Reducing Mass Degeneracy in SAR by MS by Stable Isotopic Labeling*

Chris Bailey-Kellogg †  John J. Kelley, III ‡  Cliff Stein †  Bruce Randall Donald † §

Abstract

Mass spectrometry (MS) promises to be an invaluable tool for functional genomics, by supporting low-cost, high-throughput experiments. However, large-scale MS faces the potential problem of mass degeneracy — indistinguishable masses for multiple biopolymer fragments (e.g. from a limited proteolytic digest). This paper studies the tasks of planning and interpreting MS experiments that use selective isotopic labeling, thereby substantially reducing potential mass degeneracy. Our algorithms support an experimental-computational protocol called Structure-Activity Relation by Mass Spectrometry (SAR by MS), for elucidating the function of protein-DNA and protein-protein complexes. SAR by MS enzymatically cleaves a crosslinked complex and analyzes the resulting mass spectrum for mass peaks of hypothesized fragments. Depending on binding mode, some cleavage sites will be shielded; the absence of anticipated peaks implicates corresponding fragments as either part of the interaction region or inaccessible due to conformational change upon binding. Thus different mass spectra provide evidence for different structure-activity relations. We address combinatorial and algorithmic questions in the areas of data analysis (constraining binding mode based on mass signature) and experiment planning (determining an isotopic labeling strategy to reduce mass degeneracy and aid data analysis). We explore the computational complexity of these problems, obtaining upper and lower bounds. We report experimental results from implementations of our algorithms.

Keywords: Mass spectrometry, functional genomics, experiment planning, data analysis, methods for biopolymer structure, protein-protein and DNA-protein complexes.

Introduction

We wish to develop high-throughput algorithms for the structural and functional determination of the proteome. We believe that algorithms can be designed that require data measurements of only a few key biophysical parameters, and these will be obtained from fast, minimal, and cheap experiments. We envision that, after input to computer modeling and analysis algorithms, structure and function of biopolymers can be assayed at a fraction of the time and cost of current methods. Our long-range goal is the structural and functional understanding of biopolymer interactions in systems of significant biochemical as well as pharmacological interest. An example of such computational approaches is the JIGSAW program of Donald and coworkers (Bailey-Kellogg et al. 2000) for high-throughput protein structure determination using NMR.

In this paper, we introduce new computational techniques for experiment planning and data analysis in a methodology called SAR by MS (Structure-Activity Relation by Mass Spectrometry) for use in functional genomics. SAR by MS is a combined experimental-computational protocol in which the function and binding mode of DNA-protein and protein-protein complexes can be assayed quickly. ¹ It uses a combination of accurate mass measurement of degradation products of the analyte complexes, and mathematical algorithms for data analysis and experiment planning, to maximize the information obtained by the mass measurements. In SAR by MS, a complex is first modeled computationally to obtain a set of binding-mode and binding-region hypotheses. Next, the complex is crosslinked and then

¹Biological function is a complex phenomenon. In this paper, we use the term “function” in the very limited sense of structure-activity relation (binding mode and region).
cleaved at predictable sites (using proteases and/or endonucleases), obtaining a series of fragments suitable for MS. Depending on the binding mode, some cleavage sites will be shielded by the crosslinking. Residues exposed in the isolated proteins that become buried upon complex formation are considered to be located either within the interaction regions or inaccessible due to conformational change upon binding. Thus, depending on the function, we will obtain a different mass spectrum. Analysis of the mass spectrum (and perhaps comparison to the spectra of the uncomplexed constituents) permits determination of binding mode and region.

A key issue in SAR by MS is the potential for mass degeneracy: when two potential fragments have approximately the same mass (within the resolution of the spectrum), the existence of one or the other cannot be uniquely inferred from a mass peak. To overcome this problem, we propose the use of computational experiment planning to determine how to selectively manipulate masses (isotopically label) with \(^{13}\)C and \(^{15}\)N enrichment in order to minimize or avoid potential mass degeneracy.\(^2\) Selective isotopic labeling allows, for example, all Leu and Ala residues in a protein to be labeled using either auxotrophic bacterial strains or cell-free synthesis. Mass tags — the mass differences between unlabeled and labeled proteins — can eliminate mass degeneracy by ensuring that potential fragments have distinguishable masses. For example, in Fig. 1, when X and Y are crosslinked, their combined mass is nearly identical to that of Z. By labelling, we ensure that the mass of X+Y is different from Z, thereby allowing SAR by MS to distinguish among the set of binding hypotheses. DNA can also be isotopically labeled, as is illustrated in Fig. 2. Here, we have synthesized and recorded mass spectra for an isotopically-labeled 18-mer (Kelley III 1999; Chen et al. 1999b). The \(^{13}\)C-, \(^{15}\)N-labeled oligonucleotide (top) has a mass tag when compared with its unlabeled counterpart (bottom).

