Polynomial-Time Disulfide Bond Determination Using Mass Spectrometry Data

William Murad[*], Rahul Singh[*,1], Ten-Yang Yen[**]
[*]*Department of Computer Science,* [**]*Department of Chemistry and Biochemistry,*
San Francisco State University, San Francisco, CA 94132

Abstract

We present an algorithmic approach for determining, in polynomial time, disulfide bonds in proteins using mass spectrometry data. The proposed technique is based on matching the set of all theoretically possible disulfide bonded structures with precursor ions derived from a tandem MS/MS experiment. For each match found, theoretical fragments from a disulfide bonded peptide structure are matched with precursor ion fragments obtained from the same tandem MS/MS data set in order to determine a protein's disulfide linkage pattern. A polynomial-time approximation strategy is proposed to selectively generate the set of theoretically possible disulfide bonded linkages for matching. Experiments demonstrate the efficacy of the method.

1. Introduction

A disulfide bond, also called SS-bond or disulfide bridge, is a single covalent bond formed from the oxidation of sulfhydryl groups. Among the twenty natural amino acids, cysteine is unique because it is the only amino acid involved in the formation of disulfide bridges [1].

Disulfide bonds play an important role in understanding protein folding, evolution, and structural properties, imposing length and angle constraints on the backbone of a protein. SS-bonds can be classified into three categories: structural, catalytic or allosteric [2]. Structural disulfide bonds play an important role in the folding and stabilization of proteins. Catalytic bonds mediate thiol-disulfide interchange reactions in substrate proteins and are important for regulation of enzymatic activity. Allosteric disulfide bonds, in contrast to catalytic disulfides, control the functioning of proteins by triggering changes in either the intra-molecular or inter-molecular protein structure, acting essentially as switches for protein function.

An example highlighting the importance of S-S bonds can be found by considering the polysialylation of NCAM. The neural cell adhesion molecule NCAM has an important role in neuronal development and regeneration. As reported in [3], mutations at cysteine residues involved in the disulfide bridges in ST8Sia IV completely inactivated this enzyme, thereby disrupting the polysialylation of NCAM.

It is a complex technical problem to determine the disulfide bonding pattern for a given protein. At the state-of-the-art, three classes of techniques can be broadly distinguished: (1) Protein crystallography followed by manual identification of disulfide-bonds, (2) predictive techniques using machine learning, and (3) algorithmic analysis of mass spectrometry data. The method proposed by us in this paper falls in the last of the aforementioned categories and builds on our prior research [4, 6].

2. Basics of Mass Spectrometry

Mass Spectrometry has become the standard high-throughput method for protein identification, and more recently, for protein quantification. It is a powerful analytical technique used for identification of unknown compounds, quantification of known compounds, and to elucidate the structure and chemical properties of molecules. An advantage of Mass Spectrometry is that experiments can be accomplished with small quantities of analytes (as little as 10^{-12}g, 10^{-15} moles for peptides and small proteins). Depending on the physics of the method, different types of mass spectrometry techniques can be identified. Among others, these include Inductively Coupled Plasma Mass Spectrometry (ICP-MS), Thermal Ionization Mass

[1] Corresponding author: E-mail rsingh@cs.sfsu.edu

Spectrometry (TIMS), and Spark Source Mass Spectrometry (SSMS).

The experimental data in this work is derived from a LC/ESI-MS/MS (Liquid Chromatography / Electrospray Ionization – Tandem Mass Spectrometry) experiment, where analytes of interests are first separated by Liquid Chromatography, and ionized by Electrospray Ionization (ESI). Depending on the polarity of high voltage applied to the ESI source, positive or negative ions are generated in the electrospray ion source. Positive-ion ESI spectra are most commonly recorded for analyzing peptides and proteins because they are more sensitive.

A MS/MS experiment is a Mass Spectrometry experiment in which ions are subjected to two stages of analysis to confirm their structures or sequences. In the first stage, ions of interests are selected for fragmentation through the process of Collision-Induced Dissociation (CID). The vast majority of the peptide fragments generated are either *b*-ions or *y*-ions: if the charge is retained on the fragment's N terminal, then the ion is classified as a *b*-ion. Alternatively, if the charge is retained on the fragment's C terminal, then the ion is identified as a *y*-ion. The second stage of mass analysis records the MS/MS spectrum of these fragments.

