An efficient algorithmic approach for mass spectrometry-based disulfide connectivity determination using multi-ion analysis

William Murad[1], Rahul Singh[1*], Ten-Yang Yen[2]

[1]Computer Science Department, [2]Chemistry and Biochemistry Department, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA, 94132, USA
*Corresponding author: rsingh@cs.sfsu.edu

Abstract

Background

Determining the disulfide (S-S) bond pattern in a protein is often crucial for understanding its structure and function. In recent research, mass spectrometry (MS) based analysis has been applied to this problem following protein digestion under both partial reduction and non-reduction conditions. However, this paradigm still awaits solutions to certain algorithmic problems fundamental amongst which is the efficient matching of an exponentially growing set of putative S-S bonded structural alternatives to the large amounts of experimental spectrometric data. Current methods circumvent this challenge primarily through simplifications, such as by assuming only the occurrence of certain ion-types (*b*-ions and *y*-ions) that predominate in the more popular dissociation methods, such as collision-induced dissociation (*CID*). Unfortunately, this can adversely impact the quality of results.

Method

We present an algorithmic approach to this problem that can, with high computational efficiency, analyze multiple ions types (a, b, b^o, b^*, c, x, y, y^o, y^*, and z) and deal with complex bonding topologies, such as inter/intra bonding involving more than two peptides. The proposed approach combines an approximation algorithm-based search formulation with data driven parameter estimation. This formulation considers only those regions of the search space where the correct solution resides with a high likelihood. Putative disulfide bonds thus obtained are finally combined in a globally consistent pattern to yield the overall disulfide bonding topology of the molecule. Additionally, each bond is associated with a confidence score, which aids in interpretation and assimilation of the results.

Results

The method was tested on nine different eukaryotic Glycosyltransferases possessing disulfide bonding topologies of varying complexity. Its performance was found to be characterized by high efficiency (in terms of time and the fraction of search space considered), sensitivity, specificity, and accuracy. The method was also compared with other techniques at the state-of-the-art. It was found to perform as well or better than the competing techniques. An implementation is available at: http://tintin.sfsu.edu/~whemurad/disulfidebond.

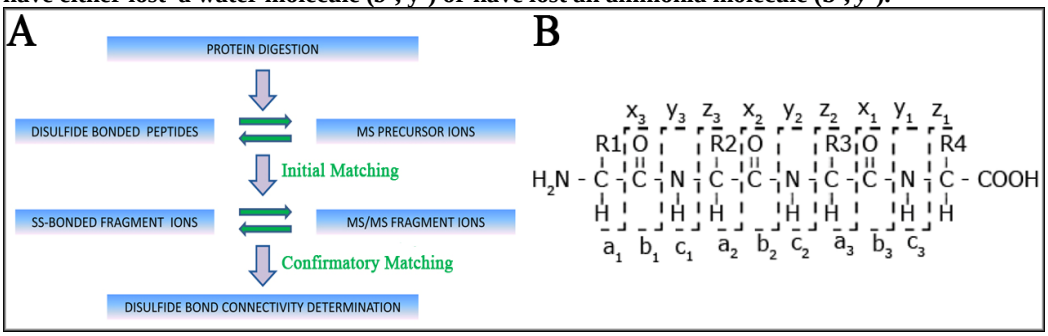
Conclusions

This research addresses some of the significant challenges in MS-based disulfide bond determination. To the best of our knowledge, this is the first algorithmic work that can consider multiple ion types in this problem setting while simultaneously ensuring polynomial time complexity and high accuracy of results.

Background

Disulfide (S-S) bonds are known to play an important role in protein structure and function. Among others, this includes: influencing protein folding and stabilization, formation of characteristic structural motifs such as the cysteine knot, mediation of thiol-disulfide interchange reactions, and regulation of enzymatic activity. Early computational approaches for S-S bond determination focused on two learning-driven formulations based on the protein primary structure[1]: *residue classification* (distinguish bonded and free cysteines) and *connectivity prediction* (determine the S-S connectivity pattern). In recent times, the increasing availability and accuracy of mass spectrometry (MS) has opened up an alternate approach; its essence lies in matching the theoretical spectra of ionized peptide fragments with experimentally obtained spectra to identify the presence of specific S-S bonds. A diagrammatic representation of the key steps of a MS-based approach is presented in Fig. 1, along with the different types of fragment ions that can be generated as an outcome of this process.

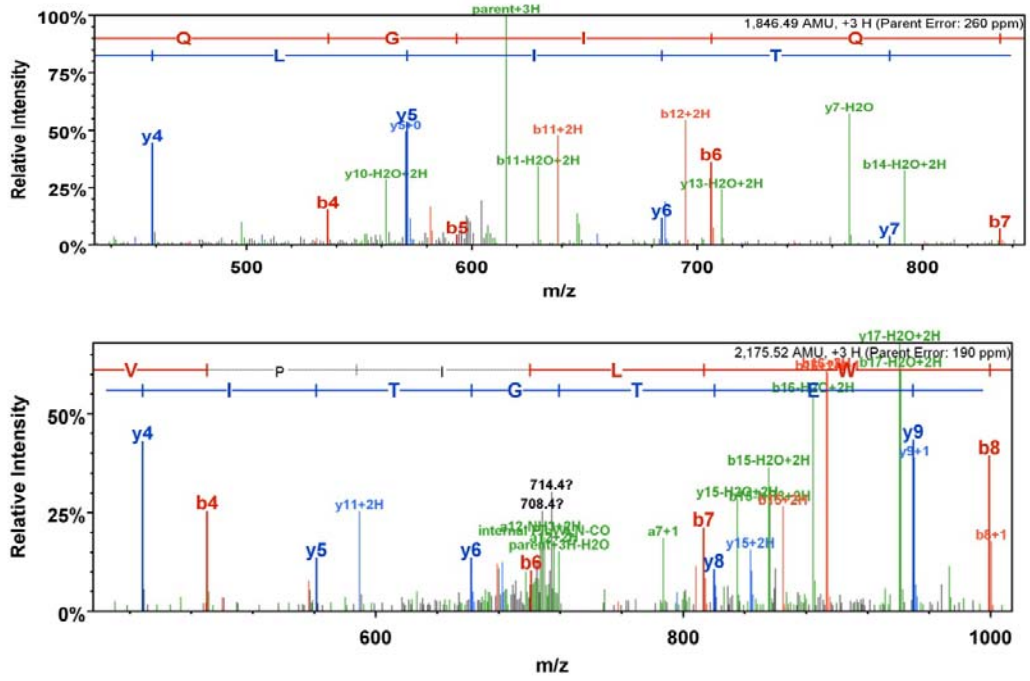
Figure 1 - (A) Once a protein is digested, the theoretically possible disulfide bonded peptides are compared with experimentally obtained precursor ions. In order to confirm each correspondence, the possible disulfide bonded fragment ions are next compared with experimentally generated MS/MS spectra. (B) Most of the different fragment ions (and their nomenclature) that can be observed. Ions types not represented here include *b* and *y* ions which have either lost a water molecule (b^o, y^o) or have lost an ammonia molecule (b^*, y^*).



