AN ALGORITHMIC APPROACH TO AUTOMATED HIGH-THROUGHPUT IDENTIFICATION OF DISULFIDE CONNECTIVITY IN PROTEINS USING TANDEM MASS SPECTROMETRY

Timothy Lee and Rahul Singh*

Department of Computer Science, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132-4025,U.S.A.

Ten-Yang Yen and Bruce Macher

Department of Chemistry and Biochemistry, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132-4025, U.S.A.

Knowledge of the pattern of disulfide linkages in a protein leads to a better understanding of its tertiary structure and biological function. At the state-of-the-art, liquid chromatography/electrospray ionization-tandem mass spectrometry (LC/ESI-MS/MS) can produce spectra of the peptides in a protein that are putatively joined by a disulfide bond. In this setting, efficient algorithms are required for matching the theoretical mass spaces of all possible bonded peptide fragments to the experimentally derived spectra to determine the number and location of the disulfide bonds. The algorithmic solution must also account for issues associated with interpreting experimental data from mass spectrometry, such as noise, isotopic variation, neutral loss, and charge state uncertainty. In this paper, we propose a algorithmic approach to high-throughput disulfide bond identification using data from mass spectrometry, that addresses all the aforementioned issues in a unified framework. The complexity of the proposed solution is of the order of the input spectra. The efficacy and efficiency of the method was validated using experimental data derived from proteins with with diverse disulfide linkage patterns.

1. INTRODUCTION

Cysteine residues have a property unique among the 20 naturally occurring amino acids, in that they can pair to form disulfide bonds. These covalent bonds occur when the sulfhydryl groups of cysteine residues become oxidized (S-H + S-H → S-S + 2H).[1] Because disulfide bonds impose length and angle constraints on the backbone of a protein, knowledge of the location of these bonds significantly constrains the searchspace of possible stable tertiary structures which the protein folds into. The disulfide linkage pattern of a protein also can have an important effect on its function. For example, the disulfide bond structures of ST8Sia IV are necessary for its polysialyation activity.[2]

Methods for determining disulfide bonds in a protein can be classified as either: (1) *purely predictive*, based completely on the protein's primary structure, or (2) *based on analyzing data from experimental methods*, such as Crystallography, NMR, and Mass Spectrometry.[3,4] Predictive approaches typically aim to infer the disulfide bonding state of cysteine residues in a protein, primarily by characterizing a heuristically defined local sequence environment. Towards this goal, predictive approaches include graph-theoretic methods,[5] combinatorial optimization formulations,[6] techniques based on efficient indexing of the search space,[7] and a variety of supervised learning formulations involving neural-networks, hidden Markov models, and support vector machines.[8-10] However, Vullo and Frasconi concluded that any prediction algorithm must have a computational time complexity bounded by $O(n(\sqrt{n}/2)^n)$, where n is the number of cysteines in the protein.[8] This limits the application of such an algorithm to proteins with only a few disulfide bonds. In addition, the prediction accuracies or these methods, defined as the fraction of the total number of proteins whose connectivity patterns are correctly predicted, are currently limited to about 60%.

By contrast, determination of disulfide bonds can also be achieved with high accuracy for any number of bonds by analyzing data from structure elucidation techniques such as X-ray crystallography and NMR. These techniques require relatively large amounts (10 to 100 mg) of pure protein in a particular solution or crystalline state and are fundamentally low-throughput in nature.

In this context, the use of information from mass spectrometric (MS) analysis constitutes an important direction for elucidation of structural features, such as disulfide bonds. For identification of disulfide linkages, the general strategy involves mass spectrometry-based

* Corresponding author. Email: rsingh@cs.sfsu.edu

analysis to make an initial identification of the putative peptides involved in a disulfide bond. These peptides are then fragmented, and a tandem mass spectrum (MS/MS) of the fragments generated. The MS/MS spectrum is subsequently analyzed to confirm the initial identification of a disulfide bond. Such an approach can offer accurate identification and, in principle, can scale to any number of bonds with much less stringent sample purity requirements when compared to NMR or X-ray crystallography. Although the aforementioned approach is conceptually straightforward, the actual task of identifying the MS/MS spectra corresponding to disulfide linkages is non-trivial.

In this paper we investigate this precise problem. The key contributions of this work lie in addressing the problem of disulfide bond identification in the context of the technical challenges arising from the use of the real-world data from tandem mass spectrometric analysis. The combination of experimental procedure and algorithmic analysis proposed is scalable to structures having a large number of disulfide bonds. Furthermore, the processing is inherently high-throughput. Other features of the proposed approach include:

- *Invariance to the topology of the disulfide bonds*: Disulfide bonds may be classified as intra-molecular bonded (within a single peptide chains) or inter-molecular bonded (between different peptide chains). The proposed methodology can identify such bonds within a single framework.
- *Analysis of experimental errors/noise at the level of the produced spectrum:* Our proposed methodology requires the mass spectra and tandem mass spectra to be converted into a finite set of discrete "mass peaks." We present algorithms to resolve such peaks from spectra having peaks of non-zero width. We also address how to obtain the optimal set of peaks from each tandem mass spectrum.
- *Accounting for neutral loss and isotopic variation:* During the collision induced disassociation step of an LC/ESI-MS/MS analysis, a peptide fragment may have undergone *neutral loss*, resulting in the loss of a small molecule such as water or ammonia. In addition, the constituent atoms that comprise an amino acid exist in a number of isotopic forms. As a result, peptides consisting of the same sequence of amino acids will be measured as a series of masses by the mass spectrometer. This must be considered when computing the expected mass of a disulfide bonded peptide fragment.
- *Interpretation of the charge state*: Precursor ions with a high charge state (triply charged ion or greater) can be misinterpreted by MS data processing programs commonly supplied as part of the MS instrumentation. For example, ion trap mass spectrometers have a relatively low resolution. In such cases, a quadruply charged ion may not be well resolved and can be misinterpreted as a triply charged ion. This error often cannot be identified unless a higher resolution scan (zoom scan) is employed during the experiment. Consequently, the mass of a disulfide bonded pair of peptides is incorrectly computed, resulting in either not identifying (false negative) or incorrectly identifying (false positive) the bond.

