Comparative Analysis of Disulfide Bond Determination Using Computational-Predictive Methods and Mass Spectrometry-Based Algorithmic Approach

Timothy Lee* and Rahul Singh

Department of Computer Science, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132-4025, U.S.A.
{timtlee@sfsu.edu, rsingh}@cs.sfsu.edu

Abstract. Identifying the disulfide bonding pattern in a protein is critical to understanding its structure and function. At the state-of-the-art, a large number of computational strategies have been proposed that predict the disulfide bonding pattern using sequence-level information. Recent past has also seen a spurt in the use of Mass spectrometric (MS) methods in proteomics. Mass spectrometry-based analysis can also be used to determine disulfide bonds. Furthermore, MS methods can work with lower sample purity when compared with x-ray crystallography or NMR. However, without the assistance of computational techniques, MS-based identification of disulfide bonds is time-consuming and complicated. In this paper we present an algorithmic solution to this problem and examine how the proposed method successfully deals with some of the key challenges in mass spectrometry. Using data from the analysis of nine eukaryotic Glycosyltransferases with varying numbers of cysteines and disulfide bonds we perform a detailed comparative analysis between the MS-based approach and a number of computational-predictive methods. These experiments highlight the tradeoffs between these classes of techniques and provide critical insights for further advances in this important problem domain.

1 Introduction

Cysteine residues have a property unique among the amino acids, in that they can pair to form a covalent bond, known as a disulfide bond. These bonds are so named because they occur when each cysteine's sulfhydryl group becomes oxidized following the reaction

$$\text{S-H} + \text{S-H} \rightarrow \text{S-S} + 2\text{H} \quad (1)$$

Because disulfide bonds impose length and angle constraints on the backbone of a protein, knowledge of the location of these bonds significantly constrains the search-space of possible stable tertiary structures which the protein folds into. The disulfide bond pattern of a protein also can have an important effect on its function. For example, in [1] it is shown that the sterical structure formed by disulfide bonds in ST8Sia

* Equal contributors.

M. Elloumi et al. (Eds.): BIRD 2008, CCIS 13, pp. 140–153, 2008.

IV is critical for it to catalyze the polysialylation of NCAM, the neural cell adhesion molecule. NCAM has an important role in neuronal development and regeneration.

At the state-of-the-art, techniques for disulfide bond determination can be classified into three primary groups: (1) Crystallographic techniques producing high-resolution three dimensional structures of proteins, where the disulfide bonds can be observed directly. (2) Algorithmic techniques that *predict* (or infer) the disulfide connectivity based on sequence data. (3) Mass-spectrometry-based techniques that detect disulfide bonded peptides by analyzing a mixture of peptides obtained by targeted digestion of an intact protein.

Crystallographic methods can be used to study a subdomain of the protein that is sufficiently soluble and may form crystals. However, such methods can rarely be used in medium or high-throughput settings. Consequently, in the recent past, significant attention has been given to computational methods that can predict disulfide connectivity based on sequence information alone [2-10, 27]. An important advantage of these predictive methods lies in the fact that they require only sequence-level data to make predictions. Recent results in this area have reported high accuracies with Q_p values (fraction of proteins in the test set with disulfide connectivity correctly predicted) in the 70 – 78% range. These methods also report high Q_c (sensitivity) values. However, in interpreting, extrapolating, and understanding these performance values, the following considerations are especially critical:

1. Most of the reported results use a dataset called SP39 of non-redundant sequences derived from the SWISS-PROT database (release no. 39) proposed in [5]. To ensure quality, this dataset only includes sequences containing information from PDB for which intra-chain disulfide bonds are annotated. Further, disulfide assignments described as "by similarity", "probable", or "potential" are excluded. Two issues emanating from the use of this standard dataset need to be emphasized. On one hand, it undeniably leads to uniformity and ease in comparing results. However, it also invariably leads to methods being optimized in context of a fixed standard. For this reason alone, care needs to be taken in extrapolating the performance on SP39 to arbitrary data. It must be noted, that certain methods (such as [7, 8]) have used multiple datasets in addition to SP39, in assessing performance.
2. In many methods, learning and testing have often been done in a cross-validated settings rather than involving independent datasets. This leaves open the issue of training bias and its possible impact on the performance of these methods on completely novel datasets.
3. In SP39 as well as the other datasets used, only a limited disulfide-bonding topology (consisting of intra-bonded cysteines) is considered. This has putative implications regarding the applicability of these methods to more complex bonding topologies.

In contrast to computational-predictive methods, mass spectrometric approaches, which involve direct measurements, provide a conceptually different approach to disulfide bond determination. The choice between these two classes of methods requires studying the tradeoffs between the *model-and-predict* strategy used in predictive methods and the *direct measurement* principle underlying mass spectrometric techniques. The investigations presented in this paper are motivated by the above

discussion. Specifically, we pursue two goals. *First*, we investigate some of the key computational challenges associated with mass-spectrometry-based disulfide bond determination. *Second*, through experimental studies conducted on nine eukaryotic Glycosyltransferases with varying numbers of cysteines and disulfide bonds, we investigate the aforementioned tradeoffs between purely predictive methods and an MS-based approach.