MS-based methods generally outperform methods using sequence-based learning formulations, as showed by Lee and Singh[3]. However, a number of algorithmic challenges remain outstanding in realizing the potential of MS-based approaches. Salient among these are: (1) *accounting for multiple ion types in the data*[4]: To avoid an exponential increase in the search space, a common simplification is to limit the analysis to the spectra of *b*-ions and *y*-ions only[3,9,10]. However, this simplification may erroneously ignore the occurrence of other ions, such as: *a*, b^o, b^*, *c*, *x*, y^o, y^*, and *z*. While the occurrence of non-*b/y* ions is minimized (though not eliminated) in collision-induced dissociation (*CID*), some of these ions can be present with greater likelihood in dissociation methods such as electron capture dissociation (*ECD)*,

electron transfer dissociation (*ETD)*, and electron-detachment dissociation (*EDD)*. In fact these ions types should be considered even in *CID* as illustrated in Figure 2.

Figure 2 - Presence of multiple ions types (in green) after *CID*. In the first spectrum, note the presence of b^o and y^o ions with high intensity in the fragmentation of the precursor ion whose sequence is FFLQGIQLNTILPDAR, for the protein Lysozyme (UniProtKB ID P11279). In the second spectrum, *a*, b^o, b^*, and y^o ions (all with high intensity) can be observed after the fragmentation of a precursor ion existing in the protein Pratelet glycoprotein 4 (UniProtKB ID P16671).



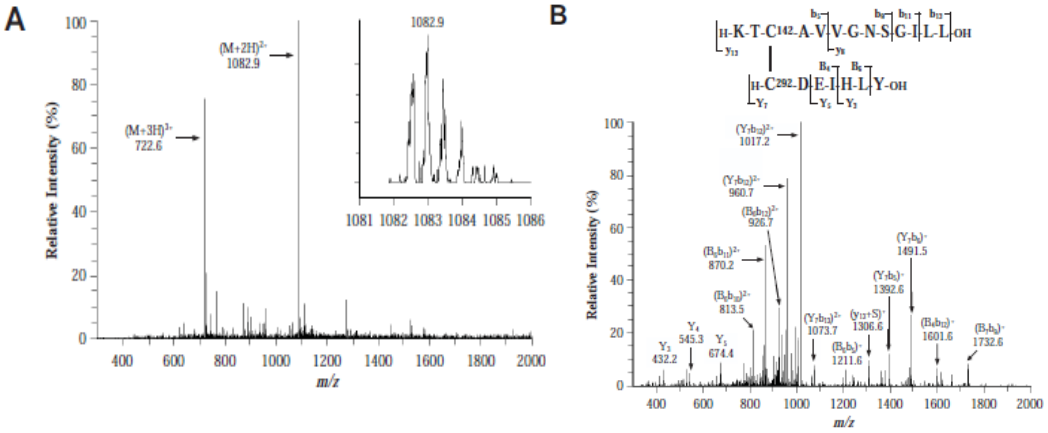
(2) *Design of efficient search and matching algorithms*: The search space of possible disulfide topologies increases rapidly not only with the number of ion types being analyzed but also with the number of cysteines as well as the types of connectivity patterns. Thus, it is imperative to have algorithms that can accommodate the richness of the entire problem domain. (3) *Automated data-driven determination of parameters:* Many advanced algorithms are intrinsically parametric. Often, determining the optimal value of these parameters automatically is in itself, a complex problem. This places the practitioner at a significant disadvantage. Support for automated and data-driven strategies for estimation of crucial parameters is therefore crucial to the real-world success of a method in this problem domain.

The contributions of this paper in context of the aforementioned challenges include: (1) Development of a highly efficient strategy for multi-ion disulfide bond analysis by considering *a*, *b*, b^o, b^*, *c*, *x*, *y*, y^o, y^*, and *z* ion types. To the best of our knowledge, this is the first algorithmic work that has considered all these ion-types in S-S bond determination. (2) A fully polynomial-time algorithm that selectively

generates only those regions of the search space where the correct solutions reside with a high likelihood. (3) A multiple-regression-based data driven method to calculate the critical parameters modulating the search, so as to ensure that the correct bonding topologies are not missed due to the truncation of the search space. At the same time, the parameter selection ensures that the search is focused on the most promising regions of the search-space. (4) A local-to-global strategy that builds a globally consistent bonding pattern based on MS data at the level of individual bonds, and (5) Assignment of probability-based scores[23] to each specific disulfide bond based on the number of MS/MS matches and their respective abundance. These scores represent an assessment of quality and reflect the significance of the disulfide bond.

At a high-level, the proposed approach can be thought of as a two-stage database-based matching technique (see Fig. 3). During the first stage, the mass values of the theoretically possible disulfide-bonded peptide structures are compared with precursor ion mass values derived from the MS-spectra. In the second (confirmatory) stage, the theoretical spectra from the disulfide-bonded peptide structures are compared with MS/MS experimental spectra. The confirmatory step is necessary since a disulfide bonded peptide may not actually correspond to a precursor ion, even though their mass values are similar. Our approach allows us to conduct this entire search process in (a low degree) polynomial time. This paper significantly extends our prior research where we had proposed efficient indexing strategies to speed-up the search[11, 12] as well as our more recent work[6], where a polynomial time approximation algorithm using hand-crafted parameters was proposed for the first stage matching.

Figure 3 - Two-stage matching spectra for protein ST8SiaIV. (A) In the first-stage (*DMS* vs. *PMS*), the theoretical disulfide-bonded structure is matched with the doubly charged precursor ion with highest intensity, whose *m/z* = 1082.9. (B) For this initial match, the disulfide-bonded peptide pair is fragmented and the fragments are matched with the MS/MS spectrum for the precursor ion (*FMS* vs. *TMS*), generating a list of validation matches.



Methods

We start the description of our method by providing, in Table 1, the key abbreviations used in the ensuing description and their respective definitions.

Table 1 - Abbreviations used and their respective definitions.