1.1. Comparison of the Proposed Approach with Related Works

Examples of techniques employing purely predictive methodology include DiANNA,[11] DISULFIND,[12] and PreCys.[13] Each of these implementations employ weighted graph matching to predict the final disulfide connectivity pattern. In fact, these implementations all use a program (by Rothberg) that implements Gabow's algorithmic solution of the maximal weighted graph matching problem.[14] Additionally, a learning strategy is involved where fundamental assumptions are made about the relationship between the cysteine residues in order to obtain the edge weights. Examples of such assumptions include length of the local sequence environment to be considered, formulation of the residue contact potential function used, and assumptions involved in defining the training set. However, their reported prediction accuracies indicate that these underlying assumptions remain open to further investigations.

Existing web-based programs such as MS-Bridge in the ProteinProspector tools,[15] X! Protein Disulphide Linkage Modeler,[16] and Peptidemap[17] are useful when analyzing MS data from MALDI-TOF (Matrix Assisted Laser Desorption Ionization-Time of Flight) experiments. However, these programs do not analyze MS/MS data, thus missing the useful structural information inherent in this data. The program MS2Assign can be used to analyze disulfide linkages from MS/MS data.[18] However, because it was designed

primarily for the analysis of results from cross-linking studies, MS2Assign requires the user to input detailed information on the specific modifications expected. As a result, there is a need for a software tool that utilizes both MS and MS/MS data to identify disulfide linkage patterns in a high throughput manner.

2. THE PROPOSED METHOD

2.1. Problem Formulation

Let a_i denote the set of amino acid residues, each with mass m(a_i). A *peptide* $p = \{a_i\}$ is then a string of amino acids with mass $m(p) = \sum_i m(a_i) + 18$ Da (Daltons). Peptides have a specific directionality: the string starts at the unbonded amide group, called the N terminus, and ends at the carboxylic acid group, called the C terminus. The term 18 Da is included in this formula to account for the masses of H and OH of the N- and C-termini of the peptide, respectively.

In a LC/ESI-MS/MS experiment, a protease is used to divide a protein into peptides. A *protein* $A = \{p_i\}$ denotes the set of all peptides. A *cysteine-containing peptide c* is a peptide of protein *A* that has at least one of its amino acids a_i identified as a cysteine residue. Thus if $C = \{c_i\}$ is the set of all cysteine-containing peptides, then $C \subseteq A$. In practice, it is very rare that C = A.

A *disulfide bonded peptide* $D_{1,2}$ is a pair of cysteine-containing peptides c_1 and c_2, with mass $m(C_{1,2}) = m(c_1) + m(c_2) - 2m(H)$, where 2m(H) accounts for the mass of the two protons that are lost when the disulfide bond is formed. A disulfide connectivity pattern can be modeled in terms of an undirected graph $G = (V, E)$. The vertex set V represents the set of bonded cysteines and an edge $e \in E$ corresponds to a disulfide bridge between its adjacent cysteines. Admissible vertex and edge sets are constrained because an even number of intra-chain bonded cysteines is required and a cysteine can only be bridged to one and only one different cysteine. Thus, we have $|V| = 2B$, $|E| = B$ and *degree*(v) = 1 for any $v \in V$ (perfect matching), where B denotes the number of disulfide bonds in a chain.

The problem of identifying the correct connectivity pattern for a given disulfide bonded chain is simply formulated as finding the best possible candidate as given by a suitable scoring function. If we consider only those cysteines that are known to be involved in a disulfide bond, it is evident that this problem is equivalent to the problem of computing the maximum-weight perfect matching. In a perfect matching, every vertex of the graph is incident to exactly one of the edges of the matching. In this formulation, we attribute a weight w_e greater or equal to zero for the edge e of G that was initially identified by the MS spectrum match to each pair of cysteines.

The *disulfide bond mass space* $BMS = \{bms_i\}$ of a protein is the set of every possible pair of cysteine-containing peptides. A *mass list* $ML = \{ml_j\}$ is the set of numbers that represent the masses of the precursor ions obtained from a LC/ESI-MS/MS experiment. A *bond match* bm_k between D and Ml occurs between bms_i and ml_j when $|bms_i - ml_j| < bm_t$, where msm_t is defined as the *bond mass tolerance*, the ± amount of experimental uncertainty that ml_j is allowed to have to determine the match. A *bond spectrum match* is the set of matches $BSM = \{bsm_k\}$ between *ML* and *BMS*.

