ A)=\sum_{\substack{B\cap C=A\\ A\neq\emptyset}} score_{m1}(B)\times score_{m2}(C)\Big/\sum_{B\cap C\neq\emptyset} score_{m1}(B)\times score_{m2}(C)$$

☑ $$score(bond\ A)=\sum_{B\cap C=A} score_{m1}(B)\times score_{m2}(C)$$

☑ $$m_{ij}(A)=\left(I\times\sum_{\substack{B\cap C=A\\ A\neq\emptyset}} m_i(B)\times m_j(C)\right)\Big/1+\log(I)\,;\ I=\frac{1}{\sum_{B\cap C\neq\emptyset} m_i(B)\times m_j(C)}$$

☑ $$score(bond\ A)=\begin{pmatrix}\alpha_{m_1}\times score_{m_1}(A)+\alpha_{m_1}\times score_{m_2}(A)+\\ \alpha_{m_1}\times score_{m_3}(A)+\cdots+\alpha_{m_1}\times score_{m_n}(A)\end{pmatrix}\Big/n$$

	MS2DB+		SVM		CSP	
Bonds	Score	Weight	Score	Weight	Score	Weight
156-356	1.0000	0.80	0.9940	0.99	1.0000	0.90
142-292	0.7612	1.00	0.9860	0.99	1.0000	0.90

Disulfide Bond found on positions: 142-292 (score:0.658)

Disulfide Bond found on positions: 156-356 (score:0.676)

```
1   M R S I R K R W T I C T I S L L L I F Y K T K E I A R T E E 30
31  H Q E T Q L I G D G E L S L S R S L V N S S D K I I R K A G 60
61  S S I F Q H N V E G W K I N S S L V L E I R K N I L R F L D 90
91  A E R D V S V V K S S F K P G D V I H Y V L D R R R T L N I 120
121 S H D L H S L L P E V S P M K N R R F K T C A V V G N S G I 150
151 L L D S E C G K E I D S H N F V I R C N L A P V V E F A A D 180
181 V G T K S D F I T M N P S V V Q R A F G G F R N E S D R E K 210
211 F V H R L S M L N D S V L W I P A F M V K G G E K H V E W V 240
241 N A L I L K N K L K V R T A Y P S L R L I H A V R G Y W L T 270
271 N K V P I K R P S T G L L M Y T L A T R F C D E I H L Y G F 300
301 W P F P K D L N G K A V K Y H Y Y D D L K Y R Y F S N A S P 330
331 H R M P L E F K T L N V L H N R G A L K L T T G K C V K Q   360
```

Fig. 1. Different parts of the MS2DB++ interface showing: (A) the S-S connectivity determination methods available, (B) ion types that can be selected when using the mass spectrometry-based MS2DB+ method, (C) the four rules available for evidence combination, and (D) graphical display of the results showing the disulfide topology determined using the Shafer rule for the molecule [Q92187].

The weights for the SVM-based predictor are defined as a function of the belief scores. For the CSP-based method, the weights are defined in terms of the number of CSP matches and the divergence between the CSPs. Further details on computing the discounting functions can be found in [2].

In MS2DB++, the disulfide bond determination process starts with the selection by the user, of specific S-S bond determination methods that are to be used [Fig. 1(A)]. If the mass spectrometry-based method is chosen, then the user also needs to indicate the various ion types that should be considered [Fig. 1(B)]. Subsequently, belief scores are computed and assigned to each S-S bond (hypothesis) determined using each of the selected methods. At this point, a user can manually alter the belief assignments based on expert knowledge, if so desired. The power set of these hypotheses is then generated and the belief assignments are made. Next, the combination rules from Table I are selected [Fig. 1(C)]. In MS2DB++, these rules can be selected and applied independently and simultaneously. Finally, the globally consistent S-S connectivity for each combination strategy is determined using the optimization

TABLE II
COMPARISON OF RESULTS OBTAINED USING MS2DB++ WITH THOSE FROM ALTERNATIVE METHODS

[UniprotID]	Known Bonds	MS2DB++	MS2DB+	SVM	CSP	MassMatrix	DISULFIND	PreCys	DiANNA	DISLOCATE
[Q92187]	*142-292, 156-356*	All	All	All	All	All	None	None	142-292	None
[P02754]	*82-176,122-135*	All	82-176	122-135	122-135	82-176	None	None	82-176	82-176
[Q11130]	*68-176,211-214,318-321*	All	All	All	All	All	None	All	None	None
[P08037]	*134-176,247-266*	All	All	None	All	All	None	None	None	None
[Q09324]	*59-413,100-172,* *151-199,372-381*	All	59-413,151-199, 372-381	100-172, 372-381	151-199	None	None	None	None	None
[P00698]	*24-145, 48-133*	All	All	24-145	None	48-133	All	None	All	None
[P21217]	*81-338,91-341*	81-338	81-338	None	None	None	None	None	None	None

strategy from [3]. The reported S-S topology is based on the belief scores of each disulfide bond, the combination rule(s) selected, and the globally optimized S-S bond assignment. The global disulfide topology is displayed in graphical, structured (XML), and text formats for each of the selected combination rules [Fig. 1(D)].