MS methods for determining the location of disulfide bonds normally require liquid chromatographic separation of the disulfide linked peptides generated from a protein by proteolytic cleavage such as tryptic or chymotryptic digestion. Subsequently, these purified disulfide-containing peptides are ionized by ESI, and identified by means of mass determination. To confirm the location of a disulfide linkage, the disulfide-containing peptides are subjected to MS/MS analysis as described previously.

In a typical LC/ESI-MS/MS analysis for mapping the location of disulfide bonds, tens of thousands of MS/MS spectra are produced from a single protein digest making manual identification prohibitively complex. Unfortunately, the problem also does not lend easily to algorithmic solutions since the number of possible disulfide bonding patterns grows exponentially with the number of residues in a given protein. Furthermore, special care needs to be taken to account for different bonding topologies wherein bonds may occur within cysteines belonging to the same chain (intra-bonds) and/or among cysteines belonging to different chains (inter-bonds). Finally, measurements of fragment-based bonding patterns can be influenced by noise and they also need to be coalesced into an overall connectivity pattern that is physically consistent as one cysteine can participate in at most one disulfide bond.

3. Problem Formulation

Disulfide bond determination can be intuitively formulated as a matching problem wherein the masses of all the theoretically possible disulfide-bonded peptide structures, formed by connecting cysteine-containing peptides obtained from the proteolytic digestion of the protein, are compared with the fragment masses from the precursor ion mass list (*PML*). The *PML* is a list containing the mass of each precursor ion obtained from the *DTA* files derived from the tandem MS/MS experiment. In a *DTA* file, the precursor peptide mass is calculated from the mass of a singly protonated MH^+ value independent of the charge state.

A 'match' can be declared when the difference between the detected mass of a targeted ion and the calculated mass of a possible disulfide bonded pair of peptides is deemed to be small in magnitude (i.e. the masses are similar). Here, a matching threshold is necessary due to practical constraints such as equipment sensitivity and experimental noise, which would typically ensure that an exact match would rarely occur. We call this threshold IM_{TH}.

A key step in realizing the above strategy lies in determining the set of all theoretically possible disulfide-bonded peptide structures of a protein. In [4], this set was called Disulfide Bond Mass Space (*DMS*) – a term, which we shall also use hereafter. Unfortunately however, the size of the *DMS* grows exponentially with the number of cysteines. Current approaches [4, 6] for the construction of a *DMS*-similar mass space are exponential, requiring $O(k^p)$ time to compute, where k is the number of cysteine-containing peptides and p is the maximum number of peptides in a disulfide-bonded structure. The values of k for the nine proteins considering up to two missing cleavages, with their respective UniProtKB reference ID according to [http://www.uniprot.org], are presented in Table 1.

Table 1. Proteins UniProtKB ID and k values

Protein	UniProtKB ID	k value
ST8Sia IV	Q92187	33
Beta-LG	P02754	20
FucT VII	Q11130	24
C2GnT-I	Q09324	65
Lysozyme	P00698	45
FT III	P21217	42
B1,4-GalT	P08037	42
Aldolase	P00883	46
Aspa	Q9R1T5	33

Due to the limitation of fragments that can be generated by dissociation methods used in mass spectrometers, we currently consider the value of p to be no more than three. However, future improvements in the MS/MS fragmentation technique including the use of more capable and accurate mass spectrometers will allow p values to increase, which will add more value to our polynomial time solution.

4. A Polynomial Time Algorithm for Determining Disulfide-Bond Patterns

4.1. Data Preprocessing

Given a protein sequence and the protease(s) used in the Mass Spectrometry experiment, we first identify the peptides resulting from the protein's digestion. Next, peptides that do not contain cysteines are discarded. These preprocessing steps run in *O(n)* time, where *n* is the length of the protein sequence.

The algorithm also treats, for each cysteine-containing peptide, the missing cleavage cases during digestion. The peptide's sequence is expanded on both sides, generating new peptides for the cases when either one or two missing cleavages could occur. This step requires *O(kn)* time to compute, where *k* is the number of cysteine-containing peptides after a digestion without missing cleavages and *n* is the length of the protein peptide sequence.

Next, for each *DTA* file, our algorithm retrieves the MH^+ value, subtracts 1Da to obtain the precursor ion mass *M* and adds the value of *M* to the *PML* list.