In a LC/ESI-MS/MS experiment, each precursor ion undergoes collision-induced disassociation, resulting in fragment ions that constitute a MS/MS spectrum. If the precursor ion is a disulfide bonded pair of peptides, the fragmentation process typically keeps this bond intact. Let $FML = \{fml_i\}$ denote the set of MS/MS values corresponding to the masses of the peptide fragments. A *peptide fragment* F is a substring of a peptide with mass $m(F) = \sum_{r \le i \le s} m(a_i)$, where r and s denote the locations of starting and ending amino acids of the peptide fragment. A *disulfide bonded fragment* $F_{1,2}$ is a pair of peptide fragments F_1 and F_2, with mass $m(F_{1,2}) = m(F_1) + m(F_2) - 2m(H)$. For there to be a disulfide bond between F_1 and F_2, each fragment must contain at least one cysteine. The *disulfide bonded fragment mass space* $FMS = \{fms_{ij}\}$ for two cysteine-containing peptides P_i and P_j is the set of every disulfide bonded fragment mass that can be obtained from these two peptides. A *fragment match* fm_k between *FML* and *FMS* occurs between fml_i and fms_{ij} when $|fms_{ij} - fml_i| < fm_t$, where fm_t is defined as the *MS/MS mass tolerance*, the ± amount of experimental uncertainty that tms_j is allowed to have to determine the match. A *MS/MS spectrum match TSM* is the set of matches $TSM = \{fsm_k\}$ between *FMS* and *FML*. The *match ratio r* is then defined as the number of matches divided by the size of the tandem mass spectrum, i.e. $r = |TSM|/|FMS|$.

Since each match ratio is a measure of how well the LC/ESI-MS/MS experimental data supports the hypothesis of a disulfide bond between two of the cysteines in the protein being analyzed, we denote $r_{i,j}$ as

the match ratio for a bond between cysteines C_1 and C_2. As a result, each r is assigned as the weight w_e of the graph G which models the overall connectivity pattern. Thus, the **disulfide linkage pattern identification problem** is to find a perfect matching in G of maximum weight.

2.2. Algorithmic Framework

Determining the disulfide linkage patterns involves solving the following four sub-problems:

1. Find the *bond spectrum match BSM* between the mass list *ML* and the disulfide bond mass space *BMS*.
2. Determine the *MS/MS spectrum match TSM* between the disulfide bonded fragment mass space *FMS* and the MS/MS mass list *FML*.
3. Find a perfect matching of maximum weight for a fully connected graph with |C| vertices, with edges of weight $r_{i,j}$.
4. Utilize experimental data the contains noise, isotopic variation, neutral loss, and charge state uncertainty to achieve the matchings in sub-problems 1 and 2.

In the following subsections, we present our approach to each of these sub-problems.

2.2.1. *Finding the MS spectrum match*

Let k denote the number of sites where an arbitrary protein A can be cleaved with a certain protease. The construction of the mass space then requires $O(k^2)$ time. This is because the k proteolytic amino acids divide the protein A into k+1 subsequences, leading to $k(k+1)/2$ unique pairs of subsequences that can be formed. For the case of disulfide bonds, we are concerned with forming unique pairs of subsequences from C as opposed to A. Because $C \subseteq A$ for almost all proteins and proteases, the disulfide bond mass space *BMS* is likely to be smaller than the mass space obtained from every peptide in A.

The quadratic time complexity can be further reduced if the data structure used to construct and search D did not require computing the mass of every member of D. The intuition lies in computing the masses of the possible disulfide bonded peptides that are expected to be close in value in the mass spectrum S. This can be done by use of the *expected amino acid mass*, as defined below:

> DEFINITION 1. *Expected Amino Acid Mass m_e.* The weighted mean of A, i.e, $m_e = \sum w_i m(a_i)$, where $\{w_i\}$denotes the relative abundance of each amino acid. Using published values for masses and relative abundances,[19] we obtain m_e = 111.17 Da.

Using this definition, we can predict that the mass of a peptide $m(p) \approx \|p\| m_e + 18$, where $\|p\|$ represents the number of amino acids contained in the peptide. The additional 18 Da was explained in Section 2. Statistically, this is justified to a first approximation because the weighted standard deviation, again using published data[19], is 28.86 Da. Thus, the number of amino acids in the bonded pair of peptides, denoted $\|d_i\|$, can be used to construct *BMS* in such a way that it is approximately mass sorted. This is the motivation for exploring the use of a hash table to construct and search *BMS*.

The hash table is a well known data structure for efficient searching of a data space.[20] If the hash function employed satisfies the assumption of simple uniform hashing, then the expected time to search for an element is $O(1)$. Simple uniform hashing means that, given a hash table T, with $|T|$ buckets, any data element d_i is equally likely to hash into any bucket, independently of where any other element has hashed to. Using the Expected Amino Acid Mass to predict the mass of a peptide, we implement the simple hashing function $h(d_i) = \|d_i\|$ as a first approximation. This results in our algorithm (which we call *MSHashID*) for this sub-problem, to have an overall complexity of $O(|C|^2 + |BMS|)$, where $|BMS|$ is the size of the mass spectrum.

Table 1 presents a toy example illustrating the construction of the hash table. In this example, the three pairs of peptides will be hashed to buckets 10, 12, and 14 respectively.

Let the MS spectrum for peptides of the protein being considered in this example contain a mass peak having the value of m(p) = 1332 Da. Following our approach, this results in an estimated number of amino acids to be 12 ($\|p\|$ = 12). Subsequently, the corresponding bucket in the hash table is accessed.

Table 1. Example showing how hash table is constructed.