Our work addresses the issue of explicitly planning experiments to minimize mass degeneracy, via the calculation and implementation of specific constraints. We incorporate selective stable isotopic labeling within the analytes. The constraints therefore reflect the partial amino acid content (or nucleotide composition) and the mass-to-charge ratio \((m/z)\) of the analytes. Note that there exist other types of constraints that could be employed in conjunction with stable isotopic labeling. For example: (i) use of a tandem mass spectrometer to generate collision induced dissociation spectra of the (peptide) analytes (Tong et al. 1999; Craig et al. 1999; Link et al. 1999); (ii) use of different enzymes to generate the fragments prior to mass analysis (Bantscheff, Weiss, & Glocker 1999; Hubbard, Beynon, & Thornton 1998); (iii) use of group-specific crosslinkers that would indicate the presence of a (constraining) amino acid in the peptide sequence (Scaloni et al. 1998) (see Ex. 2); (iv) use of a crosslinker that introduces a mass increment that reduces or eliminates mass degeneracy. None of these experimental techniques have been addressed in terms of computational experiment planning, nor as an optimization problem, nor with the goal of automation for eliminating mass degeneracy. It is important to realize that these other methods are informationally orthogonal to stable isotopic labeling. That is, selective labeling will add information content to any of the proposed methods above, by providing very fine-grained control of peptide and oligo masses. Similarly, planning selective labeling can be useful in MS planning.
protocols other than SAR by MS. In this paper, we demonstrate our technique only for SAR by MALDI-TOF MS (Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry); extensions to the methods above are planned in the future.

Experimental techniques relevant to SAR by MS have been studied by a number of researchers. For example, Pucci and coworkers (Scaloni et al. 1998) investigated a combined strategy integrating limited proteolysis and crosslinking experiments with mass spectrometry. It is hypothesized that the interface regions of two interacting proteins are accessible to the solvent in the isolated molecules, but become protected following the formation of the complex (Bantscheff, Weiss, & Glocker 1999; Hubbard, Beynon, & Thornton 1998). Therefore, the interface regions can be inferred from differential peptide maps obtained from limited proteolysis experiments on both the isolated proteins and the complex. Photo- and chemical crosslinking reactions lead to the identification of spatially close amino acid residues in the complex. Mass spectrometry can be employed both to define the cleavage sites and to identify the covalently linked fragments.

In this paper, we first formalize the problem of SAR by MS and mass degeneracy. We then study experiment planning strategies, both for optimizing a single experiment and for combining information across multiple experiments. We prove that, under some fairly natural conditions, an abstraction of the optimal experiment planning problem is NP-complete. We present results from the application of a randomized experiment planning algorithm to the proteins of the complex Ubc9/Ubiquitin-Like Protein Ubl1 (SMT3C). We next address the data analysis problem, introducing an output-sensitive polynomial-time algorithm for data analysis using the technique of spectral differencing. Finally, we present a novel probabilistic framework bridging experiment planning and data analysis, estimating actual mass degeneracy from an analysis of the statistics of hypothesis degeneracy. Some proofs have been omitted for reasons of space.

**Problem Definition**

**Experimental Setup**

We now briefly review some aspects of the experiment design.