2 Previous Work

A variety of techniques have been proposed for determining disulfide bonding patterns including crystallographic approaches, computational predictive strategies, and methods combining mass spectrometric and algorithmic techniques. A comprehensive review of algorithmic methods for this problem is presented in [11]. Broadly speaking, algorithmic approaches can be classified into two major classes: (1) techniques that *predict* (or infer) the disulfide connectivity based on sequence data, and (2) techniques that algorithmically analyze a mixture of peptides obtained by targeted digestion of an intact protein using mass spectrometry and thereby seek to *detect* disulfide bonds.

Techniques based on sequence data are based on characterizing a heuristically defined local sequence environment and address one of two correlated problem formulations. The first, *residue classification*, involves distinguishing the bonded cysteine residues from the free residues. Early techniques for residue classification either analyze the statistical frequency of amino acid residues in neighborhoods around the cysteines [12] or encode the local sequence environment of residues and solve the classification problem using machine learning methods in a supervised setting [13, 14]. Other methods [15], have combined the use of both local and global descriptors and/or hybrid classifiers [16]. While it is difficult to directly compare the prediction performance of these methods due to differences in datasets, most descriptor and classifier choices in the aforementioned works lead to prediction accuracies of greater than 78% with [16] reporting prediction accuracy of 87.4% on chains containing two or more cysteines and 88% overall accuracy. Other techniques, such as [12,15] have also reported prediction accuracies in the 82% - 84% range.

The second formulation, *connectivity prediction*, employs techniques that seek to define the complete disulfide connectivity pattern of a protein. In [17], the connectivity pattern was determined by first constructing a completely connected graph *G*. Four different formulations of contact potential were used for weighting the edges and the disulfide connectivity was defined as the solution of the maximum weight perfect matching problem on *G*. In [18], a recursive neural network (RNN) was used for scoring undirected graphs that represent connectivity patterns by their similarity to the correct graph. The idea of RNN formed the basis of the DISULFIND prediction server [19]. In [20] the notion of utilizing the specificities in the sequence neighborhood of cysteines was extended to take advantage of cysteine distributions in secondary structure elements. In [8], the chain classification problem was addressed using evolutionary information and kernel methods. Other approaches to this problem include the use of cysteine separation profiles [9, 10] and comparisons

with an annotated database, as done in the CysView server [22]. The highest Q_p scores (fraction of correctly assigned proteins) reported were in the 70% – 78% range [8, 22].

The basic strategy for determining disulfide bonds using mass spectrometry consists of the following steps: *First*, the protein of interest is cleaved in its non-reduced state between as many of the cysteine residues as possible using proteases like trypsin or chymotrypsin. *Second*, the resultant peptides, including disulfide-linked peptides, are separated and analyzed by electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI). These mass spectrometry techniques allow peptide and protein molecular ions to be put into the gas phase without fragmentation. The analysis is a two step process and involves measuring the mass-to-charge (m/z) ratio of the ionized disulfide-linked peptides (also called the parent or precursor ion) along with measurement of the m/z ratio of the product ions. Subsequently, the peptides' ions are fragmented to confirm the sequence identity of the disulfide-linked peptides and thereby the location of one of the protein's disulfide bonds.

In spite of the seeming simplicity of this process, the determination of disulfide bonds using mass spectra is complex. This is because the number of possible disulfide bonding models grows rapidly with the number of cysteines and the number of expected disulfide bonds. Furthermore, measurements of fragment-based bonding patterns can be influenced by noise and need to be coalesced into an overall connectivity pattern that is physically consistent. These issues constitute the key challenges for an algorithmic approach that seeks to utilize mass-spectrometric data for disulfide-bond determination.

3 Disulfide Bond Determination Using Mass Spectrometry Data

Based on the above discussion, we identify three main computational challenges: (1) efficiently searching the combinatorial space of peptides and fragmented peptides to determine (mass-based) correspondences with the mass spectrum/tandem mass spectrum. These correspondences would indicate putative disulfide bonds. (2) Ranking and filtering the correspondences to exclude effects of noise. (3) Determining the global pattern of disulfide bonds for the molecule.