4.2. Polynomial Time DMS Construction

As described in Section 3, the key step (and challenge) in determining the disulfide bonds lies in matching the *DMS* with the *PML*. Our key insight underpinning a non-exponential solution lies in that the *entire DMS does not need to be generated to determine the match*. Indeed, only those parts of the *DMS* need to be generated, whose mass values, defined as the sum of the masses of the disulfide bond-linked digested peptides, are potentially "close" to the *PML* value being matched. Thus, the matching problem can be cast as the subset-sum problem [5]. Unfortunately, the subset-sum problem is NP-Complete. However, it can be solved using approximation strategies to obtain near-optimal solutions. In the following, we describe the approximation strategy used by us for partial generation of the *DMS* and the matching.

Our strategy for solving this optimization problem is to use an approximation algorithm that takes as input not only the cysteine-containing peptides list derived from the protein's digestion, but also a value $\varepsilon > 0$ such that for any fixed ε, the algorithm is a $(1+\varepsilon)$-approximation scheme. That is, for any fixed $\varepsilon > 0$, the algorithm runs in polynomial time. From [5], an approximation algorithm is called fully polynomial-time, if its running time is polynomial both in $(1/\varepsilon)$ and in the size k of the list of cysteine-containing peptides.

As the value of ε is inversely proportional to the approximation algorithm's running time, we need to find a maximum value of ε, such that disulfide bonds (true positives) will not be missed. Through empirical test, cohere the value of ε was systematically varied, we discovered that setting one single static value of ε for all the proteins did not produce good results. Moreover, we found that the optimal value of ε was inversely proportional to the number of cysteine-containing peptides *k*. We also noticed a correlation between the mass range of the cysteine-containing peptides and the value of ε. We express these relationships through Eq. (1) and use it to calculate ε:

$$\varepsilon = \frac{\left(CCP_{max} - CCP_{min}\right)}{CCP_{average}} \times \frac{1}{k} \qquad (1),$$

where CCP_{max} is the mass of the cysteine-containing peptide with highest mass value, CCP_{min} is the mass of the cysteine-containing peptide with lowest mass value, and $CCP_{average}$ is the Cysteine-Containing Peptides (CCP) average mass value.

The approximation algorithm for creating the partial *DMS* is described in terms of the APPROX-DMS and TRIM procedures (Figure 1). In the pseudocode presented, the call to **unset**(*TempList*) removes from memory the list provided as the argument. The method **merge**(*DMS*, *TempList*) returns a sorted list that is the merge of its two sorted input lists *DMS* and *TempList* with duplicated values removed. We note that the routine **merge**(.,.) runs in linear time *O(|DMS|+|TempList|)*. Finally, PML_{val} denotes a precursor ion mass value in *PML* and IM_{TH} denotes the Initial Match threshold.

The TRIM routine shortens each DMS list and forms the basis of the fully polynomial-time approximation scheme. If two disulfide-bonded peptide structures have similar mass values, then for the purpose of finding an approximate solution there is no reason to maintain both of them. To trim the *DMS* list by ε, with $0 < \varepsilon < 1$, means to remove as many elements from *DMS* as possible, in such a way that if DMS^* is the result of trimming *DMS*, then for every element DMS_i that was removed from *DMS*, there is an

element DMS^*_i still in DMS^* that is "sufficiently" close in terms of its mass to DMS_i, that is,

$$\frac{DMS_i}{1+\varepsilon} \le DMS^*_i \le DMS_i \qquad (2)$$

With both *DMS* and *PML* determined, the algorithm matches the masses between *DMS* and *PML* components in order to determine precursor matches, whose mass difference falls below a threshold IM_{TH} = ±1.5Da. We call these matches Initial Matches.

```
APPROX-DMS (CCP, PML_val, ε , IM_TH )
   CCPsize ← | CCP |
   DMS_0 ← {0}
   for i ← 0 to CCPsize -1
      PM ← CCP_i        //PM = Peptide Mass
      unset (TempList)
      DMSsize ←| DMS |
      for j ←0 to DMSsize -1
         if ((PM + DMS_j) ≥ (PML_val - IM_TH))
            if ((PM + DMS_j) ≤ (PML_val + IM_TH))
                  TempList_j ← PM + DMS_j
      DMS ←merge (DMS, TempList )
      DMS ←TRIM (DMS, ε )
   return DMS

TRIM (DMS, ε )
   n ←| DMS |
   DMS*_n-1 ← DMS_n-1
   last ← DMS*_n-1
   for i ←n-1 to 0
      if last > ( ( 1+ε )× DMS_i )
         DMS*_i ← DMS_i
         last ← DMS*_i
   return DMS*
```

Figure 1. Pseudocode for the APROX-DMS and TRIM procedures

By using a similar mathematical argument as in the proof of theorem 35.8 in [5], it can be demonstrated that the running time of APPROX-DMS is polynomial in k and in $(1/\varepsilon)$; therefore our algorithm is a fully polynomial-time solution for the disulfide-bonded mass space creation.