1.Given protein sequence and protease	2. Identify cysteine-containing peptides	3. Form all possible pairs of peptides	4. determine number of amino acids in each pair
E**C**^{2}GRNVN**C**^{8}TKAIQ**C**^{14}LDEH, trypsin (cleaves after K and R)	E**C**^{2}GR	E**C**^{2}GR NVN**C**^{8}TK	10
	NVN**C**^{8}TK	E**C**^{2}GR AIQ**C**^{14}LDEH	12
	AIQ**C**^{14}LDEH	NVN**C**^{8}TK AIQ**C**^{14}LDEH	14

In our example, this bucket contains the peptide NVNCTK. The mass of this peptide is then computed, and compared with m(p) to determine if there is a match. Because there is a possibility that another bucket may contain a peptide pair with a matching mass, neighboring buckets (i.e, buckets 11 and 13) are also accessed.

2.2.2. *Finding the MS/MS spectrum match*

Based on experimental observation, when peptides undergo collision-induced dissociation (CID), the fragments produced are mostly either a b-ions (contains the N terminus) or y-ions (contains the C terminus).[3] We have also observed that the disulfide bond remains intact during CID. Let p1 denote a peptide with its possible y-ions y1 and b-ions b1, and similarly y2 and b2 for peptide p2. If p1 and p2 are in a disulfide bond, four types of fragments may occur: y1+y2, y1+b2, b1+y1, and b1+b2. The most convenient way to compute and display the disulfide bonded pair mass space is to generate four tables in which each row represents the mass of an ion of the first peptide and each column represents the mass of an ion of the second peptide. Then, each entry in this *MS/MS mass table* (subsequently referred to as *mass table*) is the sum of its row and column, minus 2m(H) Da. Next, let peptides p1 and p2 consist of m and n amino acid residues, respectively. The first step is to compute the lowest and highest masses m_{min} and m_{max} in the mass table. The former is the first row and first column of the mass table, and the latter is its last row and last column. Also, because the dynamic range of amino acid residue masses is relatively small (about 3.3:1 in the extreme case of tryptophan:glycine), the increase in mass is approximately linear as the values are read "diagonally" from the lowest to the highest value. Thus, given an MS/MS fragment ion mass, it is possible to make an initial estimate of the location of the diagonal *band* of theoretical table masses that are most likely to match this fragment ion mass.

Let s be an MS/MS fragment ion mass peak value. If either $s < m_{min}$ or $s > m_{max}$, the algorithm returns no value. Otherwise, in the second step, we compute the average amino acid residue mass $\overline{m} = (m(\text{p1}) + m(\text{p2}))/(\text{n} + m)$. This is the approximate mass difference between an element and the (up to) four elements that are a "step" away from it. A *step* is defined to be the movement of an index that points from an element to a neighboring element, either vertically or horizontally, in a mass table. Thus, the estimate of the number of steps used to index into the table to locate the band for a particular mass peak is $n_{steps} = s / \overline{m}$. While any continuous path of steps from m_{min} to m_{max} can be used to locate the band, it is simplest to step along the perimeter of the mass table. In this algorithm, we start by stepping "down" along the first column, and then "across" along the last row.

We note that the initial estimate may not index into the actual location of the band. Therefore, we need a strategy to reach the actual location starting from the initial estimate. For relatively short peptides of under one hundred amino acid residues (much longer than usually encountered in tryptic digests), one can simply generate neighboring mass table elements along the path used to index into the table until the band is reached. The location of the band is identified as the index of the element that has the mass closest to s.

Once the location of the band is identified, the remaining elements of the band are generated and compared to s. The second element will be found either directly above, or above and to the right (row=row-1, column=column+k, where k depends on the relative sizes of the peptides) of the first element.

As an example, let the two amino acid sequences be p1 = NVNCTK, and p2 = AIQCLDEH. Table 2 shows all of the possible y- and b-ions that contain a cysteine, as well as the mass of each ion. Note that for y ions, an additional 18 Da are added to the sum of the residue masses. The resulting mass table for the b1 + y2 combination is shown in Table 3.

The algorithm described by this approach, *IndexID*, has a worst case time complexity of O(n + m) to locate the band. However, because this approach usually

indexes into the mass table just a few elements away from the band, the time complexity can be estimated by a constant. Because the band is in general a diagonal along the mass table, generating the band elements has a complexity of $O(\sqrt{nm})$. Since *IndexID* is invoked for each instance of a *FSM spectrum match*, the time complexity of the solution to subproblem 2 for a protein is $O(|FML|\ (\sqrt{nm}))$, where $|FML|$ is the size of the tandem mass spectrum.

Table 2. Example fragment mass space

Peptide	Ion type	Sequence	Mass (Da)
1	y	CTK	351
		NCTK	446
		VNCTK	564
		NVNCTK	678
	b	NVNC	431
		NVNCT	532
		NVNCTK	660
2	y	CLDEH	501
		QCLDEH	639
		IQCLDEH	752
		AIQCLDEH	813
	b	AIQC	316
		AIQCL	429
		AIQCLD	544
		AIQCLDE	673
		AIQCLDEH	810

Table 3. Example mass table.

b1+y2-2	501 CLDEH	639 QCLDEH	752 IQCLDEH	813 AIQCLDEH
431 NVNC	930	1068	1181	1242
532 NVNCT	1031	1169	1282	1343
660 NVNCTK	1159	1297	1410	1471

2.2.3. *Finding a perfect matching of maximum weight for a fully connected graph*

Sub-problem 3, the maximum-weight perfect matching problem, is a well understood problem in graph theory. At present, the best performing algorithm that solves this problem for a fully connected graph was designed by Gabow.[21] This algorithm has a worst-case bound of $O(|C|^3)$.