3.1 Basic Definitions and Computational Formulation

A *cysteine-containing peptide C* is a defined to be a peptide that has at least one of its amino acids identified as a cysteine residue. A *disulfide bonded peptide structure* consists of one or more cysteine-containing peptides that contain one or more disulfide bonds. The *disulfide bond mass space DMS* = $\{Dm_i\}$ is the set of masses of every possible disulfide bonded peptide structure for a protein. During LC/ESI, precursor ions are generated. A *precursor ion* is a peptide or disulfide bonded peptide structure that has been ionized, so that a charge to mass ratio associated with it appears as part of the mass spectrum of a protein. A *precursor ion mass list PML* = $\{Pm_j\}$ is the set of numbers that represent the masses of the precursor ions obtained from a LC/ESI-MS/MS experiment. The *PML* is a representation of the mass spectrum that has been processed to remove noise from the experimental procedure, and

has been expanded to address uncertainties in the charge state of the ion, as well as neutral loss. A *precursor match* between *DMS* and *PML* occurs when the difference between their values is below a predefined threshold. We denote the set of precursor matches as the *Initial Match IM* = between *PML* and *DMS*. During the MS/MS step, peptides undergo collision-induced dissociation (CID), generating peptide fragments. The fragments produced are mostly either b-ions containing the peptide's N-terminus or y-ions containing its C-terminus. A *cysteine-containing peptide fragment* is a peptide fragment that has at least one of its amino acids identified as a cysteine residue. A *disulfide bonded peptide fragment structure* consists of one or more cysteine-containing peptide fragments that contain one or more disulfide bonds. The *disulfide bonded fragment mass space FMS* = $\{Fm_i\}$ is the set of the masses of every disulfide bonded fragment structure that can be obtained from a disulfide bonded peptide structure. A *MS/MS mass list TML* = $\{Tm_i\}$ is the set of numbers corresponding to the masses of the peptide fragments obtained from a precursor ion in a LC/ESI-MS/MS experiment. A *MS/MS match* between *TML* and *FMS* occurs when the difference between the corresponding elements of TML and FMS is less than a predefined threshold. We denote the set of MS/MS matches as the *Confirmed Match CM* between *TML* and *FMS*.

The number of elements in a Confirmed Match is an indication of the degree to which the LC/ESI-MS/MS data supports the presence of a particular disulfide bonded peptide fragment. In our case the identification of the peptide structure shows us which cysteine residues are participating in disulfide bonds. Thus, by aggregating the all the Confirmed Matches for a protein analyzed by LC/ESI-MS/MS, we can arrive at the overall disulfide bond pattern for the protein. In doing so, we need to ensure not only that the overall connectivity pattern is physically consistent (no cysteine participates in more than one disulfide bond) but also that the pattern is the most likely one given the data. The primary challenges for determining the disulfide-bond connectivity therefore include:

1. Finding an efficient way to obtain the *initial match IM* between the *PML* and the *disulfide bond mass space DMS*.
2. For each initial match, efficiently determining the *confirmed match* between the *disulfide bonded fragment mass space FMS* and the *TML*.
3. Aggregating the *confirmed matches* into a weighted graph enabling the computation of the overall disulfide bond pattern.

3.1.1 Determining the Initial Match

We first examine how to construct the DMS for a disulfide bond topology consisting of an arbitrary number of peptides. For this analysis, let k denote the number of sites where an arbitrary protein A can be cleaved with a certain protease. As a result, A is divided into k+1 peptides. For most proteins and proteases, each peptide contains at most one cysteine residue. These peptides can form *interbonded* disulfide bonds with other peptides. If a peptide contains two or more cysteines, an *intrabonded* disulfide bond may be present. For this case, the time needed to construct of the DMS equals the time required to search each peptide to determine which peptides contain two or more cysteine residues. Because there are k+1 peptides, the overall complexity to construct the DMS for this topology is $O(k)$. Extending this line of argument, it can be shown that for the n-peptide case, the mass space requires $O(k^n)$ time to compute.

This complexity can be reduced if the data structure used to construct and search the DMS does not require computing the mass of every member of the DMS. Such a data structure can be constructed by identifying every possible disulfide bonded peptide structure and then storing each as an element in a pre-computed state. For example, if a protein has the three cysteine-containing peptides p1, p2, and p3, this space consists of {p1+p2, p1+p3, p2+p3}. Here, each element contains the amino acid sequence of each unique peptide combination. The next step is to arrange these elements in such a way that they are approximately sorted by mass. This can be done without computing the mass of each peptide combination by noting that the number of amino acids in a peptide is directly proportional to its mass. Based on this idea, we store the DMS in a hash table with its key the number of amino acids in the peptide combination. Next the elements of the PML must be converted to an equivalent number of amino acids in order to index into the DMS for matches. This can be done by use of the *expected match index*, as defined below:

Definition 1. *The Expected Match Index* i_e is defined as the number used to index into a sorted or approximately sorted data structure to arrive at the region where a match is likely to be found. The match index is constructed for mass tables that represent strings of amino acids a by $i_e = m_j/m_e$, where m_j is a value from a mass list and m_e is the *expected amino acid mass*. We defined the expected amino acid mass in [23] as the weighted mean of all 20 amino acids, i.e, $m_e = \sum_i w_i m(a_i)$, where $\{w_i\}$denotes the relative abundance of each amino acid, and $m(a_i)$ is the mass of an amino acid residue. Using published values for masses and relative abundances of each amino acid [24], we obtain m_e = 111.17 Da, with a weighted standard deviation of σ_e = 28.86 Da.