Abbreviation	Definition
DMS	Set of mass values corresponding to all possible disulfide-bonded peptide structures that can be obtained from a digested protein.
PMS	Set of mass values of ions that undergo dissociation to produce product ions (set of precursor ions).
IM	Correspondence obtained when the difference between the detected mass of a targeted ion from the *PMS* and the calculated mass of a possible disulfide-bonded peptide structure from the *DMS* is less than a match threshold T_{IM}.
T_{IM}	Initial Match threshold. Threshold used to define a mass window centered on a *PMS* value within which a correspondence between a *DMS* value and a *PMS* value may be found.
ε	*DMS* trimming parameter used to trim the *DMS* set. To trim the *DMS* set by ε means to remove as many elements from *DMS* as possible without losing meaningful mass values.
TrimSet	Set of trimmed mass values from the *DMS* set.
PM	Peptide Mass: cysteine-containing peptide mass value.
TempSet	Temporary mass set containing possible disulfide bonded peptide structures.
FMS	Set of mass values of every disulfide-bonded fragment structure that can be obtained from fragment ions, which can be of types *a*, *b*, b^o, b^*, *c*, *x*, *y*, y^o, y^* and *z*.
TMS	Set of mass values of the product ions obtained after the MS/MS step (MS/MS spectra).
VM	Correspondence obtained when the difference between a precursor ion fragment mass from *TMS* and a disulfide-bonded fragment structure mass from *FMS* falls below a validation match threshold T_{VM}.
T_{VM}	Validation Match threshold. Threshold used to define a mass window centered at a *TMS* value in which a correspondence between a *FMS* value and a *TMS* value may be found.
δ	*FMS* trimming parameter used to trim the *FMS* set. To trim the *DMS* set by δ means to remove as many elements from *FMS* as possible without losing meaningful fragment ions mass values.
FragSet	Set containing the mass values of fragment ions generated by the method GENFRAGS(.) in the APROX-FMS routine.

In the first stage of the method, an *Initial Match* (*IM)* is obtained when the difference between the detected mass of a targeted ion from the *PMS* and the calculated mass of a possible disulfide-bonded peptide structure from the *DMS* is found to be less than a threshold T_{IM}. The second stage validates (or rejects) the initial matches found. For each Initial Match, the validation occurs by searching for matches between product ions from the *TMS* and the theoretical spectra *FMS*. A *Validation Match* (*VM*) is obtained when the difference between a precursor ion fragment mass from *TMS* and a disulfide-bonded fragment structure mass from *FMS* is below a validation match threshold T_{VM}.

Unfortunately, the sizes of both *FMS* and *DMS* grow exponentially. For a disulfide-bonded peptide structure consisting of *k* peptides, considering that there are *f* different fragment ion types possible, up to f^k types of fragment arrangements may occur in the *FMS*. If each fragment ion consists of $\|p_i\|$ amino acid residues, then the complexity to compute the entire *FMS* for a disulfide-bonded peptide structure is $O\left(f^k \times \prod \|p_i\|\right)$ using a brute-force approach. The *DMS* also grows exponentially. To

understand this, let $P = \{p_1, p_2, \ldots, p_k\}$ be the list of cysteine-containing peptides in a polypeptide chain. Further, let $C = \{c_1, c_2, \ldots, c_i\}$ be the list of the number of cysteines per cysteine-containing peptide p_i. If $n = \sum_1^i c_i$ is the total number of cysteines in a protein, the number of possible disulfide connectivity patterns (*DMS* size) is:[1,14]

$$(n-1)!! = \prod_{i=1}^{n/2} (2i-1).$$

The subset-sum formulation: Towards polynomial-time matching

Given the growth characteristics of the *DMS* and the *FMS*, an exhaustive search-and-match strategy is clearly infeasible in the general case. This is especially true in practice if multiple ion types are considered. Our approach towards designing an efficient algorithm for this problem is based on the key insight that the *entire search space (DMS or FMS) does not need to be generated to determine the matches*. That is, we only want to generate the few disulfide bonded peptides whose mass is close to the (given) experimental spectra rather than generate all possible peptide combinations and subsequently testing and discarding most of these. This insight allows us to re-cast the *DMS* and *FMS* generation as instances of the subset-sum problem[5]. Recall, that given the pair (S, t), where S is a set of positive integers and $t \in \mathbf{Z}^+$, the subset-sum problem asks whether there exists a subset of S that adds up to t. While the subset-sum problem is itself NP-Complete, it can be solved using approximation strategies to obtain near-optimal solutions, in polynomial-time.

Polynomial time DMS mass list construction

Our strategy lies in obtaining an approximate solution to the subset-sum problem by trimming as many elements from *DMS* as possible based on a parameter ε. To trim the *DMS* set by ε means to remove as many elements from *DMS* as possible such that if DMS^* is the resultant trimmed set, then for every element DMS_i removed from *DMS*, there will remain an element DMS^*_i in DMS^* which is "sufficiently" close in terms of its mass to the deleted element DMS_i. Specifically,

$$(DMS_i / 1+\varepsilon) \le DMS_i^* \le DMS_i \qquad (1)$$

The approximation algorithm for creating the partial *DMS* is described by the APPROX-DMS and TRIM routines (Fig. 4). APPROX-DMS takes the following parameters: (1) a list of cysteine-containing peptides mass values, (2) a value from the *PMS* list, (3) the trimming parameter ε, and (4) the Initial Match threshold T_{IM}.

Figure 4 - Pseudo code for APROX-DMS and TRIM routines.

```
01. APPROX-DMS (CCP, PMS_val, ε , T_IM )
02.     DMS_0 ← {0}
03.     IM ← { }
04.     TrimSet ← { }
05.     for i ← 0 to (|| CCP || – 1)
06.         PM ← CCP_i
07.         TempSet ← { }
08.         DMSsize ← || DMS ||
09.         for j ← 0 to (DMSsize – 1)
10.             if ((PM + DMS_j) ≤ (PMS_val + T_IM))
11.                 TempSet_j ← PM + DMS_j
12.                 if ((PM + DMS_j) ≥ (PMS_val - T_IM))
13.                     IM ← PM + DMS_j
14.         DMS ← MERGE-SORT (DMS, TempSet)
15.         Lists ← TRIM (DMS, TrimSet, ε )
16.         DMS ← Lists["DMS"]
17.         TrimSet ← Lists["TrimSet"]
18.     return {DMS, TrimSet}
19. TRIM (DMS, TrimSet, ε )
20.     n ← | DMS |
21.     max_value ← DMS_n-1
22.     last ← max_value
23.     for i ← n-2 to 0
24.         if last > (( 1+ ε ) × DMS_i )
25.             DMS* ← DMS_i
26.             last ← DMS_i
27.         else
28.             TrimSet ← DMS_i
29.     sort (DMS*)
30.     DMS* ← max_value
31.     return {DMS*, TrimSet }
```

In lines 2-8 of Fig. 4, all the variables and data structures are initialized. In lines 9-13, the theoretical disulfide-bonded peptide structures are formed, stored in a temporary set *TempSet*, and the Initial Matches are computed. Lines 14-17 increments the *DMS* by invoking the routine MERGE-SORT, which returns a sorted set that is the merge of its two sorted input sets *DMS* and *TempSet*, with duplicated values removed. This is also where the TRIM routine is called to shorten the *DMS* set.