**Resolution and Mass Range** MALDI and ESI (Electrospray Ionization) produce gas-phase ions of biomolecules for their analysis by MS. ESI produces a distribution of ions in various charge states, whereas MALDI yields predominantly singly-charged ions. Therefore, ESI spectra are correspondingly more complex. Smith and coworkers (Scalf et al. 1999) have shown how to reduce the charge state of ESI ions, to obtain greatly simplified spectra in which fragments are manifested as single mass peaks (similar to MALDI). The decreased spectral complexity afforded by charge reduction facilitates the analysis of mixtures by ESI MS. While the mass limit for MALDI is about a megadalton, charge-reduction TOF ESI has a mass limit of about 22 kDa. ESI appears to respect weak covalent interactions (such as the hydrogen bonds) (Loo 1997), whereas complexes for MALDI must be covalently crosslinked.

MALDI MS is orders of magnitude better than traditional gel techniques in terms of mass resolution, cycle time, and sample sizes. For example, its mass resolution is one dalton in $10^{-5}$ (or $10^5$ with FT-ICR (Solouki & others 1996)). Indeed, MALDI FT-ICR allows distinguishing reduced vs. oxidized states of Cys residues in large proteins, although to obtain this resolution, depletion of the naturally abundant $^{13}$C and $^{15}$N isotopes is often necessary (Marshall & others 1997). These quantitative differences make SAR by MS an attractive method for high-throughput functional genomics (Scalf et al. 1999; Loo 1997).

**Crosslinking** Crosslinking (the covalent linking of a multimer) is most commonly used for DNA-protein complexes. For protein-protein complexes, a residue can be mutated to a photoreactive amino acid such as p-benzoyl L-phenylalanine (BPA) (Cao & others 1997). After exposure to UV light, the complex is crosslinked. Proteins interact with their substrates on the basis of their 3D fold. If protein complexes are digested, generally the 3D structure of the interacting segments gets distorted or destroyed and the interactions are disrupted. Without crosslinking it is unlikely that the interactions would be preserved in the fragments to be observed by MS. For this reason, we crosslink our complexes, and we restrict our attention in this paper to MALDI MS. It is worth noting that selective isotopic labeling can add information content to ESI MS, in which the experiment planning algorithm would be similar. This is an interesting direction for future work.

**Stable Isotopic Labeling** Uniform and selective labeling of proteins is a standard molecular biology protocol (e.g., for heteronuclear protein NMR). Until recently, the methodology for the uniform and selective labeling of DNA needed to perform these MS experiments was not available. However, recent advances in the enzymatic synthesis of $^{13}$C and $^{15}$N-labeled DNA in milligram quantities have the potential to revolutionize the NMR and MS analysis of nucleic acids (see Fig. 2 and (Chen et al. 1999b)). The feasibility of selective labeling for stable isotope assisted mass spectrometry has been experimentally demonstrated by (Chen et al. 1999a; Chen, Smith, & Bradbury 2000). The experiments were planned and interpreted manually; this paper gives algorithms for automating both processes.

**Computational Model**

This section introduces a mathematical abstraction capturing the essence of the biological problem. In this investigation of SAR by MS, we focus on the problem of determining the binding mode of a protein-protein complex using $^{13}$C- and $^{15}$N-selective label-
ing followed by MS. We defer the problem of planning cleavage strategies, and assume the use of a fixed protease (e.g. trypsin, which cleaves the peptide bond following Lys and Arg residues). We also defer generation of a priori binding mode hypotheses. This type of data is available from several sources, including docking studies such as (Bohm & Klebe 1996; Sridharan, Lilien, & Donald 1999), together with homology searching, DNA footprinting, and mutational analysis. When available, these hypotheses provide priors that restrict the set of fragment interpretations.

**Fragments.** A protein or protein-protein complex is digested by a protease, yielding a set of fragments. There may be many more potential fragments \( F \) than observed fragments \( F^* \) — exposed cleavage sites in the isolated proteins might be inaccessible in the complex, due to incomplete digestion, conformational change upon binding, or shielding within an interaction region. The regions of the primary sequence between adjacent (accessible) cleavage sites are called segments. Protein 1-fragments are formed of sequential unions of segments.