For each member of the PML, an index is calculated by dividing the member by i_e. These indices are then used to access the corresponding buckets in the hash table. Finally, the mass of each peptide pair in a bucket is computed and compared with the corresponding peak value to determine a match. Because only the disulfide bonded peptide configurations that are indexed have their masses computed, we call this technique *generative indexing*. As discussed earlier, the construction of the mass space requires $O(k^p)$ time, where k is the number of cysteine-containing peptides following proteolytic digestion, and p is the maximum number of peptides in a disulfide bonded peptide structure. Thus the overall time complexity of this step is $O(k^p+|DMS|+|PML|)$. In nature, p greater than 3 are rarely observed, and p greater than 5 has not been reported to our knowledge. Consequently the effective complexity of this step is cubic.

3.1.2 Determining the Confirmed Match

Consider a peptide with intrabonded cysteines. For the general case, the total number of y- and b-ions combined is a constant and depends only on the number of amino acid residues in the peptide, denoted $||p||$. Thus, the construction of the disulfide bonded fragment mass space for this case requires $O(||p||)$ time. We note that the expected match index can be used to improve on this time complexity by only considering those elements that are likely to match an element of the TML. For an interbonded pair of peptides, let p1 denote a peptide with its set of possible y-ions denoted y1 and b-ions denoted b1, and y2 and b2 denotes the y-ions and b-ions for peptide p2. Since p1 and p2 are in a disulfide bond, four types of fragments may occur: y1+y2, y1+b2, b1+y1, and b1+b2. A simple way to compute and display the FMS is to generate four tables based on these four types of fragment combinations. Then, for this

MS/MS mass table the mass of any element T[i, j] equals m(i) + m(j) - 2 Da. If the two peptides consist of ||p1|| and ||p2|| amino acid residues, respectively, the total number of elements to compute is (||p1||+1)(||p2||+1). If the ions used to form the mass tables are arranged in order of increasing number of amino acids, the matrices will be sorted. Again, the expected match index can be used to generate only those elements that are likely to match an element of the TML. These elements correspond to a diagonal region in a mass table. This leads to a time complexity to search for a match of O (||p||), if ||p1|| ≈ ||p2||. The extension of this analysis to construct the FMS for a n- peptide disulfide bonded structure is now straightforward. The FMS for a n-peptide structure consists of 2^n n-dimensional sorted tables. Given an expected match index value, the region where matches are likely to be found has n-1 dimensions. Thus, the time complexity to determine a match with an element of the TML is $O(2^n ||p||^{n-1})$. Based on the previous discussion on the number of fragments that have been observed in the disulfide bonded peptide structure, the effective complexity reduces to cubic.

3.1.3 Aggregating Results to Compute the Overall Disulfide Bond Pattern

The output of the previous step is a collection of confirmed matches between pairs of cysteines. Let the *confirmed match* $CM_{a,b}$ denote a match obtained from a disulfide bonded peptide structure with cysteines C_a and C_b. To convert each $CM_{a,b}$ into a single number that is assigned to the weight of an edge between the pair, we apply the notion of the *match ratio r*, which is defined as the number of matches divided by the size of the tandem mass spectrum, i.e. $r = |CM|/|TML|$. To compute the overall disulfide connectivity, we construct a weighted graph G where each vertex represents a cysteine residue in the protein. If there is a match ratio $r_{a,b}$ that is greater than 50%, this number is assigned to the weight of the edge between vertex a and vertex b. Thus each edge represents a Confirmed Match for a disulfide bond between a pair of cysteine residues. Subsequently, the maximal weight matching problem is solved on this graph (using the algorithm by Gabow [25]) to obtain the overall disulfide-bond topology. The complexity of this step is $O(|C|^3)$. This leads to an overall cubic complexity for our method, which we call MS2DB.

4 Experimental Evaluation

The data used in experiments consisted of the primary sequences (obtained from the Swiss-Prot database [24]) and the DTA files obtained from LC/MS/MS analysis using an LCQ ion trap mass spectrometer (Finnigan, San Jose, CA) for nine eukaryotic Glycosyltransferases with varying numbers of cysteines and disulfide bonds. For each protein, all DTA files are used collectively from an LC run. The proposed method was used with the following parameters: bond mass tolerance bm_t = 3.0 Da, maximum peak width p_w = 2 Da, threshold t = 2% of the maximum intensity, and the limit l = 50 peaks. Further, the MS/MS mass tolerance was set to fm_t = 1.0 Da, except when intramolecular bonded cysteines were identified, when a value of 1.5 Da was used. The protease was set to what was used in the actual experiment(s). The number of missed cleavages allowed was set to m_{max}= 1. Three different sets of experiments were performed. In the first experiment the gains in efficiency that are achieved by

utilizing the generative indexing technique were experimentally studied. In the second experiment, the proposed method was compared with MS2Assign [26], which is a mass spectrometry-based method for determining cross-linkages. In the final experiment, a detailed comparative study was conducted where the disulfide connectivity determination capabilities of the proposed mass spectrometry-based method was compared with three well established methods using the model-and-predict methodology, namely DiANNA [7], DISULFIND [27], and PreCys [28]. The results from each of the methods were analyzed in terms of well established statistical metrics of sensitivity, specificity, accuracy, and Matthew's correlation coefficient.