Table 2 presents an example showing the effectiveness of the APROX-DMS. In this specific case, 16.7% of the entire search space (all feasible combinations of cysteine-containing peptides) was successfully trimmed, while ensuring that the correct *IM* was not missed. Another example illustrating the action of APPROX-DMS on the Beta-LG protein is available as supplemental information at: http://tintin.sfsu.edu/~whemurad/disulfidebond/papers/suptable1.pdf.

Table 2 – Running APROX-DMS on the ST8SiaIV C^{142}-C^{292} bond. *CCP:* the mass values of all cysteine-containing peptides. PMS_{val} : a disulfide-bonded precursor ion mass. *TrimSet*: all the disulfide-bonded structures trimmed from the set of feasible combinations of cysteine-containing peptides. For this example, 16.7% of the structures were trimmed and the correct *IM* was found.

Property	Value
CCP	{716, 728, 749, 863, 864, 891, 976, 1096, 1105, 1161, 1204, 1274, 1359, 1367, 1418, 1480, 1593, 1733, 1754, 1846, 1863, 1864, 1976, 2179, 2292, 2351, 2617, 2737, 2822}
PMS_{val}	{2640} *(Precursor $M+H^+$* mass and charge state*: 2638.121 3)*
ε	0.02530
T_{IM}	1.0
TrimSet	*{716, 863, 1096, 1443, 1589, 1590, 1611, 1725, 1832, 1838, 1844, 1853, 1866, 1888, 1909, 1958, 1995, 2001, 2022, 2051, 2066, 2070, 2086, 2094, 2107, 2135, 2164, 2178, 2221, 2248, 2249, 2307, 2333, 2363, 2462, 2519}*
IM	{2640} *(KTCAVVGNSGILL – ATRFCDEIHLY) – SS-bond: C^{142}-C^{292}*

The complexity of both routines MERGE-SORT and TRIM, based on the comparison-based sorting algorithm merge sort, is

$O(|DMS|+|TempSet|\times lg(|DMS|+|TempSet|))$. Further, for any fixed $\varepsilon > 0$, our algorithm is a $(1+\varepsilon)$-approximation scheme. That is, for any fixed $\varepsilon > 0$, the algorithm runs in polynomial time. The proof of the polynomial time complexity of APPROX-DMS can be obtained by direct analogy to the proof of the polynomial time complexity of the subset sum approximation algorithm from[5].

Parameters Estimation

APPROX-DMS depends on two important parameters, namely, the match threshold T_{IM} and the trimming parameter ε. The match threshold is responsible for defining a "matching window". This is necessary due to practical constraints such as the sensitivity of the instrument (i.e. 0.01Da, 0.1Da, and 1.0Da) and experimental noise, which would typically ensure that an exact match would rarely occur. We conducted an empirical study by using different values of T_{IM} for all our datasets. Based on the results, the T_{IM} value of $\pm 1.0Da$ was found to minimize missing matches as well as the occurrence of false positives. Considering the smallest precursor ion mass involved, in these studies, the above value of T_{IM} guarantees a matching accuracy of 99.86%.

The second parameter ε is much more important as it is crucial to the running time of the algorithm and its accuracy as evident from Eq. 1. To determine ε, we note that it is inversely proportional to the algorithm's running time. However, a large value of ε would cause meaningful fragments to be left out of the *DMS*. At the same time, a small value for ε will lead to few data points being trimmed. Thus "guessing" good values of ε can be complicated and suboptimal choices can significantly impact the quality of the results. We address the problem of data-driven estimation of ε using a regression framework where ε is treated as a dependent variable and based on the data, a functional relationship is obtained between it and the other (independent) variables. We model this functional relationship using the following independent variables: (1) the cysteine-containing peptides (*CCP*) mass range defined by CCP_{max} and CCP_{min} corresponding to the peptides with highest and lowest mass respectively. (2) The number of cysteine-containing peptides *k*. A large *k* implies that the average difference in the mass of any two peptide fragments is small. Conversely, a small *k* implies fewer fragments with putatively larger differences in their masses. (3) The cysteine-containing peptides average mass value $CCP_{average}$. The relationship between ε and these other variables is then obtained using multiple-variable

regression. The data for the regression was obtained using a leave-*n*-out cross-validation setting where, groups of four proteins were randomly picked from the set of 9 proteins available to us. The functional relationship defining ε was obtained to be:

$$\varepsilon = 1.3939\times10^{-2} \times \frac{(CCP_{\max} - CCP_{\min})}{CCP_{average}} - 1.0824\times10^{-3} \times k + 3.9094\times10^{-2} \qquad (2)$$

Figure 5 - Pseudo code for APROX-FMS routine.

```
01. APPROX-FMS (peptides, TMS_val , δ , T_VM )
02.     FMS_0 ← {0}
03.     VM ← { }
04.     for i ← 0 to (|| peptides || - 1)
05.         TempSet ← { }
06.         Pep_sequences ← peptides_i
07.         FragSet ← GENFRAGS (Pep_sequences)
08.         for j ← 0 to (|| FMS || - 1)
09.             for k ← 0 to (||FragSet|| - 1)
10.                 if ((FragSet_k + FMS_j) ≤ (TMS_val + T_VM))
11.                     TempSet[ ] ← FragSet_k + FMS_j
12.                 if ((FragSet_k + FMS_j) ≥ (TMS_val − T_VM))
13.                     VM ← FragSet_k + FMS_j
14.         FMS ← MERGE-SORT (FMS, TempSet)
15.         FMS ← TRIM (FMS, δ )
16.     return { FMS, VM }
```

Polynomial time FMS construction

In creating the *FMS*, a strategy similar to the one used for generating the *DMS* is implemented. This involves using an approximation algorithm, this time, to generate the theoretical spectra for all the *IMs* found during the first-stage matching. We define another trimming parameter δ to trim the *FMS* mass list. It can be expected that the functional form of δ depends on the fragments mass range, as well as their granularity (extent to which fragments are broken down into smaller ions). In a manner similar to the case for estimating ε, we use regression to obtain the specific functional form for the dependent variable δ in terms of the variables AA_{max} (the largest amino acid residue mass), AA_{min} (the smallest amino acid residue mass), $AA_{average}$ (the average amino acid residues mass), and $||p||$ (average number of amino acid residues per fragment). Leave-*n*-out cross-validation was again used, yielding the relationship shown in Eq. 3.

$$\delta = 6.1744\times10^{-3} \times \frac{(AA_{\max} - AA_{\min})}{AA_{average}} - 3.0936\times10^{-3} \times \|p\| + 5.0731\times10^{-2} \qquad (3)$$

The pseudocode of the APPROX-FMS procedure used for generating the *FMS* is shown in Figure 5. The function/procedure GENFRAGS(.), in line 7, generates multiple fragment ions (*a*, *b*, b^o, b^*, *c*, *x*, *y*, y^o, y^*, and *z*) for peptide sequences in $Pep_{sequences}$, which contains the disulfide-bonded peptides involved in the *IM* being

analyzed. Next, for each element in the *FMS* and for each fragment in the *FragSet* (lines 8-13), new disulfide-bonded peptide fragment structures are formed. A Validation Match *VM* is declared when a correspondence is found between the mass of a generated structure and an experimentally determined mass value TMS_{val}. Once all *FMS* and *FragSet* values are processed, the current *FMS* set is combined with the disulfide-bonded peptide fragments set *TempSet* using MERGE-SORT (line 14). Lastly, *FMS* is trimmed using the TRIM routine.