**Example 1.** If a peptide of 20 residues has cleavage sites 5 and 10, then the segments are \((1,5), (6,10), \) and \((11,20). \) The 1-fragments are these 3 segments, plus \((1,10), (1,20), \) and \((6,20).\)

When two interacting proteins are crosslinked and cleaved, a 2-fragment may be formed by the binding of one 1-fragment from each protein. The mass spectrum will then exhibit a peak at the mass of the 2-fragment. 2-fragment masses are not simply the sum of 1-fragment masses, since crosslinking can increase or decrease the mass of both crosslinked and exposed residues. However, since the change is predictable, it can easily be incorporated into our framework and modeled as a mass shift. We take a peak at a 2-fragment mass as evidence that the two constituent 1-fragments are implicated in the interface region of the protein-protein complex. In particular, such a 2-fragment is formed by crosslinking the interface regions, followed by cleavage on each protein strand.

**Example 2.** Consider the interaction of 1-fragments \( \{g_1, g_2, g_1 \cup g_2\} \) of one protein with 1-fragment \( h \) of another. One binding hypothesis is that \( h \) binds \( g_1 \) or \( g_2 \) and the cleavage site \( g_1 \cup g_2 \) is shielded (either by \( h \) or some other fragment). This hypothesis is encoded as the single 2-fragment \( g_1 \cup g_2 \cup h. \) Let \( m(g_i) \) denote the mass of \( g_i, \) etc. If the hypothesis is false, the mass spectrum should contain three peaks \( \{m(g_1), m(g_2), m(h)\}, \) else it should contain one peak \( m(g_1) + m(g_2) + m(h) + \Delta m(g_1, g_2, h), \) where \( \Delta m(g_1, g_2, h) \) denotes the change in mass due to crosslinking.

**Example 3.** Another binding hypothesis for Ex. 2 is that the complex shields the proteolytic site \( g_1 \cup g_2, \) but without \( h \) binding \( g_1 \) or \( g_2. \) This hypothesis is supported by a spectrum containing two peaks \( \{m(g_1) + m(g_2), m(h)\}. \) However, this spectrum could also support other hypotheses (e.g. \( g_1 \) and \( g_2 \) are in the core, shielded from proteolytic digestion). Thus we must compare the spectrum for the cleaved \( g \)-protein in isolation. If the isolated spectrum contains \( \{m(g_1), m(g_2)\} \) then the \( g_1/g_2 \) cleavage site is exposed in isolation and protected (Zappacosta et al. 1996; Cohen et al. 1995) in the complex. Therefore the residues at the \( g_1/g_2 \) site are considered to be located either within the interaction regions in the complex, or inaccessible due to conformational change upon binding. On the other hand, if it contains a peak \( m(g_1) + m(g_2), \) there is no evidence that the \( g_1/g_2 \) site is implicated in the interaction region.

Our algorithm is based on the assumption that the sequence segments responsible for the interactions are (a) contiguous and (b) preserved if the proteins are digested. (a) is assumed strictly for combinatorial reasons. If (a) fails, then our method still works, but with a penalty in combinatorial complexity. Refer to the experimental setup subsection on the use of crosslinking for a discussion of (b).

**Mass Degeneracy** Mass degeneracy results when the masses of two fragments are indistinguishable within the resolution of a particular spectrum. Our goal is to use selective labeling to force the fragment masses to be distinct. A selective labeling scheme uses different isotopes in specific amino acids (e.g. Arg with \(^{15}\text{N} \) instead of \(^{14}\text{N} \)) to affect the resulting mass spectrum.

Given (any) two fragments \( k, l \in F, \) we wish to plan a labeling such that their masses are distinct whenever \( k \neq l. \) That is

\[
\sum_{i \in R} n_{ki}(m_i + x_i) \neq \sum_{i \in R} n_{li}(m_i + x_i), \tag{1}
\]

where \( R \) is the set of residues \( \{\text{Ala, Arg, Asn, Asp...}\} \) plus a “pseudo-residue” term for the appropriate crosslinker (see Ex. 2), \( m_i \) is the unlabeled monoisotopic integer mass of residue type \( i, \) \( x_i \) is the additional mass of residue \( i \) after labeling, and \( n_{ki} \) (resp. \( n_{li} \)) is the number of residues of type \( i \) in fragment \( k \) (resp. \( l \)).