4.1 Experimental Analysis of the Proposed Approach

To quantify the gains in efficiency achieved by utilizing the generative indexing technique, the fraction of the MS mass space that was actually searched for each of the Glycosyltransferases was determined. For this, the number of mass computations was tracked and divided by the total number of entries in the hash table (i.e. the MS mass space). Fig. 1 (left plot) shows the results obtained. It may be noted that the fraction of the mass space that had to be searched decreased as the number of precursor ions increased, thus underlining the effectiveness of the proposed search strategy. For data obtained after the tandem mass spectrometry step, the efficiency gain was measured by dividing the number of mass computations made by the size of the MS/MS mass space, which is essentially the size of the four tables of b- and y-ion combinations. Fig. 1 (right plot) shows that while a larger fraction of the mass space is accessed by a MS/MS mass peak, a saturation level of about 50% is rapidly achieved. This is because the proposed approach saves mass table entries across searches so that the same element is not recomputed.

In order to quantify the ability of the proposed method to efficiently determine the overall bonding pattern, we first must determine the size of the solution space from which the disulfide bond pattern has to be identified. In Table 1, the first column shows the size of this space if there is no knowledge as to whether any one cysteine is bonded with any other. In other words, the cysteine graph for this protein is fully connected. The second column shows the number of possible patterns that are obtained if all edges with match ratios less than .50 are removed in the cysteine graph.

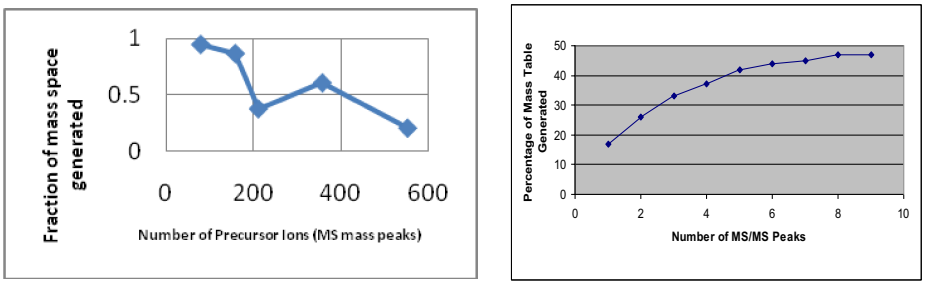

Fig. 1. Experimental analysis of the proposed indexing-based search strategy. The generative indexing approach results in the computation/search of only a fraction of the theoretical mass space.

Table 1. Effectiveness of the proposed approach in reducing the number of disulfide bond patterns that need to be considered for determining the final connectivity

Protein	Number of theoretically possible disulfide bond patterns from fully connected cysteine graph	Number of theoretically possible disulfide bond patterns from cysteine graph with edges for match ratios exceeding 0.50
C2GnT-I	945	67
ST8Sia IV	15	4
FT VII	15	11
Lysozyme	105	61
Lactoglobulin	15	2
FT III	15	10
β1,4-GalT	61	25
Aldolase	124	2
Aspa	124	1

4.2 Comparison with MS2Assign

To compare the proposed method with MS2Assign we identified the DTA files that were used to obtain match ratios for C13 to C59 (true positive identification) and C199 to C413 (false positive identification) of C2GnT-I. The fragment ion m/z portions of the file were then copied to use for the Peak List in MS2Assign. In the true positive identification case, for MS2Assign, the number of matches obtained was 1646 out of 1774 peaks input, giving a match ratio of 0.93. For our method, the corresponding number of matches was 48 out of 50 peaks, giving a match ratio of 0.96. In the false positive identification case, for MS2Assign, the number of matches we obtained was 1791 out of 2169 peaks, giving a match ratio of 0.78, while for our method, the number of matches we obtained was 44 out of 50 peaks, giving a match ratio of 0.72. While preliminary, the results from this study seem to indicate that the accuracy of the proposed approach is indistinguishable from MS2Asssign (the proposed approach performs marginally better in recognizing true positives and scores false positives lower than MS2Assign). However, it should be noted that unlike MS2Assign, the proposed method is fully automated; in MS2Assign the DTA files have to be manually analyzed to identify the mass spectrum and retain the mass values (MS2Assign does not provide such parsing functionality).