Determining the globally consistent bond topology

Once all the *Initial Matches* and *Validation Matches* are calculated, we have a "local" (putative bond-level) view of the possible disulfide connectivity. This local information needs to be integrated to obtain a globally consistent view. Our approach to this problem is motivated by Fariselli and Casadio[14]. Specifically, we model the location of the putative disulfide bonds by edges in an undirected graph $G(V, E)$, where the set of vertices V corresponds to the set of cysteines. To each edge, we assign a match score. This score represents the combined importance of each single peak match within two spectra. Each specific peak match is weighted according to its intensity. The match score is given by:

$$V_S = \left(\sum_{i=1}^{n} (VM_i \times I_N) / \sum_{i=1}^{n} (TMS_i \times I_N) \right) \times 100 \qquad (4)$$

In Eq. (4), the numerator corresponds to the sum of each validation match for a disulfide bond multiplied by the matched MS/MS fragment normalized intensity value (I_N). Here, VM_i is a binary value which is set to 1 if a confirmatory match was found for fragment i. The denominator similarly contains the sum of each experimental MS/MS fragment ion from *TMS* multiplied by I_N. Here, TMS_i is a binary variable which indicates the presence of a fragment i in the MS/MS spectrum.

Results

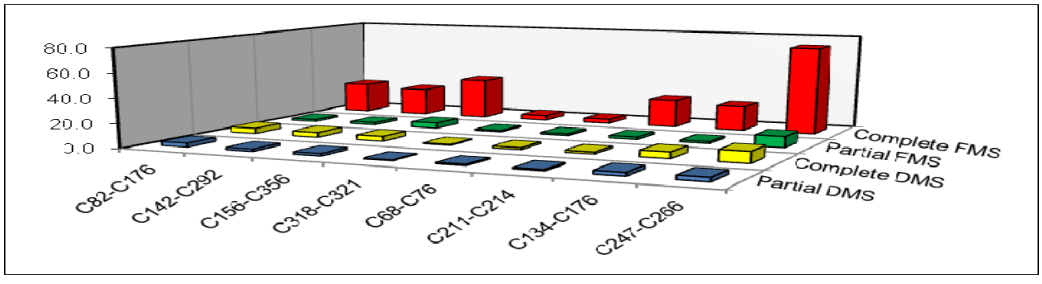
The proposed method was validated utilizing experimental data obtained using a capillary liquid chromatography system coupled with a Thermo-Fisher LCQ ion trap mass spectrometer LC/ESI-MS/MS system. Details of the experimental protocols can be found in [8,19]. We used data from nine eukaryotic glycosyltransferases. These molecules (and their UniProtKB ID) were: ST8Sia IV (Q92187), Beta-lactoglobulin (P02754), FucT VII (Q11130), C2GnT-I (Q09324), Lysozyme (P00698), FT III (P21217), β1-4GalT (P08037), Aldolase (P00883), and Aspa (Q9R1T5).

We conducted five sets of experiments to investigate the proposed method and its efficacy. These experiments included: (1) Analysis of method's efficiency, showing how the method successfully reduced the *DMS* and *FMS* search spaces. (2) Analysis of the effect of incorporating multiple ion types, demonstrating the importance of considering non-*b*/*y* ions in the determination of disulfide bonds. (3) Comparative analysis of the proposed method with established predictive techniques. (4) Comparative analysis of the method with MassMatrix, an established MS-based approach which can be used for determining S-S bonds. In both experiment 3 and experiment 4, the aforementioned set of glycosyltransferases and their known S-S bond topology provided us with the ground truth. (5) Analysis of the method in terms of established performance measures: *Accuracy* (Q_2), *Sensitivity* (Q_c), *Specificity* (Q_{nc}), and *Matthew's correlation coefficient (c)*.

Table 3 - *DMS* and *FMS* mass space sizes comparison.

Protein	Disulfide Bond	Full Search (exponential)		Proposed Search (polynomial)		DMS decrease	FMS decrease
		DMS size	FMS size	DMS size	FMS size		
Beta-LG	$C^{82}C^{176}$	2152	2169	1870	78	13.1%	96.4%
ST8Sia IV	$C^{142}C^{292}$	1246	1792	1038	106	16.7%	94.1%
	$C^{156}C^{356}$	1246	2640	1038	255	16.7%	90.3%
FucT VII	$C^{318}C^{321}$	581	115	528	34	9.1%	70.4%
	$C^{68}C^{76}$	879	103	681	41	22.5%	60.2%
	$C^{211}C^{214}$	879	1819	681	107	22.5%	94.1%
B1,4-GalT	$C^{134}C^{176}$	2149	1189	1127	77	47.6%	93.5%
	$C^{247}C^{266}$	2149	5480	1127	426	47.6%	92.2%
Average DMS and FMS decrease						21.8%	86.4%

Figure 6 – Comparison of the computational time (in seconds) for the exhaustive and partial generation of *DMS* and *FMS* of the proteins from Table 3. On average there was a 49.5% decrease in time to compute the *DMS* and 88.7% decrease in time to compute the *FMS*. The computations were carried out on an Intel T2390 1.86 GHz single-core processor with 1GB RAM.



Analysis of efficiency of the search

One of the most important characteristics of the proposed method is its efficiency in terms of excluding significant portions of a large and rapidly expanding search space. In Table 3 we compare the size of the complete *DMS* (containing all the disulfide-bonded peptide structures generated for each protein) and the complete *FMS*

(containing all the disulfide-bonded fragment ions) with the truncated *DMS* and *FMS* obtained using the proposed approach.