4.3. Applying the Algorithm to MS/MS Data

For each Initial Match, a Disulfide Bonded Fragment Mass Space (*FMS*) and a MS/MS Mass List are calculated in order to compute the Confirmed Matches (*CM*). The *FMS* is the set of masses of every disulfide-bonded fragment structure that can be obtained from the arrangement of *b*-ions and *y*-ions among the peptides involved in the Initial Match. Considering an average case where all fragments have approximately the same length or the same number of amino acids residues $||f||$ (i.e. for a two fragments structure, let's consider $||f1|| \approx ||f2|| \approx ||f||$), this step runs in $O(2^n||f||^n)$, where n is the number of fragments in the disulfide-bonded peptides. Typically, simultaneous values of n greater than 3 and $||f||$ greater than 15 are extremely rare. In fact, a configuration of 3 fragments with length 15 each would not be identified by the Mass Spectrometry apparatus used by us due to its mass range limitation. The value of n also decreases with an increase of $||f||$, and vice-versa. In most of the cases, n will be 2 and $||f||$ will be less than 25; therefore the effective complexity for this step can be considered as cubic for the worst-case scenario and does not represent a bottleneck for our implementation. Here, a similar idea to the *DMS* construction could also be applied as only those arrangements of *b*-ions and *y*-ions, whose mass values, defined as the sum of the masses of each fragment ion, are potentially close to a MS/MS Mass List value. By using an approximation algorithm that generated the *FMS* in polynomial time, the method would become both theoretically and practically polynomial-time, independent of the disulfide bond linkage pattern. Part of our current research is directed towards this goal.

A *TML* or MS/MS Mass List is the set of masses of the peptide fragments obtained from a precursor ion in a tandem MS/MS experiment *DTA* file. A *DTA* file usually contains hundreds of fragment ions with the ion intensity ranging from 50 to about 10^7. We select the 50 measured fragments' mass-to-charge ratio (*m/z*) with the highest intensity for each DTA file. The screening of the peaks begins by selecting all the peaks whose intensities are higher than at least 5% of the highest peak's intensity. If 50 or more *m/z* values are returned, the algorithm stops the screening process, discards any extra value, and returns an array with the top 50 *m/z* values. If less than 50 values are returned, the algorithm lowers the screening threshold by 1% and re-run the screening procedure, until at least 50 *m/z* values are obtained. This step requires $O(tk)$ time, where t is the number of iterations necessary to select 50 *m/z* values and k is the number of fragments in the *DTA* file. Usually only one iteration (t=1) is needed.

Once the 50 *m/z* values with highest intensity are selected, the mass-to-charge ratio is converted to a mass value, so that a match between elements from the *FMS* and elements from the MS/MS list becomes possible. The mass of a fragment can be calculated as:

$$M = \left(\frac{m}{z}\right) \times n - n \qquad (3),$$

where *m/z* is the mass-to-charge ratio and *n* is the charge state. In a *DTA* file, the charge state for each fragment is not provided, so the algorithm predicts it based on a fragment's *m/z* and on the precursor ion mass. Depending on the charged state of the precursor ion, fragments can be a mixture of different charged ions. For a triply charged precursor ion, its fragments can be singly, doubly or triply charged. All these possible configurations are stored in the *TML* mass list. This step runs in *O(k)*, where *k* is the number of fragments in the *DTA* file. It may be noted that the size of a MS/MS mass list can not exceed 150, since we use only 50 peaks and the precursor ions are singly, doubly or triply charged.

At this point, the algorithm searches for matches between *FMS* and *TML*. A Confirmed Match *CM* occurs when the difference between a *FMS* value and a *TML* value falls below a confirmed match threshold (CM_{TH}). This condition is formally described in Eq. (4).

$$|FMS_{val} - TML_{val}| \leq CM_{TH} \quad (4)$$

In our experiments we use the confirmed match threshold (CM_{TH}) value of 1 and the the output from this step is a collection of confirmed matches *CM* for each initial match *IM* and it takes O(||*FMS*||||*TML*||) time to compute, where ||*FMS*|| is the number of entries in the *FMS* list and ||*TML*|| is the number of entries in the *TML* list.