Note that

\[
x_i \in \{0, \hat{c}_i, \hat{n}_i, \hat{c}_i + \hat{n}_i\}, \tag{2}
\]

where \( \hat{c}_i \) and \( \hat{n}_i \) are the additional mass after labeling residue type \( i \) with \(^{13}\text{C} \) and \(^{15}\text{N} \), respectively. Thus, for example, for \( i = 2 \) (Arginine), \( m_2 = 156, \hat{c}_2 = 6, \) and \( \hat{n}_2 = 4. \) Now, let

\[
N_{kl} = (n_{kl} - n_{l1}, n_{k2} - n_{l2}, n_{k3} - n_{l3}, \ldots), \tag{3}
\]

\[
C_{kl} = N_{kl} \cdot (m_1, m_2, \ldots), \tag{4}
\]

\[
X = (x_1, x_2, \ldots). \tag{5}
\]

Then Eq. (1) can be written as the constraint

\[
f_{kl}(X) \neq 0, \quad \text{where} \quad f_{kl}(X) = N_{kl} \cdot X + C_{kl}. \tag{6}
\]

We have a constraint of the form Eq. (6) for every pair of distinct fragments \( k \) and \( l. \) Whenever a constraint \( f_{kl} \) is violated, we obtain mass degeneracy (two fragments with the same mass). This constraint can be expressed as a disjunction of inequality relations (that is, \( \leq \) or \( \geq \)). Inequalities can also enforce peak separation.
in the spectrum. For example, to ensure a peak separation of at least $\delta$, Eq. (6) becomes the disjunction $f_{kl}(X) > \delta$ or $f_{kl}(X) < -\delta$.

**Basic Combinatorics** Let $p = |F|$ be the number of potential fragments after crosslinking and trypsin cleavage, and $n = |R|$ be the size of the set $R$, that is, the number of residue types. Then the number of constraints $m$ of type Eq. (6) is $O(p^2)$. Although in theory the number of residue types. Then the number of potential fragments after crosslinking and trypsin cleavage of at least 5, Eq. (6) becomes the disjunction $3$.

**Computational methods can be employed to form hypotheses about binding modes.** However, in terms of experiment planning, we must be able to distinguish between any pair of hypotheses. Hence, we have $O(s^2) = O(s^3)$ constraints.

It is clear that not all 1-fragment/1-fragment interactions are possible. Some may be excluded based on 1-fragment length. For example, it may be impossible to shield two cleavage sites that are $t$-apart with a single $u$-mer if $u \ll t$. Such reasoning requires careful modeling: for example, the longer strand may be heavily kinked. Computational methods can be employed to form hypotheses about binding modes. These should greatly help the combinatorics, since an experiment would only need to distinguish the fragments identified by hypothesis, and could allow degeneracy in unrelated fragments. In this model, predictions of docking and binding would be made on the computer, and labeling+MS would be performed as a way of screening these hypotheses to test which are correct.

**Experiment Planning**

**Single-Experiment Planning**

The goal of single-experiment planning is to find a labeling $X$ that minimizes the amount of mass degeneracy. To do this, we attempt to minimize the number of constraint violations of the form $f_{kl}(X) = 0$ (refer to Eq. (6)). An exact solution to this optimization problem would find the best labeling—that is, the labeling that minimizes the number of constraint violations, and hence the “amount” of mass degeneracy. An approximate solution would come “close”—for example, within an $(1 + \varepsilon)$ factor of the minimum, for some small $\varepsilon$.

The problem of planning a single-experiment labeling plan can be viewed as an optimization problem. We call this problem OMSEP for OPTIMAL MASS SPECTROMETRY EXPERIMENT PLANNING. Experimentally, OMSEP appears difficult to solve efficiently. OMSEP is an instance of the NP-complete problem MINIMUM UNSATISFYING LINEAR SUBSYSTEM (MULS) (Arora et al. 1997; Feige et al. 1992; Arora & Safra 1992; Arora et al. 1992; Ausiello & others 1999; Halldorsson 1995; Amaldi & Kann 1995). We show that a variant of OMSEP is NP-complete (the proof is in the appendix):

**Lemma 1** OMSEP, even restricted to using only $^{13}\text{C}$ selective labeling, is NP-complete.

**Multiple-Experiment Planning**

The single-experiment planning problem OMSEP is intractable. Even if we could solve it, the resulting labeling might have too much mass degeneracy. Therefore, we pursue a different approach, allowing experiment plans to use several different labelings. First, we explore a necessary condition for experiment planning. Next, we present a stronger, sufficient condition and then discuss how a practical, necessary and sufficient condition may be obtained.