It may be noted that across the molecules, on an average, the proposed approach required examining about 78% of the entire *DMS* and only about 14% of the entire FMS. It is crucial to note that this reduction in search was achieved without impacting the accuracy *and* having considered all multiple fragment ion types (a, b, b^o, b^*, c, x, y, y^o, y^*, and z). The *DMS* decrease was less than the *FMS decrease* because the disulfide-bonded structures in the *DMS* were bigger and fewer in number and consequently dispersed across the spectra mass range. In Figure 6, we show the actual time taken to obtain a solution by generating the complete *DMS* and *FMS*, as well as their truncated counterparts, for each of the molecules.

Effects of incorporating multiple ion types: A case study

In this experiment, we investigated the effect of incorporating multiple ion types (a, b, b^o, b^*, c, x, y, y^o, y^*, and z) in determining the S-S bonds as opposed to considering only b/y-ions. We found that multiple instances of combinations between b/y ions and other ions types occurred by analyzing the confirmatory matches for the different disulfide bonds. Supplemental information showing these combinations is available at: http://tintin.sfsu.edu/~whemurad/disulfidebond/papers/suptable2.pdf

The consideration of multiple ion types also contributed to the method's accuracy in terms of determining specific S-S bonds. Disulfide bonds previously missed due to their low match score could be identified when all ten different ion types were considered. The tryptic-digested protein FucT VII (which underwent *CID*) constituted one such example. In FucT VII the bond C^{318}-C^{321} was missed when considering only b/y ions (match score 29, *pp*=11, *pp2* =15). However, as shown in Fig. 7, this bond was identified when multiple ions types were included (match score 100, *pp*=31, *pp*2=70). The confidence measures *pp* and *pp2* are described in the following section. To explain this improvement we note that C^{318}-C^{321} was an intra-bond involving cysteines that were close together. Consequently, CID-based fragmentation was poor and the consideration of other ion types essentially improved the signal-to-background contrast. In this particular case, five other ion types - a_4, a_5, a_6, b^o_7, y^*_7 - were present in the FucT VII MS/MS data besides the b ions represented in the spectrum on the right in Fig. 7. In the following, we present details of how these ions contribute to the match score V_s (from Eq. 4). We present the two cases: consideration

of only *b/y*-ions (Eq. (5)) and consideration of multiple ion types (Eq.(6)) In the numerator we specify the contribution of each spectrum peak from Fig. 7 (the ion corresponding to each $VM_i \times I_N$ term is showed in brackets).

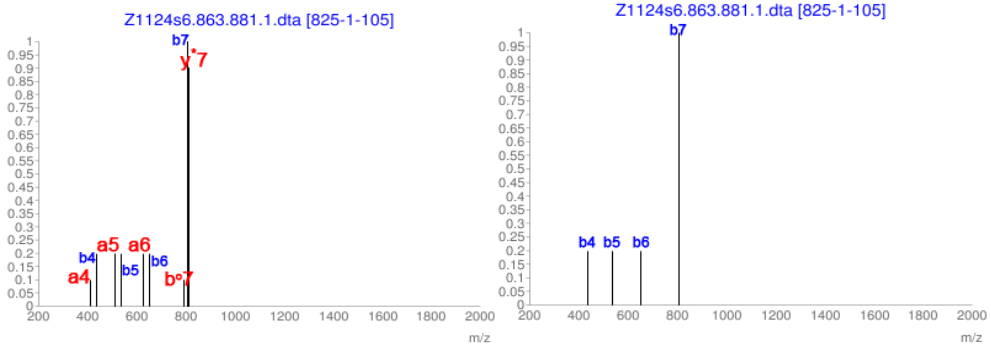
$$V_S = \left(\frac{0.1784[b_4] + 0.1732[b_5] + 0.1574[b_6] + 1.4324[b_7]}{6.7321} \right) \times 100 = \left(\frac{1.9414}{6.7321} \right) \times 100 = 28.8 \qquad (5)$$

$$V_S = \left(\frac{0.1784[b_4] + 0.1732[b_5] + 0.1574[b_6] + 1.4324[b_7] + 0.1863[a_4] + 0.6787[a_5] + 0.1996[a_6] + 0.1464[b^o{}_7] + 3.5797[y^*{}_7]}{6.7321} \right) \times 100$$

$$= (6.7321/6.7321) \times 100 = 100.0 \qquad (6)$$

We also observed that consideration of multiple ion-types led to significant increase in the match scores of the true disulfide bonds, whereas only a modest increase was noticed for false positives. This allowed us to increase the threshold we use on the match score V_s to identify high-quality matches from 30 to 80 (a 166% increase). The positive effect of this increment on the specificity of the method can be illustrated by considering the protein Aldolase. In this molecule, consideration of only *b*/*y* ions led to a false positive S-S bond identification between cysteines C^{135}-C^{202} (V_s=30.8, with (original) threshold 30) However, when the multiple ions-types were considered with the (increased) threshold on the match score, no S-S bond was found between C^{135}-C^{202} (V_s= 53.2, (incremented) threshold 80).

Figure 7 – Spectra (m/z vs. normalized intensity) illustrating the confirmatory matches (whose intensity values were at least 10% of the maximum intensity) found for the disulfide bond between cysteines C^{318}-C^{321} in protein FucT VII. The spectrum in the left shows the matches found when multiple ions were considered. The spectrum in the right shows the matches when only *b*/*y*-ions were considered.



Comparative studies with predictive techniques

In this experiment we compared the proposed method with three well known predictive methods DiANNA[20], DISULFIND[21], and PreCys[22]. The results from each of the methods are shown in Table 4 along with the with the known disulfide bond linkages according to the UniProtKB knowledgebase. As it can be seen, in terms of correct identifications (as well as minimizing false positives), the proposed approach outperformed all the predictive techniques.

Table 4 – Comparison with predictive methods.