2.2.4. *Consideration of missed proteolytic cleavages and intra-molecular bonded cysteines*

In the laboratory, a protease used to digest a protein may sometimes miss a cleavage point. For example, a protein with sequence NRDKTA should be digested by trypsin into three peptides: NR, DK, and TA. However, if one cleavage point is missed, two peptides are created: either NRDK and TA, or NR and DKTA. We model this behavior by including the parameter m_{max}, the *maximum number of missed cleavages allowed*.

It can be inferred by induction that a protein with k cleavage sites and a $m_{max} = m$ will digest into $(m + 1)k$ unique peptides, assuming $k >> m$. Note that m_{max} includes all smaller values of missed cleavage levels, e.g., $m_{max} = 2$ includes $m = 1$ and $m = 0$ as well. If m_{max} is small (e.g., three or smaller), missed cleavages can be considered to be a constant multiplicative factor in our time complexity analysis as described earlier.

Since the proteolytic digestion process produces peptides that contain two or more cysteine residues, there is the possibility that intra-molecular bonds may occur, i.e. disulfide bonds exist within a single peptide. These peptides must be included into the mass list *ML*, with mass $m(p) = \sum_i m(a_i) - 2$, if at most one disulfide bond per peptide is considered. The impact on time complexity is simply the larger disulfide bond mass space D, which can be modeled as an additive factor, $f(|P|, |C|, m_{max})$. The disulfide bonded fragment mass space *DF* for an intra-molecular bonded peptide consists of the union of the mass spaces of the possible b-ions and y-ions that can result from its fragmentation. For example, for the peptide ASI**C**QQN**C**QY, the possible b-ions are b1, b2, b3, b8, b9, and b10, and the possible y-ions are y1, y2, y7, y8, y9, and y10. Thus the complexity of the solution to subproblem 2 is increased by an additive factor, $O(|ML| \max[n, m])$.

2.2.5. *Peak finding in the presence of noise*

Using Bioworks software from Thermo-Fisher, the raw data obtained from a LC/ESI-MS/MS analysis of a single protein is converted to a series of of DTA files. The DTA format is very simple; the first line contains the mass of the precursor ion and the peptide charge state as a pair of space separated values. Subsequent lines contain space separated pairs of fragment ion

mass-to-charge ratios (denoted m/z) and intensity values. These lines are sorted in order of increasing m/z. Typically hundreds of DTA files are produced per analysis.

A typical DTA file contains on the order of 10^2 to 10^3 lines of fragment ion information. The intensity vales can range from 1 to the order of 10^7. It is expected that only a fraction (<100) of the measured fragment ion readings are derived from the actual b- and y-related fragments of the precursor ion. Most of the other less intense ion readings reflect instrumental or chemical noise. To account for these effects, our proposed methodology takes the following steps:

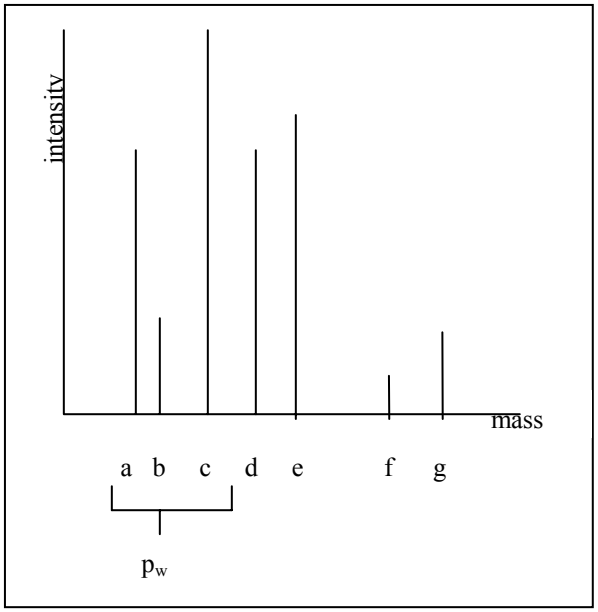
1. We do not consider fragment ion lines with intensity values less than a certain *threshold t*, defined as a percentage of the maximum intensity found in the DTA file.
2. If the number of remaining lines is still large (>100), a *limit l* is placed on the number of peaks to be considered for matching.

Next, we consider the correlation between MS/MS spectrum peaks and the mass/intensity lines in the associated DTA file. In the graphical representation of a MS/MS spectrum the peaks are very sharp. In the DTA file the more intense mass peaks typically occupy several neighboring lines, reflecting the slightly differing masses of the isotopes of a fragmented ion. If each line in a DTA file is considered to be a separate mass peak, the data analysis would be biased towards masses associated with more intense peaks.

To correct for this bias, we represent a set of neighboring lines with similar intensity as a single peak. We formalize the concept of "neighborhood" by defining the *maximum peak width* p_w as the maximum difference in mass-to-charge ratio that two consecutive lines in the DTA file can have and yet be considered to a single peak. "Similar" is defined as the absolute difference in intensity of two neighboring peaks less than 50% of the larger intensity. We denote the set of peaks that result as p_i, where $0 \leq i \leq l$.

Figure 1 illustrates an example of how *threshold t*, *limit l*, and *maximum peak width* p_w work together to find the best mass peaks. Let the masses of the six peaks shown here be labeled a through f, and let t = 10%, l = 2 and p_w be the mass range as shown. Peak c has the maximum intensity, so peak f is eliminated, since its intensity is less than 10% of c's. Because $p_w \geq$ c – a, peaks a, b, and c would have been replaced by a single peak with mass of average mass of these three peaks. However, because the intensity of peak b is less than 50% of peak a, this is not done. Instead, the peak window moves to peaks c, d, and e. In this case, these peaks are replaced by a single peak of mass = (c + d + e)/3. Since the limit is two, this peak and peak a are identified as the peaks to use for subsequent analysis.