4.3 Comparison with Predictive Methods

In this experiment the proposed mass spectrometry-based method was compared with three predictive methods (DiANNA [7], DISULFIND [27], and PreCys [28]) and the results extensively analyzed. The disulfide bonding pattern determined using each method is shown in Table 2. It may be noted that across the entire dataset, using the

Table 2. Dataset and comparison of MS2DB with some prediction servers. The first column displays the name (or abbreviation) of the protein, followed by its Swiss-Prot accession number. Column 3 lists the experimentally determined disulfide bond pattern of each protein. For example for protein C2GnT-I, eight cysteines are engaged in four disulfide bonds, while the remaining three are unbonded. One of these bonds is between the cysteine at amino acid position 59 and the cysteine at amino acid position 413. Columns 4 to 6 show the results for three prediction servers, given the primary sequence of each protein as input. Note that DISULFIND does not support the prediction of proteins with more than 10 cysteines.

Protein (Swiss-Prot ID #)	Number of Cysteines	Disulfide Bond Structure	*DiANNA 1.1*	*DISULFIND*	*PreCys*	MS2DB
C2GnT-I (Q09324)	11	C59-C413 C100-C172 C151-C199 C372-C381 C13, C217, C234 free	C13-C172, C59-C217, C151-C234, C199-C372, C381-C413	Not supported	C59-C381 C100-C372 C151-C172 C199-C413	C59-C413 C100-C172 C151-C199 C372-C381 C13, C217, C234 free
ST8Sia IV (Q92187)	6	C142-C292 C156-C356 C11, C169 free	C11-C156, C142-C292, C169-C356	all free	C142-C356 C156-C292	C142-C292 C156-C356 C11, C169 free
FT VII (Q11130)	6	C68-C76 C211-C214 C318-C321	C68-C321, C76-C211, C214-C318	C76-C318	C68-C76 C211-C214 C318-C321	C68-C76 C211-C214 C318-C321
Lysozyme (P00698)	9	C24-C145 C48-C143 C82-C98 C94-C112 C10 free	C24-C145, C48-C133, C82-C98, C94-C112	C24-C145 C48-C133 C82-C98 C94-C112	C82-C145	C24-C145 C48-C143 C10, C82, C94, C98, C112 free
Lactoglobulin (P02754)	7	C82-C126 C3, C12, C135, C137, C176 free	C12-C137, C82-C176, C126-C135	all free	all free	C82-C126 C3, C12, C135, C137, C176 free
FT III	7	C81-C338 C91-C341, C16, C129, C143 free	C16-C91, C81-C143, C129-C338	none	C81-C91	C81-C338 C91-C341
β1,4-GalT	7	C134-C176 C247-C266 C23, C30, C342 free	C23-C176, C30-C144, C266-C341	none	C134-C247 C176-C266	C134-C176 C247-C266
Aldolase	8	C73, C135, C115, C178, C202, C240, C290, C339 free	C73-C339, C135-C290, C115-C240, C178-C202	none	none	none
Aspa	8	C4, C60, C66, C123, C145, C151, C217, C275 free	C4-C275, C60-C217, C66-C151, C123-C145	none	none	C145-C349
Q_p			0.0	0.0	0.22	0.78

proposed method a Q_p score (representing the fraction of molecules with disulfide bonds correctly identified) of 0.89 was obtained. While DiANNA, DISULFIND, and PreCys are known to perform well on the SP39 dataset, their performance on these nine Glycosyltransferases was significantly inferior.

To further analyze these results, we created connectivity tables for all of the proteins that we studied in a manner similar to what is shown in Table 2. Table 3 is one of the connectivity tables we created. Subsequently the four commonly used metrics of sensitivity, specificity, accuracy, and Matthew's correlation coefficient were calculated.

These four metrics are defined as:

- Sensitivity = TP/(TP+FN)
- Specificity = TN/(TN+FP)
- Accuracy = (TP+TN)/(TP+FP+TN+FN)
- Matthew's correlation coefficient =

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

In the above formulae the following abbreviations are used: TP (true positive), TN(true negative), FP (false positive), and FN (false negative). In seven out of the nine cases, the metrics for the proposed mass spectrometry-based method were superior those of the predictive methods. However, the proposed method had difficulty with Lysozyme, where two disulfide bonds were observed to occur in a complex topology with one inter-peptide bond sandwiched by cysteines participating in an intra-peptide bond. Currently, MS-based methodologies lack the resolution to disambiguate such patterns. Interestingly however, the predictive methods all performed well for this case. It should also be noted that in practice, researchers consider false negative results to have a more deleterious effect on protein characterization than false positive results. Our results, summarized in Table 4, show that MS2DB generates fewer false negative results than the prediction servers we considered.

Table 3. Connectivity table summarizing validation testing results for two proteins. Below diagonal: β1,4-GalT. Above diagonal: Lactoglobulin. The experimentally determined disulfide bond pattern is shaded in gray. The diagonal is shaded black. Only match ratios greater than 0.5 are included into the table.