4.4. Overall Disulfide Bond Pattern Determination

The collection of Confirmed Matches for each Initial Match represents all possible disulfide bonds between pairs of cysteines residues. Once all Initial Matches and, consequently, all Confirmed Matches are calculated, the disulfide bonds found need to be aggregated into a weighted graph in order to obtain an overall connectivity pattern which is physically consistent and also represents the most likely pattern, given the data analyzed and the fact that one cysteine can only participate in at most one disulfide bond.

The overall disulfide bond pattern can be detected by finding the maximum weight matching on a weighted graph. The cysteines occupy the vertices of the graph and the weight of an edge *E* between two vertices C_1 and C_2 is defined as the number of Confirmed Matches calculated between cysteines C_1 and C_2. We obtain the maximum weight matching, which represents the protein's overall disulfide bond pattern, by integrating to our application the MATHPROG Solver for Maximum Weight Matching Problem, which is an implementation of the Edmonds-Gabow algorithm [7] and can be found at [http://elib.zib.de/pub/Packages/mathprog/matching/weighted/]. This step runs on $O(V^3)$, where *V* is the number of cysteines or vertices for all the Confirmed Matches.

5. Experiments and Case Studies

5.1. Quantitative Results

Nine proteins and their respective mass spectrometry data files were used to test our algorithm: ST8Sia IV, Beta-lactoglobulin (Beta-LG), FucT VII, C2GnT-I, Lysozyme, FT III, β1-4GalT, Aldolase, and Aspa. All UniProtKB IDs for the aforementioned proteins are presented in Table 1.

Table 2 summarizes the values encountered for the number of cysteine-containing peptides after protein's digestion (*CCP*), ε (approximation scheme trimming parameter), partial *DMS* size (number of disulfide-bonded peptide structures obtained using the fully polynomial-time approximation scheme), *PML* size (number of precursor ions from the tandem MS/MS *DTA* files) and Initial Matches (matches between *DMS* and *PML* values) for all nine proteins.

Table 3 shows the substantial decrease in *DMS* size (number of disulfide-bonded peptide structures) when we compare the entire *DMS*, determined by using the exponential-time approach, and the partial *DMS*, determined by using the polynomial-time solution proposed in this paper. The entire DMS is formed by possible disulfide bonded pair of peptides, whose mass does not exceed the mass of the precursor ion, as a Initial Match (*IM*) is only declared when the difference between the detected mass of a targeted precursor ion and the calculated mass of a possible disulfide bonded pair of peptides is smaller than the Initial Match threshold (*IM*). The partial *DMS* is generated by the approximation algorithm proposed in section 4.2.

For the protein Beta-LG, the *DMS* size difference is smaller than the size difference for other proteins' entire *DMS* and partial *DMS*, because the peptides derived from the tryptic digestion of Beta-LG have long amino acid chains. Therefore, the masses of most of the disulfide-bonded structures formed in this case exceed the maximum mass threshold (PML_{max}) and are excluded.

The enhancements achieved are again noticed when we compare the CPU time necessary to determine the disulfide bond arrangement using the

fully-polynomial time disulfide bond determination algorithm and the exponential time algorithm. For this comparison, we used a laptop equipped with an Intel 1.86Ghz Pentium Dual-Core Mobile processor T2390, 1MB L2 cache, and 2GB of DDR2 RAM memory, running Windows XP Professional SP3 and Apache 2.2 as the web server.

The CPU time results, measured in seconds, are reported in Table 4.

Table 2. Initial matches' summary for all nine proteins analyzed

Protein	*CCP*	$\mathcal{E}$	*Partial DMS*	*PML*	*IM*
ST8Sia IV	33	0.0384	9	79	9
Beta-LG	20	0.0749	14	552	14
FucT VII	24	0.0632	31	158	31
C2GnT-I	65	0.0267	33	212	33
Lysozyme	45	0.0370	35	121	35
FT III	42	0.0650	97	357	97
B1,4-GalT	42	0.0404	29	211	29
Aldolase	46	0.0499	6	184	6
Aspa	33	0.0812	5	673	5

Table 3. DMS size comparison: total number of disulfide-bonded peptide structures for each protein analyzed