Figure 1. Illustrative example of peak finding.



2.2.6. *Addressing isotopic variation and neutral loss*

To account for the possibility of neutral loss, for each element fml_i of the MS/MS mass space *FML* computed in the preceding section, we add three more elements: m(fml_i) + m(H_2O), m(fml_i) + m(NH_3), and m(fml_i) + m(H_2O) + m(NH_3), where m(H_2O) is the mass of a water molecule and m(NH_3) is the mass of an ammonia molecule. This accounting increases the size of the disulfide bonded fragment mass space *FMS* by a factor of four.

In addition, we use the average masses for amino acid residues to compute the mass of peptides and their fragments with molecular weights greater than 1500 Da. Our experiments using an ion trap indicate that this results in more accurate correlations with observed fragment ion peaks than by simply using monoisotopic masses. As a consequence of this step, we empirically observed a much closer correlation between the MS/MS

mass space *FML* and the disulfide bonded fragment mass space *FMS* values.

2.2.7. *Interpretation of peaks given charge state uncertainty*

For some low resolution mass spectrometers, it has been observed that the charge state of the precursor ion used to generate the MS/MS spectra may be reported incorrectly. An incorrect number for the charge state will significantly impact the MS/MS mass space that is searched for matches. To address such cases our system is implemented such that the user can intervene and correct the mass assignment.

Next, we examine how to process the values of fragment ion m/z in the DTA file to obtain the MS/MS mass space *FML* used to search for matches with the disulfide bonded fragment mass space *FMS*. Let the charge state value (reported or corrected) for a DTA file be denoted as c. No fragment of the precursor ion can have a charge larger than c. Then each element of *FML* is obtained by computing $FML_i(z) = zp_i-(z-1)m(H)$, where $1 \le z \le c$ for each z, $1 \le i \le l$, and $m(H)$ is the mass of a single proton. The second term is needed because *FMS* is computed for singly protonated ions.

2.2.8. *Overall complexity*

The overall time complexity of our algorithmic approach is computed as follows:

- Finding the *bond spectrum match BSM* between the mass list *ML* and the disulfide bond mass space *BMS* is performed once per analysis, with a time complexity of $|\mathrm{ML}|O(MSHashID) = O(|\mathrm{ML}|(|C|^2 + |BMS|))$.
- Determining the *MS/MS spectrum match TSM* between the disulfide bonded fragment mass space *FMS* and the MS/MS mass list *FML* is performed each time there is a bond spectrum match, or |BSM| times, with time complexity of $|\mathrm{BSM}|O(IndexID) = O(|\mathrm{BSM}||FML|(\sqrt{nm}))$.
- Finding a perfect matching of maximum weight for a fully connected graph with |C| vertices has a time complexity of $O(|\mathrm{C}|^3)$.
- The techniques developed to utilize experimental data constitute a constant factor multiplying ML and FML.

Thus, the overall complexity of our approach is $O(|\mathrm{ML}|(|C|^2 + |BMS|) + (|\mathrm{BSM}||FML|(\sqrt{nm}) + |C|^3)$. Since n, m and C are typically small (< 100), the performance of this algorithm is dominated by $|ML|$, $|FML|$ and the I/O cost to process the spectrum data.

3. EXPERIMENTAL RESULTS

3.1. Description of the Data and Experimental Procedures

The proposed method was validated utilizing MS and MS/MS data obtained by LC/ESI-MS/MS analysis for three eukaryotic glycosyltransferases with varying numbers of cysteines and disulfide bonds:

1. Mouse Core 2 β1,6-N-Acetylglucosaminyl-tranferase I (C2GnT-I) [22]
2. ST8Sia IV Polysialytransferase (ST8Sia IV) [2]
3. Human Fucosyltransferase VII (FT VII) [23]

The disulfide linkage pattern for each of these proteins is known and reported in each cited reference. The experimental data was obtained using a capillary liquid chromatography system coupled with a Thermo-Fisher LCQ ion trap mass spectrometer LC/ESI-MS/MS system was used to obtain the MS and MS/MS data. Further details of the experimental protocols used are available.[24]

We obtained the primary sequences from the Swiss-Prot database,[25] and DTA files were obtained from LC/ESI-MS/MS analyses of each protein. For each experiment, we set the *bond mass tolerance* bm_t = 3.0 Da, the *maximum peak width* p_w = 2 Da, the *threshold t* = 2% of the maximum intensity, and the *limit* l = 50 peaks. We used *MS/MS mass tolerance* fm_t = 1.0 Da, except when intramolecular bonded cysteines were identified, when a value of 1.5 Da was used. The protease is set to what was used in the actual experiment. We set *maximum number of missed cleavages allowed* m_{max}= 1, except for one case where a combination of trypsin and chymotrypsin was used, where we set m_{max}= 3.

3.2. Summary of Results

The proposed method was applied to determine the disulfide-bonding patterns of three proteins, with varying numbers of cysteines and disulfide bonds. Our results are presented in the form of a connectivity matrix, as proposed in.[26] Each matrix element below the diagonal corresponds to a possible disulfide bond. In this matrix we indicate the "known" linkage patterns by a gray shaded matrix element. If our method computes a match ratio of over 50% for a particular combination,

we record it in the table. In addition, we assign one of the values TP, FP, FN, or TN to each matrix element per the following conventions:

- For match ratios of at least 50%, true positive (TP) is assigned if the same matrix element is shaded gray.
- A false positive (FP) is assigned if the matrix element is not shaded.
- A false negative (FN) is assigned to a matrix element if the matrix element is shaded but its match ratio is less than 50%.