**A Necessary Condition** In the Necessary Condition approach, we label the proteins in several different ways, to produce several samples. MALDI MS is performed on each sample. We do not require that each pair of fragments have distinct masses in every labeling-MS experiment. However, we do require that for every pair of fragments, there exists some labeling in which their masses are distinct.$^4$

Let $L$ be a set of labelings. $L$ may be represented by a set $L = \{X_1, X_2, \ldots\}$ where each $X_i$ is a point of the form $X$ in Eq. (5). For a pair of fragments $k$ and $l$, and a labeling $X \in L$, we can ask whether their masses are distinct under labeling $X$. That is:

$$f_{kl}(X) \neq 0?$$

(The constraint $f_{kl}$ is given in Eq. (6).) Hence, our necessary condition is:

**Feasibility Condition:** Find a set of labelings $L = \{X_1, X_2, \ldots\}$ such that for every pair of fragments $k$ and $l$, either $k = l$ or there exists some labeling $X_{kl} \in L$, such that $f_{kl}(X_{kl}) \neq 0$. We call $L$ a Feasible Set of Labelings.

The Feasibility Condition can be converted into an optimization problem—for example, minimizing the number of experiments or the number of different amino acids labeled in each experiment. Let us focus on the first. The Feasibility Condition requires that we find a set of labelings such that for every pair of fragments, there is at least one labeling in which the pair is not...
mass degenerate. If there are \( p \) fragments, the feasible labeling set \( L \) (when it exists), could be large, which would not be practical. Obviously, the smaller \( p \) is, the better. This leads to the optimization version of our problem, which can be given as follows:

**Labeling-Set Optimization:** Minimize the size \(|L|\) of the Feasible Set of Labelings \( L \).

**Necessary vs. Sufficient Conditions** We say that ambiguity occurs when, in a data spectrum, it is impossible to assign each mass peak to a unique fragment, due to mass degeneracy. This makes it impossible to infer which fragment caused each peak, and therefore we cannot infer which fragments are experimentally present.

**Claim 2** The Feasibility Condition is worst-case necessary and sufficient to eliminate ambiguity in the case \(|L| = 1\).

**Claim 3** For \(|L| > 1\), the Feasibility Condition is necessary but not sufficient.

**Proof:** Necessity is definitional. We show it is not sufficient. Suppose \( L = \{X_1, X_2\} \). Let \( k, g_1, g_2 \) be fragments, and let \( \psi_k(g) \) denote the mass of fragment \( g \) in labeling scheme \( X_k \). Suppose \( \psi_k(g) = \psi_k(g') \), \( \psi_k(g) \neq \psi_k(g') \). Then the Feasibility Condition holds, but it is impossible to assign the \( k-g_1 \) or \( k-g_2 \) peaks. In particular, we cannot guarantee that \( k \)'s presence or absence can be inferred.

**Claim 4** A Sufficient Condition for \(|L| > 1\) is given as follows: Find a set of labelings \( L \) such that for every fragment \( k \), there exists a labeling \( X_k \in L \) such that, for every fragment \( g \neq k \), \( f_{k,g}(X_k) \neq 0 \).

In practice, the sufficient condition in Claim 4 is much stronger than we need. One intuitive reason is the potential for use of negative evidence: the absence of a peak in one labeled spectrum can disambiguate a potential mass degeneracy in another. For example, in the proof of Claim 3, if fragment \( g_1 \) does not occur, then the peak \( \psi_2(g_1) \) will be missing if \( \psi_2^{-1}(\psi_2(g_1)) \) is a singleton. In this case, the \( k-g_1 \) peak in labeling \( X_1 \) can be unambiguously assigned to \( k \). Thus, the sufficient condition does not take into account the expected information content of negative evidence. Note that this assumes that the quantity of a particular fragment is dramatically reduced or completely absent. Since MS is not a quantitative method, a reduction in peak size under some conditions could not be construed as negative evidence. The key point is that we do not require that any peak must be absent: however, when a peak is experimentally absent, the algorithm can exploit that information to make valid inferences about function. Since roughly \( s^4 - s \) fragments will not occur in any experiment, we expect to find a great deal of negative evidence. In the next section, we incorporate negative evidence into the data analysis phase.