Protein	Known Pattern	Proposed Algorithm	DiANNA 1.1	DISULFIND	PreCys
ST8Sia IV	$C^{142}C^{292}$, $C^{156}C^{356}$,	**$C^{142}C^{292}$, $C^{156}C^{356}$**	$C^{11}C^{156}$,$C^{142}C^{292}$,$C^{169}C^{356}$	None	$C^{142}C^{356}$, $C^{156}C^{292}$
Beta-LG	$C^{82}C^{176}$, $C^{122}C^{135}$	**$C^{82}C^{176}$**	$C^{12}C^{137}$, $C^{82}C^{176}$, $C^{126}C^{135}$	None	None
FucT VII	$C^{68}C^{76}$, $C^{211}C^{214}$, $C^{318}C^{321}$	**$C^{68}C^{76}$, $C^{211}C^{214}$, $C^{318}C^{321}$**	$C^{68}C^{321}$, $C^{76}C^{211}$, $C^{214}C^{318}$	$C^{76}C^{318}$	$C^{68}C^{76}$, $C^{211}C^{214}$, $C^{318}C^{321}$
B1,4-GalT	$C^{134}C^{176}$, $C^{247}C^{266}$	**$C^{134}C^{176}$, $C^{247}C^{266}$**	$C^{23}C^{176}$, $C^{30}C^{144}$, $C^{266}C^{341}$	None	$C^{134}C^{247}$, $C^{176}C^{266}$
C2GnT-I	$C^{59}C^{413}$, $C^{100}C^{172}$, $C^{151}C^{199}$, $C^{372}C^{381}$	**$C^{59}C^{413}$, $C^{151}C^{199}$, $C^{372}C^{381}$**	$C^{13}C^{172}$, $C^{59}C^{217}$, $C^{151}C^{234}$, $C^{199}C^{372}$, $C^{381}C^{413}$	Not supported	$C^{59}C^{381}$, $C^{100}C^{372}$, $C^{151}C^{172}$, $C^{199}C^{413}$
Lysozyme	$C^{24}C^{145}$, $C^{48}C^{133}$	**$C^{24}C^{145}$, $C^{48}C^{133}$,**	$C^{24}C^{145}$, $C^{48}C^{133}$, $C^{82}C^{98}$, $C^{94}C^{112}$	$C^{24}C^{145}$, $C^{48}C^{133}$, $C^{82}C^{98}$, $C^{94}C^{112}$	$C^{82}C^{145}$
FT III	$C^{81}C^{338}$, $C^{91}C^{341}$	**None**	$C^{16}C^{91}$, $C^{81}C^{143}$, $C^{129}C^{338}$	None	$C^{81}C^{91}$
Aldolase	None	**None**	$C^{73}C^{339}$, $C^{135}C^{290}$, $C^{115}C^{240}$, $C^{178}C^{202}$	None	None
Aspa	None	**None**	$C^{4}C^{275}$, $C^{60}C^{217}$, $C^{66}C^{151}$, $C^{123}C^{145}$	None	None

Table 5 – Comparison with MassMatrix. The score (V_s) of each disulfide bond and the confidence scores (*pp* and *pp2* values) are shown in brackets, respectively.

Protein	Known Pattern	Proposed Method	MassMatrix
ST8Sia IV	$C^{142}C^{292}$, $C^{156}C^{356}$	**$C^{142}C^{292}$**[V_s*:131;pp:109;pp2:41*], **$C^{156}C^{356}$**[V_s*:100;pp:97;pp2:6*]	$C^{142}C^{292}$[V^*_s*:54;pp:15;pp2:13*], $C^{156}C^{356}$[V^*_s*:77;pp:23;pp2:15*]
Beta-LG	$C^{82}C^{176}$, $C^{122}C^{135}$	**$C^{82}C^{176}$**[V_s*:100;pp:49;pp2:16*]	$C^{82}C^{176}$[V^*_s*:68;pp:14;pp2:14*]
FucT VII	$C^{68}C^{76}$, $C^{211}C^{214}$, $C^{318}C^{321}$	**$C^{68}C^{76}$**[V_s*:105;pp:41;pp2:98*], **$C^{211}C^{214}$**[V_s*:100;pp:13;pp2:20*], **$C^{318}C^{321}$**[V_s*:100;pp:31;pp2:70*]	$C^{68}C^{76}$[V^*_s*:12;pp:9;pp2:3*], $C^{211}C^{214}$[V^*_s*:78;pp:16;pp2:11*], $C^{318}C^{321}$[V^*_s*:46;pp:28;pp2:16*]
B1,4-GalT	$C^{134}C^{176}$, $C^{247}C^{266}$	**$C^{134}C^{176}$**[V_s*:100;pp:61;pp2:29*], **$C^{247}C^{266}$**[V_s*:195;pp:88;pp2:177*]	$C^{134}C^{176}$[V^*_s*:34;pp:9;pp2:7*], $C^{247}C^{266}$[V^*_s*:31;pp:7;pp2:7*]
C2GnT-I	$C^{59}C^{413}$, $C^{100}C^{172}$, $C^{151}C^{199}$, $C^{372}C^{381}$	**$C^{59}C^{413}$**[V_s*:158;pp:237;pp2:61*], **$C^{151}C^{199}$**[V_s*:100;pp:93;pp2:15*], **$C^{372}C^{381}$**[V_s*:100;pp:81;pp2:79*]	None
Lysozyme	$C^{24}C^{145}$, $C^{48}C^{133}$	**$C^{24}C^{145}$**[V_s*:140;pp:65;pp2:88*], **$C^{48}C^{133}$**[V_s*:100;pp:62;pp2:55*]	$C^{48}C^{133}$[V^*_s*:135;pp:51;pp2:33*]
FT III	$C^{81}C^{338}$, $C^{91}C^{341}$	**None**	None
Aldolase	None	**None**	None
Aspa	None	**None**	None

Comparative studies with MassMatrix

At the state-of-the-art MS2Assign[9] and MassMatrix[10] are two MS-based methods that can be applied to the problem of determining S-S bond connectivity. In our previous work[3], the MS2DB system developed by us was found to be comparable to MS2Assign[9], albeit, in limited testing. Since the proposed method improves upon MS2DB and due to space limitations, we only present detailed comparative results with MassMatrix[10] in table 6. As part of this experiment, for each S-S bond, in addition to the empirical match score (Eq. 4), a probability based scoring model proposed in [23] was implemented. This model provided two scores called *pp* and *pp2* scores. The *pp* score helps to evaluate whether the number of *VMs* could be a random. The *pp2* score evaluates whether the total abundance (intensity) of *VMs* could be a random. We refer the reader to [23] for a detailed description and formulae of the *pp* and

pp2 scores. The reader may note that the proposed method had better *pp* and *pp2* scores when compared to MassMatrix. While the match scores (V_s) obtained with the proposed method were higher than those obtained with MassMatrix (V^*_s), no inferences should be drawn as these scores are calculated differently. As can be seen from Table 6, every bond correctly determined by MassMatrix was also found by us. However, there were S-S bonds in C2GnT-I and Lysozyme that were found by the proposed method but not by MassMatrix.

Quantitative assessment and analysis of the method's performance

If the set of disulfide bonds are denoted by *P* and the set of cysteines not forming disulfide bonds by *N*, then true positive (*TP*) predictions occur when disulfide bonds that exist are correctly predicted. False negative (*FN*) predictions occur when bonds that exist are not predicted as such. Similarly, a true negative (*TN*) prediction correctly identifies cysteine pairs that do not form a bond. Finally, a false positive (*FP*) prediction, incorrectly assigns a disulfide link to a pair of cysteines, which are not actually bonded. Based on these definitions, we use the following four standard measures to analyze the proposed method.