Table 4 summarizes our results for an analysis of 233 DTA files of C2GnT-I. For this dataset, the charge state reported in two DTA files needed to be reinterpreted in order to avoid false negative results. In Table 5 we present the results from the analysis of 79 DTA files of ST8Sia IV, and table 6 contains the results obtained from the analysis of 158 DTA files of FucT VII.

We evaluate the performance using the following metrics:

- Precision P = TP/(TP+FP)
- Recall R = TP/(TP+FN)
- Sensitivity S = TN/(TN+FP)

Table 7 summarizes our results for these metrics. Although our precision result for C2GnT-I is low compared to the precision results for ST8Sia IV and FucT VII, it still compares favorably with the results reported by the purely predictive methods.[11-13] In addition, we note that we can improve the precision from P = 0.40 to P = 0.70 if we chose to ignore all match ratios less than 85%.

Table 4. C2GnT-I 7validation testing results.

Cysteine location	59	100	151	172	199	372	381	413
59								
100	**TN**							
151	**TN**	**TN**						
172	**TN**	**.98 TP**	**TN**					
199	**.84 FP**	**.72 FP**	**TP .96**	**TN**				
372	**TN**	**TN**	**TN**	**TN**	**TN**			
381	**TN**	**.76 FP**	**TN**	**.86 FP**	**TN**	**.86 TP**		
413	**.96 TP**	**TN**	**TN**	**.72 FP**	**.88 FP**	**TN**	**TN**	

Table 5. ST8Sia IV validation testing results.

Cysteine location	142	156	292	356
142				
156	**TN**			
292	**.68 TP**	**TN**		
356	**TN**	**.88 TP**	**TN**	

Table 6. FucT VII validation testing results.

Cysteine location	68	76	211	214	318	321
68						
76	**.94 TP**					
211	**TN**	**TN**				
214	**TN**	**TN**	**.54 TP**			
318	**.TN**	**TN**	**TN**	**TN**		
321	**TN**	**TN**	**TN**	**TN**	**.66 TP**	

Table 7. Overall performance results.

Protein	Precision	Recall	Specificity
C2GnT-I	0.40	1.0	0.75
ST8Sia IV	1.0	1.0	1.0
FT VII	1.0	1.0	1.0

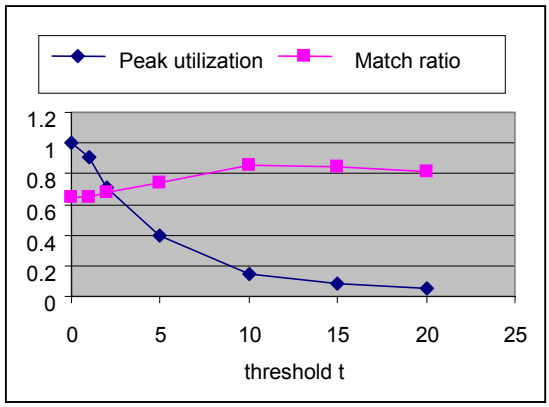
Following the implementations of the purely predictive methodology, we adapted WMATCH, Rothberg's implementation of Gabow's algorithm[14,21] to find the maximum weight matching. This analysis component was only conducted for the C2GnT-I intermediate results, as the linkage patterns for ST8Sia IV and FucT VII are already evident. Our result was in agreement with the published bonding pattern.[22]

3.2.1. *Analysis of the effect of varying threshold t on results*

The values we used for many of the parameters introduced in this paper, such as *threshold t*, *limit l*, and *maximum peak width* p_w, were based on heuristics

developed by experimenters. In this section, we examine the effect of varying the *threshold t* on our results. We used the C156-C356 bond in ST8Sia IV for data. Figure 2 consists of two graphs: (1) a plot of match ratio vs. *t*, and (2) a plot of the fraction of total peaks used vs. *t*. The intersection of these two graphs is close to $t = 2$, confirming that the heuristic value used in our experiments optimizes performance and data utilization.

Figure 2. Match ratio and peak utilization vs. threshold t.



3.2.2. *Comparison with MS2Assign program*

As discussed in Section 1, the program MS2Assign can be configured to process MS/MS data to identify disulfide bonds in protein. However, we note that while MS2Assign automates the identification of disulfide bonds, it does not do so in a high throughput manner. For example:

- The two peptides MS2Assign takes as input must be obtained from another program, such as Peptidemap.
- MS2Assign accepts the input of only one MS/MS mass list (from one DTA file).

Also, because MS2Assign does not account for experimental noice, isotopic variation, or the intensity of the fragmented ion, the accuracy of its results may not be as high as the accuracy of a program that takes these factors into consideration. To investigate this, we identified the DTA files that MS2DB used to obtain match ratios for C13 to C59 (true positive identification) and C199 to C413 (false positive identification) of C2GnT-I. We then copied the fragment ion m/z portion of the file to use for the Peak List in MS2Assign. Our results are summarized in Tables 7 and 8.

Table 7. Comparison of true positive identification.

Program	Number of peaks utilized	Number of matches	Match ratio
MS2Assign	1774	1646	0.93
MS2DB	50	48	0.96

Table 8. Comparison of false positive identification.