IV. THE MS2DB++ WEBSERVER

MS2DB++ has been implemented using PHP and JavaScript. In it, sequence information is entered in FASTA format and a number of formats (e.g., mzXML, mzML, mzData or *Sequest* DTA files) are supported for entering mass spectrometry data. The S-S bonds determined by the webserver are presented both textually and graphically. These results can also be exported in TXT and XML formats. MS2DB++ supports interactive data analysis in that users can tune the weights assigned to different methods as well as select different combination rules without having to (re)start the entire analysis process.

V. COMPARISON WITH OTHER METHODS

In Table II, we present a comparison of the S-S bonds found using MS2DB++, using the Shafer rule, and eight other methods at the state-of-the-art (MS2DB+, SVM, CSP, MassMatrix, DISULFIND, PreCys, DiANNA, and DISLOCATE) on a set of glycosyltransferase molecules. As it can be seen, MS2DB++ outperformed all other methods, finding 16 out of 17 known disulfide bonds (listed on UniProt) for these molecules. On this dataset, MS2DB++ attained the following performance measures: sensitivity: 0.93, specificity: 0.99, accuracy: 0.99, and Mathews' correlation coefficient (MCC): 0.91.

Table III compares the results obtained by MS2DB++ (using three combination rules) and DISLOCATE [8], a predictive approach that is based on protein localization, on 20 different molecules. These molecules were randomly picked from the UniProt SP47 dataset, and excluded the molecules which were used to train the SVM and CSP sequence-based classifiers in MS2DB++. The performance measures show that MS2DB++ outperformed DISLOCATE using the Dempster, Campos, and Shafer rules of evidence combination. The Yager rule alone led to lower sensitivity and MCC scores. However, the specificity score obtained with the Yager rule was higher than those obtained with the Shafer, Dempster, and Campos rules.

TABLE III
COMPARISON BETWEEN MS2DB++ AND DISLOCATE

METHOD	Sensitivity	Specificity	Accuracy	MCC
DISLOCATE	0.58	0.92	0.84	0.54
MS2DB++ (Shafer, Dempster, and Campos rules)	0.63	0.92	0.85	0.57
MS2DB++ (Yager rule)	0.42	0.97	0.84	0.51

REFERENCES

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

[2] R. Singh and W. Murad, "Protein disulfide topology determination through the fusion of mass spectrometric analysis and sequence-based prediction using dempster-shafer theory," *BMC Bioinformatics*, vol. 14, no. Suppl. 2, p. S20, 2013.

[3] 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, pp. S1–S12, 2011.

[4] W. Murad and R. Singh, "MS2DB+: A software for determination of disulfide bonds using multi-ion analysis," *IEEE Trans. NanoBiosci.*, vol. 12, no. 2, pp. 69–71, 2013.

[5] B. J. Chen, C. H. Tsai, C. H. Chan, and C. Y. Kao, "Disulfide connectivity prediction with 70% accuracy using two-level models," *Proteins: Struct., Funct., Bioinformatics*, vol. 64, pp. 246–252, 2006.

[6] E. Zhao, H. L. Liu, C. H. Tsai, H. K. Tsai, C. H. Chan, and C. Y. Kao, "Cysteine separations profiles on protein sequences infer disulfide connectivity," *Bioinformatics*, vol. 21, pp. 1415–1420, 2005.

[7] C. C. Chuang, C. Y. Chen, J. M. Yang, P. C. Lyu, and J. K. Hwang, "Relationship between protein structures and disulfide-bonding patterns," *Proteins: Struct., Funct., Bioinformatics*, vol. 52, pp. 1–5, 2003.

[8] C. Savojardo, P. Fariselli, M. Alhamdoosh, P. L. Martelli, A. Pierleoni, and R. Casadio, "Improving the prediction of disulfide bonds in eukaryotes with machine learning methods and protein subcellular localization," *Bioinformatics*, vol. 27, no. 16, pp. 2224–2230, 2011.