More intuition as to why the sufficient condition might be stronger than needed follows from recognizing that the necessary condition imposes \( O(s^8) \) constraints on \( O(s^4) \) fragment hypotheses. However, in any physical experiment, only \( O(s) \) fragments will appear. These fragments are so constrained by the \( O(s^3) \) clauses of the necessary condition, that mass degeneracy under a feasible labeling is rare. The randomized experiment planning algorithm described above can be viewed as "satisfying a necessary condition," as opposed to optimally satisfying a necessary condition (which would mean minimizing \(|L|\)), or satisfying a worst-case sufficient condition like Claim 4 (which would be so pessimistic as to demand a very large number of experiments). Our goal is to minimize or reduce the ambiguity from mass degeneracy in an \( O(s) \)-size sample \( F \) that is selected "randomly" from a larger, \( O(s^4) \)-sized set \( F \) of fragment hypotheses, given statistics on the mass degeneracy in \( F \). In the probabilistic framework section below, we quantitate these observations by modeling the statistical properties of mass degeneracy.

**Experimental Results**

It follows immediately from Lemma 1 that Labeling-Set Optimization is NP-hard. Therefore, we explored how feasibility (without optimality) could be computed (i.e., to obtain a "small" number of unsatisfied constraints), with the randomized algorithm in Table 1. This algorithm merely checks the necessary condition like Claim 4 (which would be so pessimistic as to demand a very large number of experiments). Our goal is to minimize or reduce the ambiguity from mass degeneracy in an \( O(s) \)-size sample \( F \) that is selected "randomly" from a larger, \( O(s^4) \)-sized set \( F \) of fragment hypotheses, given statistics on the mass degeneracy in \( F \). In the probabilistic framework section below, we quantitate these observations by modeling the statistical properties of mass degeneracy.

**Table 1:** Randomized experiment planning algorithm.

| Let \( L = \emptyset \). |
| Let \( D = \mathcal{F} \times \mathcal{F} \). |
| Repeat |
| Let \( X \) = a random labeling. |
| Set \( L \leftarrow L \cup \{X\} \). |
| Set \( D \leftarrow \{(k,l) \in D \mid f_{k,l}(X) = 0\} \). |
| Until \( D = \emptyset \). |

We applied the randomized algorithm to experiment planning for the proteins Ubiquitin Carrier Protein (UBL9)\(^5\) and Ubiquitin-Like Protein (UBL1)\(^6\) under trypsin cleavage. The algorithm was run for 1000 trials, with each trial identifying a set of experiments that disambiguate the fragments. A minimal-sized experiment set (not necessarily unique) was chosen from this group. Two fragments were considered ambiguous if their masses differed by less than one part per thousand. The computation required about three minutes of real time on a 400MHz Pentium II machine, running interpreted Scheme code. Results, detailed in Table 2, show that fragments of UBL1 can be disambiguated with one correctly-chosen isotopic labeling, and fragments of

\(^5\)UBC9 (or Human UBC1), Accession # P50550/Q15698.

\(^6\)UBL1 (or Human SM33), Accession # P55856/Q93068.
Since optimal experiment planning appears generacy, the data analysis may yield ambiguous re-
to a table-lookup. The process is designed to mini-
sign the experiments so that the data analysis devolves
complexes of interest.

hypotheses, which should be sufficient for many com-
feasible fragments in the complex. A reasonable set of
importance of computational modeling and prediction of
of experiments to disambiguate, underscoring the im-
For the UBL1-UBC9 complex, the program identified
UBC9 can be disambiguated with no more than five la-
elings: the first labeling leaves 18 ambiguous pairs, of
which only 10 are ambiguous with respect to the second
labeling, and so forth. In a later section, we calculate
a probabilistic measure of how well these planned ex-
periments are expected to eliminate mass degeneracy
(P(interp)) is the probability that spectral differ-
cencing can eliminate all incorrect fragments (Eq. (14)).