Sensitivity (Q_c) = TP/P (7), Specificity (Q_{nc}) = TN/N (8), Accuracy (Q_2) = $(TP+TN/(P+N)$ (9)

Matthew's correlation coefficient (c) = $$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \quad (10)$$

Table 5 - Sensitivity, specificity, accuracy and Mathew's correlation coefficient results for all nine proteins analyzed.

Protein	Q_c	Q_{nc}	Q_2	c
ST8Sia IV	1.00	1.00	1.00	1.00
Beta-LG	0.50	1.00	0.95	0.69
FucT VII	1.00	1.00	1.00	1.00
C2GnT-I	0.75	1.00	0.98	0.86
Lysozyme	1.00	1.00	1.00	1.00
B1,4-GalT	1.00	1.00	1.00	1.00
FT III	0.00	1.00	0.89	X
Aldolase	X	1.00	1.00	X
Aspa	X	1.00	1.00	X

In Table 5 we present the results obtained for our framework. With maximum specificity and high accuracy (98% average), the method correctly reported the connectivity for most of the proteins. The method only failed to identify three disulfide bonds. One intra-bond in the Beta-LG protein could not be found due to a blind spot caused by the same intra-bond, making the protein's fragmentation difficult during mass spectrometry. Peptides with intra-chain disulfide bonds might suffer from too few product ions generated in the MS/MS spectrum. In this case, the precursor ion fragmentation produces different fragments only in the outside boundaries of the

intra-disulfide bond. Fragmentation that occurs in the sequence connecting the two linked sulfides results in the same product ion and, thus, no useful sequence information. This region is referred to as a *blind spot*. Two cross-linked bonds in the FT III protein also could not be identified because this particular connectivity configuration creates a large disulfide-bonded structure, which is poorly fragmented by tandem mass spectrometry. One bond in the C2GnT-I protein could not be found, since the precursor ion cannot be formed by chymotryptic digestion, which was the digestion carried for C2GnT-I. It is important to note that neither MassMatrix nor MS2Assign were able to identify these bonds.

Conclusions

We have presented an algorithmic framework for determining S-S bond topologies of molecules using MS/MS data. The proposed approach is computationally efficient, data driven, and has high accuracy, sensitivity, and specificity. It is not limited either by the connectivity pattern or by the variability of product ion types generated during the fragmentation of precursor ions. Furthermore, the approach does not require user intervention and can form the basis for high-throughput S-S bond determination.

Authors' contributions

The algorithmic solution framework was designed by RS and implemented by WM. Computational studies and experiments were carried out by WM and RS. T-YY developed the experimental protocols and generated the data. The paper was written by RS with help from WM.

Acknowledgements

WM and RS were supported by funding from NSF grant IIS-0644418 (CAREER). T-YY was supported by NSF grant CHE-0619163 and NIH grant P20MD000544.

References

01. Singh R: **A review of algorithmic techniques for disulfide-bond determination**. *Briefings in Functional Genomics and Proteomics*, 7(2), pp. 157-172.
02. Nesvizhskii AI, Vitek O, Aebersold R: **Analysis and validation of proteomic data generated by tandem mass spectrometry**. *Nature Methods*, 4(10), pp. 787-797.
03. Lee T, Singh R: **Comparative Analysis of Disulfide Bond Determination Using Computational-Predictive Methods and Mass Spectrometry-Based Algorithmic Approach**. *BIRD*, CCIS 13, 2008, pp. 140-153.

04. Johnson RS, Martin SA, Biemann K, Stults JT, Watson JT: **Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine**. *Anal Chem*, 1987, 59:2621
05. Cormen TH, Leiserson CE, Rivest RL, Stein C: Introduction to Algorithms, 2nd edition, MIT Press, Cambridge, MA, U.S.A, 2001.
06. Murad W, Singh R, Yen TY: **Polynomial-Time Disulfide Bond Determination Using Mass Spectrometry Data**. *IEEE Computational Structural Bioinformatics Workshop*, 2009, 79-86.
07. Steen H, Mann M: **The abc's (and xyz's) of peptide sequencing**. *Nature Reviews, Molecular Cell Biology*, 5, 2004, pp.699-711.
08. Thomas S, Yen TY, Macher BA: **Eukaryotic glycosyltransferases: cysteines and disulfides**. *Glycobiology*, 12, 2002, pp. 4G-7G.
09. Schilling B et al: **MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides**. *J Am Soc Mass Spectrom.*, 14(8):834-50, 2003.
10. Xu H, Zhang L, Freitas MA: **Identification and Characterization of Disulfide Bonds in Proteins and Peptides from Tandem MS Data by Use of the MassMatrix MS/MS Search Engine**. *Journal of Proteome Research*, 7, 13-144, 2008.
11. Lee T, Singh R, Yen TY, Macher B: **An Algorithmic approach to Automated High-Throughput Identification of Disulfide Connectivity in Proteins Using Tandem Mass Spectrometry**. *Computational Systems Bioinformatics*, 2007, pp. 41-51.
12. Lee T, Singh R, Yen R, Macher B: **A mass-based hashing algorithm for the identification of disulfide linkage patterns in protein utilizing mass spectrometry data**. *IEEE International Symposium on Computer-Based Medical Systems*, 2007, pp. 397-402.
13. Gabow HN: **An efficient implementation of Edmonds' Algorithm for Maximum Matching on Graphs**. *Journal of the ACM*, 23, 2006, 221-234.
14. Fariselli P, Casadio R: **Prediction of disulfide connectivity in proteins**. *Bioinformatics* 2001, 17, 957-64.
15. Chen T, Jaffe JD, Church GM: **Algorithms for Identifying Protein Cross-links via Tandem Mass Spectrometry**. *RECOMB*, 2001, pp. 95-102.
16. Frank A, Tanner S, Pevzner P: **Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry**. *RECOMB* 2005, LNBI 3500, pp. 326-341, 2005.
17. Keller A, Nesvizhskii AI, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS database search**. *Analytical Chemistry*, 74(20), 2002.
18. Rothberg E: http://elib.zib.de/pub/Packages/mathprog/matching/weighted/
19. Yen TY, Macher BA: **Determination of glycosylation sites and disulfide bond structures using LC/ESI-MS/MS analysis**. *Methods in enzymology* 2006; 415:103-113.
20. Ferre F, Clote P: **DiANNA: A Web Server for Disulfide Connectivity Prediction**. *Nucleic Acids Research* 2005; 33:230-232.
21. Ceroni A et al: **DISULFIND: A Disulfide Bonding State and Cysteine Connectivity Prediction Server**. *Nucleic Acids Research* 2006; 34:177-181.
22. Tsai CH et al: **Improving disulfide connectivity prediction with sequential distance between oxidized cysteines**. *Bioinformatics* 2005, 21:4416-4419.
23. Xu H, Freitas MA: **A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data**. *BMC Bionformatics 2007*, 8:133-142.