Protein	Entire *DMS*	Partial *DMS*
ST8Sia IV	294	9
Beta-LG	17	14
FucT VII	720	31
C2GnT-I	316	33
Lysozyme	1818	35
FT III	5396	97
B1,4-GalT	235	29
Aldolase	107	6
Aspa	31	5

Table 4. CPU Time comparison (in seconds) between polynomial-time algorithm and exponential-time algorithm solutions

Protein	Exponential-time	Polynomial-time
ST8Sia IV	3 s	1 s
Beta-LG	5 s	3 s
FucT VII	5 s	3 s
C2GnT-I	40 s	27 s
Lysozyme	38 s	14 s
FT III	27 s	6 s
B1,4-GalT	15 s	8 s
Aldolase	5 s	3 s
Aspa	12 s	4 s

5.2. Comparison with known disulfide bonding patterns and with MS2DB

In this section we present the results from our algorithm, and compare them with the known disulfide bonding patterns of the molecules used in the experiments according to the UniProtKB database and according to [11]. We also present comparative results on these molecules with MS2DB application in [4]. While our implementation predicted 6 out of 7 disulfide bonds for the three proteins (ST8Sia IV, Beta-LG, and FucT VII) analyzed in the case studies section 5.3, MS2DB only predicted 4 out of 7 SS-bonds. The detailed disulfide bond arrangement for all nine proteins is presented in Table 5.

Table 5. Disulfide bonding patterns comparison for all nine proteins analyzed

Protein	Known Pattern	Proposed Algorithm	MS2DB
ST8Sia IV	$C^{142}C^{292}$	$C^{142}C^{292}$	$C^{142}C^{292}$
	$C^{156}C^{356}$	$C^{156}C^{356}$	-
Beta-LG	$C^{82}C^{176}$	$C^{82}C^{176}$	$C^{82}C^{176}$
	$C^{122}C^{135}$	$C^{3}C^{12}$	-
FucT VII	$C^{68}C^{76}$	$C^{68}C^{76}$	$C^{68}C^{76}$
	$C^{211}C^{214}$	$C^{211}C^{214}$	-
	$C^{318}C^{321}$	$C^{318}C^{321}$	$C^{318}C^{321}$
C2GnT-I	$C^{59}C^{413}$	$C^{59}C^{413}$	$C^{59}C^{413}$
	$C^{100}C^{172}$	$C^{100}C^{172}$	$C^{100}C^{172}$
	$C^{151}C^{199}$	$C^{151}C^{199}$	$C^{151}C^{199}$
	$C^{372}C^{381}$	$C^{372}C^{381}$	$C^{372}C^{381}$
Lysozyme	$C^{24}C^{145}$	$C^{24}C^{145}$	$C^{24}C^{145}$
	$C^{48}C^{133}$	$C^{48}C^{133}$	$C^{48}C^{143}$
	$C^{82}C^{98}$	$C^{82}C^{98}$	$C^{82}C^{98}$
	$C^{94}C^{112}$	$C^{94}C^{112}$	$C^{94}C^{112}$
FT III	$C^{81}C^{338}$	$C^{81}C^{338}$	$C^{81}C^{338}$
	$C^{91}C^{341}$	$C^{91}C^{341}$	$C^{91}C^{341}$
B1,4-GalT	$C^{134}C^{176}$	$C^{134}C^{176}$	$C^{134}C^{176}$
	$C^{247}C^{266}$	$C^{247}C^{266}$	$C^{247}C^{266}$
Aldolase	-	-	-
Aspa	-	-	$C^{145}C^{349}$

For the proteins C2GnT-I, FT III, and β1,4-GalT, the results produced by both solutions were identical and they matched the results expected according to the UniProtKB database. For the protein Lysozyme, our proposed algorithm results match with UniProtKB data as it found a disulfide bond between cysteine C^{48} and cysteine C^{133}, while MS2DB found a disulfide bond between cysteines C^{48} and C^{143}.

No disulfide bonds were found for the protein Aldolase, an expected result according to the UniProtKB database, which reinforces the accuracy of

both applications. Lastly, our algorithm, again in concordance with UniProtKB, didn't find any disulfide bond for the protein Aspa, where MS2DB returned an incorrect bond between cysteines C^{145} and C^{349}. These comparisons are expanded in [4], where MS2DB is compared with MS2Assign, which is a mass spectrometry-based method for determining cross-linkages, and it is also compared to other well-known predictive methods, such as DiANNA, DISULFIND, and PreCys, to determine proteins' disulfide connectivity pattern.