Program	Number of peaks utilized	Number of matches	Match ratio
MS2Assign	2169	1791	0.78
MS2DB	50	44	0.72

These studies suggest that MS2DB may be slightly better than MS2Assign at discriminating between a true positive and a false positive result. More studies are needed to support this conclusion.

4. CONCLUSIONS AND DISCUSSION

In this paper we have presented a comprehensive algorithmic framework for the determination of disulfide bonds by utilizing data from tandem mass spectrometry. The proposed approach involves addressing four key sub-problems. First, the match between a given mass spectrum and the set of every possible pair of cysteine-containing peptides of the given protein is obtained. Next, the correspondence between the tandem mass spectrum and the set of every disulfide bonded fragment mass is determined. The actual disulfide connectivity pattern is determined by solving the maximal weight matching problem. The salient contribution of our approach is the use of real-world data from mass spectrometry in the above steps. Doing so, requires addressing a series of algorithmic challenges that include peak finding in noise spectra, addressing issues of isotopic variation and neutral loss, peak interpretation in the presence of charge state uncertainty, consideration of both inter-peptide and intra-peptide bonds, and consideration of missed proteolytic cleavages.

Until now, techniques for disulfide bond identification have tended to remain on either sides of the *model-or-measure* dichotomy. The proposed work seeks to span this divide and identifies the core algorithmic challenges at the intersection of purely computational and purely experimental strategies. Experimental results highlight the high precision and recall that can be obtained with such a hybrid strategy. Another advantage of this approach is its data-driven

and high-throughput nature. An implementation of our approach is available for public use at: http://tintin.sfsu.edu:33191/ms2db/.

Acknowledgments

The research presented in this paper was partially supported by the grants IIS-0644418 and CHE-0619163 of the National Science Foundation, a grant from the Center for Computing in Life Science of San Francisco State University, and the grant P20MD000262 from the NIH. The authors also than the anonymous reviewers for their comments.

References

1. Creighton TE, Zapun A and Darby NJ. Mechanisms and catalysts of disulfide bond formation in proteins. *Trends in biotechnology* 1995; **13**: 18-23.
2. Angata K, Yen TY, El-Battari A, Macher BA, Fukuda M. Unique disulfide bond structures found in ST8Sia IV polysialyltransferase are required for its activity. *J Biol Chem.* 2001; **18**:15369-15377.
3. Gorman JJ, Wallis TP, Pitt JJ. Protein disulfide bond determination by mass spectrometry. *Mass spectrometry reviews* 2002; **21**: 183-216.
4. Brunger, AT. X-ray crystallography and NMR reveal complementary views of structure and dynamics. *Nature structural biology* 1997; **4 Suppl**: 862-865.
5. Klepeis JL, Floudas CA. Prediction of β-sheet topology and disulfide bridges in polypeptides. *J. Comput. Chem.* 2003; **24**:191–208.
6. Taskar B, Chatalbashev V, Koller D, Guestrin C. Learning structured prediction models: A large margin approach. *Proc. of the International Conference on Machine Learning*; 2005.
7. Ferre F, Clote P. Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics* 2005; **21**: 2336-2346.
8. Vullo A, Frasconi P. Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics* 2004; **20**, 653-659.
9. Baldi P, Cheng J, Vullo A. Large-Scale Prediction of Disulphide Bond Connectivity. *Advances in Neural Information Processing Systems* 2004; **17:** 97–104.
10. Tsai CH, Chen BJ, Chan CH, Liu HL, Kao CY, Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. *Bioinformatics* 2005, **21**:4416-4419.
11. DiANNA: http://clavius.bc.edu/~clotelab/DiANNA/
12. DISULFIND: http://disulfind.dsi.unifi.it/
13. PreCys: http://bioinfo.csie.ntu.edu.tw:5433/Disulfide/
14. WMATCH: Solver for the Maximum Weight Matching Problem: http://elib.zib.de/pub/Packages/mathprog/matching/weighted/
15. MS-Bridge: http://prospector.ucsf.edu/prospector
16. X! Protein Disulphide Linkage Modeler: http://www.systemsbiology.ca/x-bang/DisulphideModeler/DisulphideModeler.html
17. Peptidemap: http://prowl.rockefeller.edu/prowl
18. MS2Assign: http://roswell.ca.sandia.gov/~mmyoung/ms2assign.html
19. http://prowl.rockefeller.edu/aainfo/struct.htm
20. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms*, MIT Press, 2001: 224-229.
21. Gabow H. Implementation of Algorithms for Maximum Matching on Nonbipartite Graphs. Ph.D. thesis, Stanford University, 1973.
22. Yen TY, Macher BA, Bryson S, Chang X, Tvaroska I, Tse R, Takeshita S, Lew AM, and Datti A. Highly Conserved Cysteines of Mouse Core 2 1,6-N-Acetyl glucosaminyltransferase I Form a Network of Disulfide Bonds and Include a Thiol That Affects Enzyme Activity. *J Biol Chem.* 2003; **278**:45864-81.
23. De Vries T, Yen T, Joshi RK, Storm J, van den Eijnden DH, Knegtel RMA, Bunschoten H, Joziasse DH, Macher BA. Neighboring cysteine residues in human fucosyltransferase VII are engaged in disulfide bridges, forming small loop structures: a proposed 3D model based on location of cysteines, and threading and homology modeling. *Glycobiology* 2001; **11**:423-432.
24. 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.
25. Swiss-Prot database: http://cs.expasy.org/
26. Fariselli P, Casadio R. Prediction of disulfide connectivity in proteins. *Bioinformatics* 2001; **17**:957-964.