	3	12	82	122	135	137	176	Cysteine location
23		**TN**	**TN**	**TN**	**TN**	**TN**	**TN**	3
30	**TN**		**TN**	**TN**	**TN**	**TN**	**TN**	12
134	**.71 FP**	**TN**		**TN**	**TN**	**TN**	**.88 TP**	82
176	**.62 FP**	**TN**	**.92 TP**		**TN**	**TN**	**TN**	122
247	**TN**	**TN**	**TN**	**TN**		**TN**	**TN**	135
266	**TN**	**TN**	**TN**	**.6 FP**	**.82 TP**		**TN**	137
342	**TN**	**TN**	**.64 FP**	**TN**	**TN**	**TN**		176
Cysteine location	23	30	134	176	247	266	342	

Table 4. Overall performance results, shown as a collection of sub-tables, one for each protein. The results for the protein C2GnT-I using DiANNA are not reported as proteins with > 10 cysteines are not supported. A zero in the denominator of the performance metric results in it having a value of Undefined.

C2GnT-I	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	0	46	5	4	0.84	0	0.9	-0.09
DISULFIND	> 10 cys							
PreCys	0	47	4	4	0.85	0	0.92	-0.08
MS2DB	4	45	6	0	0.89	1	0.88	0.59
Lysozyme	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	3	31	1	1	0.94	0.75	0.97	0.72
DISULFIND	4	32	0	0	1	1	1	1
PreCys	1	32	0	3	0.92	0.25	1	0.48
MS2DB	2	23	9	2	0.69	0.5	0.72	0.15
Aldolase	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	0	24	4	0	0.86	?	0.86	?
DISULFIND	0	28	0	0	1	?	1	?
PreCys	0	28	0	0	1	?	1	?
MS2DB	0	27	1	0	0.96	?	0.96	?
ASPA	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	0	24	4	0	0.86	?	0.86	?
DISULFIND	0	28	0	0	1	?	1	?
PreCys	0	28	0	0	1	?	1	?
MS2DB	0	28	0	0	1	?	1	?
ST8Sia IV	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	0	9	3	3	0.6	0	0.75	-0.25
DISULFIND	0	13	0	2	0.87	0	1	?
PreCys	0	11	2	2	0.73	0	0.85	-0.15
MS2DB	2	13	0	0	1	1	1	1
FucT VII	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	0	9	3	3	0.6	0	0.75	-0.25
DISULFIND	0	11	1	3	0.73	0	0.92	-0.13
PreCys	3	12	0	0	1	1	1	1
MS2DB	3	12	0	0	1	1	1	1
Lactoglobulin	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	0	17	3	1	0.81	0	0.85	-0.09
DISULFIND	0	20	0	1	0.95	0	1	?
PreCys	0	20	0	1	0.95	0	1	?
MS2DB	1	20	0	0	1	1	1	1
FT III	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	1	17	2	1	0.86	0.5	0.89	0.33
DISULFIND	0	19	0	2	0.9	0	1	?
PreCys	0	19	1	1	0.9	0	0.95	-0.05
MS2DB	4	16	1	0	0.95	1	0.94	0.87
b1,4-GalT	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity	Matthew's Corr. Coeff.
DiANNA	1	17	2	1	0.86	0.5	0.89	0.33
DISULFIND	0	19	0	2	0.9	0	1	?
PreCys	0	17	2	2	0.81	0	0.89	-0.11
MS2DB	2	15	4	0	0.81	1	0.79	0.51

5 Conclusions

In this paper we have presented a comparative analysis of disulfide bond determination using computational-predictive and mass spectrometry-based methods. The proposed mass spectrometry-based method seeks to efficiently search the combinatorial space of possible peptide fragments and find high-quality correspondences with measurements from tandem mass spectra. Subsequently, the correspondence scores

(match ratios) are used to solve a maximal weight matching problem to obtain a globally optimal disulfide bond assignment. This approach contrasts significantly from the core philosophy of computational predictive methods, where the challenge lies in determining the optimal machine learning algorithm, the features to be used, and selection of the training data set. The experimental results show that in general, the direct measurement philosophy underlying mass spectrometry-based methods can outperform the model-and-predict method. At the same time, specificities of protease-dependent digestion combined with specificities of collision-based fragmentation imply that certain bonding topologies can be more reliably discerned using prediction-based methods. To the best of our knowledge, the comparative investigations presented in this paper (and the underlying questions researched) are unique at the current state-of-the-art. We believe that these results provide important cues for future development of both computational-predictive methods as well as mass spectrometry-based algorithmic techniques.

Acknowledgements

This work is supported in part by funding from the NSF grants IIS-0644418 (CAREER) and CHE-0619163, and the Center for Computing for Life Sciences at San Francisco State University. The authors thank Bruce Macher and Ten-Yang Yen for supplying the data we used and for their input they provided in discussions about the method.