It should also be noted that the proposed technique, unlike MS2DB, is fully automated. The user needs to provide only a protein sequence, the protease used, and the *DTA* files from the Mass Spectrometry experiment. All other parameters and thresholds are automatically calculated as described by us earlier.

To better illustrate the results obtained by our application, here we present more detailed results for three of the proteins analyzed: ST8Sia IV, Beta-Lactoglobulin (Beta-LG), and FucT VII.

5.3. Case Studies

5.3.1. ST8Sia IV

For all 9 Initial Matches, there were a total of 30 Confirmed Matches (*CM*s), resulting in two different disulfide bonds (Figures 2.A and 2.B): between cysteines 142-292 (12 *CM*s) and cysteines 156-356 (18 *CM*s). According to [3], these two disulfide bonds represent the true disulfide bond pattern for the protein ST8Sia IV. Therefore, our algorithm obtained 100% accuracy in predicting the disulfide linkage for the protein ST8Sia IV, digested with chymotrypsin.

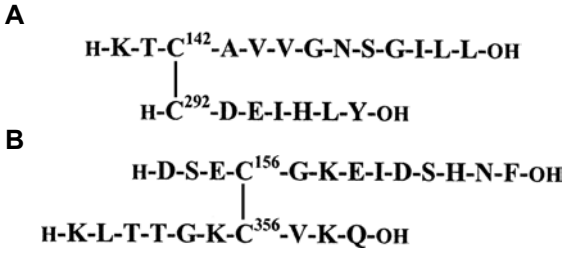

Figure 2. (A) ST8Sia IV SS-bond $Cys^{142}Cys^{292}$, (B) ST8Sia IV SS-bond $Cys^{156}Cys^{356}$

5.3.2. Beta-LG

For 14 Initial Matches, there were a total of 50 Confirmed Matches, corresponding to two distinct disulfide bonds. Based on [8], our algorithm successfully identified the disulfide bond between cysteines 82-176 (35 *CM*s) as shown in Figure 3; however it did not identify a disulfide bridge that might exist alternatively either between cysteines 122-137 or between cysteines 122-135.

One possible explanation of not being able to predict the second disulfide bond recognized in [8] is that peptides with intrachain 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 occurred 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 in [1] as the blind spot. In such cases the prediction of the disulfide bond using tandem MS/MS data becomes very difficult.

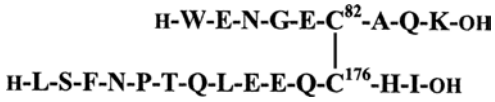

Figure 3. Beta-LG SS-bond $Cys^{82}Cys^{176}$

5.3.3. FucT VII

FucT VII was digested by a combination of trypsin and chymotrypsin. For all 31 Initial Matches, there were a total of 19 Confirmed Matches, resulting in six different disulfide bonds. After aggregating all the disulfide bridges into a weighted bipartite graph and finding the maximum weighted match, three disulfide bonds (Figures 4.A, 4.B, and 4.C) were recognized: between cysteines 68-76 (5 *CM*s), between cysteines 211-214 (4 *CM*s), and between cysteines 318-321 (5 *CM*s). Using the pattern determined in [9] as the gold standard, the disulfide bond arrangement was again successfully determined by our algorithm.

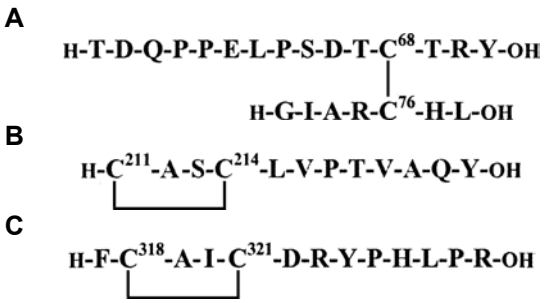

Figure 4. (A) FucT VII SS-bond $Cys^{68}Cys^{76}$, (B) FucT VII SS-bond $Cys^{211}Cys^{214}$, (C) FucT VII SS-bond $Cys^{318}Cys^{321}$

6. Conclusions and Future Work

In this paper, we have extended the prior works from our research group [4, 6] in disulfide bond determination using data from mass-spectrometry. The key contribution of this paper is the development and implementation of a fully polynomial-time approximation algorithm for solving this matching problem.

The polynomial-time method ran up to 3 times faster when compared to the exponential-time method, considering only a maximum of two disulfide-bonded peptides per structure ($p = 2$). For increasing values of p, the time to compute the disulfide bond arrangement of a protein using the exponential-time method is expected to rise very rapidly, giving the values of k presented in Table 1.

The method was tested on nine proteins with known disulfide bond patterns and the majority of the disulfide bonds were successfully predicted. One exception was an intrabond in the beta-lactoglobulin protein due to a blind spot caused by the same intrabond, making the protein's fragmentation difficult during the Mass Spectrometry experiment. While the proposed algorithm is theoretically approximate, we have suggested schemes for parameter selection that empirically ensure that no disulfide bonds are missed due to the approximation strategy.

When compared to our earlier work MS2DB, the proposed algorithm not only provides for better time-complexity, but in experiments it also demonstrated higher accuracy in disulfide bond determination for the entire data set consisting of ST8Sia IV, Beta-LG, FucT VII, C2GnT-I, Lysozyme, FT III, β1,4-GalT, Aldolase, and Aspa.

As part of our future research, we aim to: (1) investigate computational/statistical aspects of how the weight of the disulfide bonds (edges) in a weighted graph is computed, (2) refine the calculation of ε (trimming parameter) in order to maximize the accuracy of our implementation with the minimum CPU time necessary to predict disulfide bonds, and (3) implement an approximation algorithm, similar to the algorithm presented in section 4.2, to generate in polynomial time the Fragment Mass Space (*FMS*), thus making the entire solution both theoretically and practically polynomial, independent of the disulfide bond linkage pattern. Finally, we target to validate the proposed method on larger data sets and make a web-enabled version publicly available.

7. 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.

8. References

[1] H. Xu, L. Zhang, 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", *Journal of Proteome Research* 2008, 7, 13-144

[2] R. Singh, "A review of algorithmic techniques for disulfide-bond determination", Briefings in Functional Genomics and Proteomics, VOL 7. NO 2. 157-172

[3] K. Angata, T.Y. Yen, A. El-Battari, B. Macher, M. Fukuda, "Unique Disulfide Bond Structures Found in ST8Sia IV Polysialyltransferase Are Required for Its Activity", *The Journal of Biological Chemistry*, VOL. 276, No. 18, May 4, 2001, pp. 15369-15377

[4] T. Lee, R. Singh, "Comparative Analysis of Disulfide Bond Determination Using Computational-Predictive Methods and Mass Spectrometry-Based Algorithmic Approach", *BIRD* 2008, CCIS 13, 2008, pp. 140-153

[5] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., *Introduction to Algorithms*, 2nd edition, MIT Press, Cambridge, MA, U.S.A, 2001

[6] T. Lee, R. Singh, T.Y. Yen, B. Macher, "An Algorithmic approach to Automated High-Throughput Identification of Disulfide Connectivity in Proteins Using Tandem Mass Spectrometry", *6th Annual Conference on Computational Systems Bioinformatics* (CSB 2007), 2007, pp. 41-51

[7] H. N. Gabow, "An efficient implementation of Edmonds' Algorithm for Maximum Matching on Graphs", *Journal of the ACM*, VOL. 23, April, 2006, pp. 221-234

[8] H. A. Mackenzie, G.B. Ralston, D.C. Shaw, "Location of sulfhydryl and disulfide groups in bovine beta-lactoglobulins and effects of urea", *Biochemistry* VOL. 11, November, 1972, pp. 4539-4547

[9] T. de Vries, T.Y. Yen, R. K. Josh, J. Storm, D. H. van den Eijnden, R. M. A. Knegtel, H. Bunschoten, D. H. Joziasse, B. A. Macher, "Neighboring cysteine residues in human fucosyltransferase VII are engaged in disulfide bridges, forming small loop structures", *Glycobiology*, VOL. 11, No. 5, December 12, 2000, pp. 423-432

[10] T. Chen, J.D. Jaffe, G.M. Church, "Algorithms for Identifying Protein Cross-links via Tandem Mass Spectrometry", *Annual Conference on Research in Computational Molecular Biology*, 2001, pp. 95-102

[11] S. Thomas, T-Y Yen, B.A. Macher, "Eukaryotic glycosyltransferases: cysteines and disulfides", *Glycobiology*, VOL. 12, February 2002, pp. 4G-7G