References

[1] Angata, K., Yen, T.Y., El-Battari, A., Macher, B.A., Fukuda, M.: Unique disulfide bond structures found in ST8Sia IV polysialyltransferase are required for its activity. J. Biol. Chem. 18, 15369–15377 (2001)

[2] Fariselli, P., Riccobelli, P., Casadio, R.: Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins: Structure, Function, and Genetics 36, 340–346 (1999)

[3] Frasconi, P., Passerini, A., Vullo, A.: A Two-Stage SVM Architecture for Predicting the Disulfide Bonding State of Cysteines. In: Proc. of the IEEE Workshop on Neural Networks for Signal Processing, pp. 25–34 (2002)

[4] Martelli, P.L., Fariselli, P., Malaguti, L., et al.: Prediction of the Disulfide Bonding State of Cysteines in Proteins with Hidden Neural Networks. Protein Engineering 15, 951–953 (2002)

[5] Fariselli, P., Casadio, R.: Prediction of disulfide connectivity in proteins. Bioinformatics 17, 957–964 (2001)

[6] Vullo, A., Frasconi, P.: Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 20, 653–659 (2004)

[7] Ferre, F., Clote, P.: DiANNA: A Web Server for Disulfide Connectivity Prediction. Nucleic Acids Research 33, 230–232 (2005)

[8] Cheng, J., Saigo, H., Baldi, P.: Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins 62, 617–629 (2006)

[9] Zhao, E., et al.: Cysteine Separation Profiles on Protein Sequences Infer Disulfide Connectivity. Bioinformatics 8, 1415–1420 (2005)
[10] Chen, Y.-C., Hwang, J.-K.: Prediction of Disulfide Connectivity from Protein Sequences. Proteins 61, 507–512 (2005)
[11] Singh, R.: A Review of Algorithmic Techniques for Disulfide-Bond Determination. Briefings in Functional Genomics and Proteomics 1(1) (to appear, 2008)
[12] Fiser, A., Simon, I.: Predicting the Oxidation State of Cysteines by Multiple Sequence Alignment. Bioinformatics 16, 251–256 (2000)
[13] Muskal, S.M., Holbrook, S.R., Kim, S.-H.: Prediction of the Disulfide-bonding state of cysteine in proteins. Protein Engineering 3, 667–672 (1990)
[14] Fariselli, P., Riccobelli, P., Casadio, R.: Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins: Structure, Function, and Genetics 36, 340–346 (1999)
[15] Frasconi, P., Passerini, A., Vullo, A.: A Two-Stage SVM Architecture for Predicting the Disulfide Bonding State of Cysteines. In: Proc. of the IEEE Workshop on Neural Networks for Signal Processing, pp. 25–34 (2002)
[16] Martelli, P.L., Fariselli, P., Malaguti, L., et al.: Prediction of the Disulfide Bonding State of Cysteines in Proteins with Hidden Neural Networks. Protein Engineering 15, 951–953 (2002)
[17] Fariselli, P., Casadio, R.: Prediction of disulfide connectivity in proteins. Bioinformatics 17, 957–964 (2001)
[18] Vullo, A., Frasconi, P.: Disulfide connectivity prediction using recursive neural networks and evolutionary information. Bioinformatics 20, 653–659 (2004)
[19] Ceroni, A., Passerini, A., Vullo, A., et al.: DISULFIND: A Disulfide Bonding State and Cysteine Connectivity Prediction Server. Nucleic Acids Research 34, 177–181 (2006)
[20] Ferre, F., Clote, P.: DiANNA: A Web Server for Disulfide Connectivity Prediction. Nucleic Acids Research 33, 230–232 (2005)
[21] Cheng, J., Saigo, H., Baldi, P.: Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins 62, 617–629 (2006)
[22] Lenffer, J., Lai, P., Mejaber, W.-E., et al.: CysView: Protein Classification Based on Cysteine Pairing Patterns. Nucleic Acids Research 32, 350–354 (2004)
[23] Lee, T., Singh, R., Yen, T.Y., Macher, B.: An Algorithmic Approach to Automated High-Throughput Identification of Disulfide Connectivity in Proteins Using Tandem Mass Spectrometry. In: 6th Annual International Conference on Computational Systems Bioinformatics (CSB 2007) (2007)
[24] Swiss-Prot database web site, `http://expasy.org/sprot/`
[25] Gabow, H.: Implementation of Algorithms for Maximum Matching on Nonbipartite Graphs. Ph.D. thesis, Stanford University (1973)
[26] Schilling, B., Row, R.H., Gibson, B.W., et al.: MS2Assign, Automated Assignment and Nomenclature of Tandem Mass Spectra of Chemically Crosslinked Peptides. Journal of American Society of Mass Spectrometry 14, 834–850 (2003)
[27] Ceroni, A., Passerini, A., Vullo, A., et al.: DISULFIND: A Disulfide Bonding State and Cysteine Connectivity Prediction Server. Nucleic Acids Research 34, 177–181 (2006)
[28] Tsai, C.H., Chen, B.J., Chan, C.H., Liu, H.L., Kao, C.Y.: Improving disulfide connectivity prediction with sequential distance between oxidized cysteines. Bioinformatics 21, 4416–4419 (2005)