UBC9 can be disambiguated with no more than five la-
elings: the first labeling leaves 18 ambiguous pairs, of
which only 10 are ambiguous with respect to the second
labeling, and so forth. In a later section, we calculate
a probabilistic measure of how well these planned ex-
periments are expected to eliminate mass degeneracy
(P(interp)) is the probability that spectral differ-
cencing can eliminate all incorrect fragments (Eq. (14)).

Given a complex and a fixed cleavage agent, let $S_i$
be a mass spectrum, represented as a set of masses (at
observed peaks) \{s_1, s_2, \ldots\}, under labeling scheme $X_i$.
$X_i = \{x_1, x_2, \ldots\} is a vector of labels as in Eq. (5). Let \phi_i(s) be the set of fragments which could have produced peak s:

$$\phi_i(s) = \{k \in \mathcal{F} | s \approx \psi_i(k)\}$$

where $\psi_i(k)$ is the mass of fragment $k$ under $X_i$. Spect-
ral differencing then identifies pairs of peaks in two
different spectra $S_1$ and $S_2$ such that the same frag-
ment could have caused both peaks. We define the set of
interpretations of the mass shift $(s_1, s_2)$ for peaks
$s_1 \in S_1$ and $s_2 \in S_2$ as $I(s_1, s_2) = \phi_1(s_1) \cap \phi_2(s_2)$. Due
to mass degeneracy, $s_1$ in spectrum $S_1$ could have mul-
tiple explaining fragments $k \in \phi_1(s)$, and $s_2 \in \phi_2(s_2)$, in order to be consistent with the spe-
tral difference. This approach uses negative evidence
to rapidly prune the fragments being considered.

We now develop a fast algorithm for spectral differ-
cencing. The difference spectrum of $S_1$ and $S_2$ is
obtained from the Minkowski difference $S_2 \ominus S_1 =$
\{ $s_2 - s_1 | s_2 \in S_2, s_1 \in S_1 \} as follows. In general, there
will be constraints on which pairs of peaks in $S_1 \times S_2$
can participate in the difference spectrum. In the ex-
ample above, $X_1 = 0$ (i.e., $S_1$ is unlabeled) and $X_2$
contains only positive and zero increments. This means
that all mass shifts must be between 0 and some maxi-
mum value $t$ that depends on the primary sequence (for
example, if all Arginine residues are labeled with $^{15}$N,
then the upper bound $t$ for the mass shift of a fragment
is given by the maximum number of Arg residues in any
fragment times $n_2 = 4$). There will be a lower bound $l$
well (for example, $l = 0$ if there is any Arginine-free
fragment), and in general $l$ and $t$ can be made tighter
by varying as functions of $s_1$. Hence, we define the
difference spectrum as

$$D(S_1, S_2) = \{(s_1, s_2) \in S_1 \times S_2 | s_2 - s_1 \in l([s_1], t(s_1))\}.$$  

Now, suppose a peak $s \in S_1$ is caused by a fragment $f_k$. Following Eq. (5), we obtain

$$s = N(f_k) \cdot (M + X_1).$$  

Hence, $N(f_k)$ is simply the vector encoding the counts of each residue type. Now, because of mass degeneracy, we may also have other fragments $f_2, f_3$, etc. that can cause $s$. That is, $s = N(f_2) \cdot (M + X_1), s = N(f_3) \cdot (M + X_1), s = N(f_3) \cdot (M + X_1).$

<table>
<thead>
<tr>
<th>13C-labeled</th>
<th>15N-labeled</th>
<th>P(interp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlabeled</td>
<td>Unlabeled</td>
<td>0.43</td>
</tr>
<tr>
<td>ARCEGLKSWV</td>
<td>NCQHLSWV</td>
<td>0.88</td>
</tr>
<tr>
<td>QOSWV</td>
<td>ACQGIEHPY</td>
<td>0.99</td>
</tr>
<tr>
<td>ANDGCEHLSV</td>
<td>ACQGILMPSWV</td>
<td>0.9999</td>
</tr>
<tr>
<td>ARNQEHMKSV</td>
<td>ACQGILMWWY</td>
<td>0.99999</td>
</tr>
<tr>
<td>DQELSW</td>
<td>ANSGNLKPTWY</td>
<td>0.9999997</td>
</tr>
</tbody>
</table>

Table 2: Isotopically-labeled experiment planning re-
sults from the randomized algorithm. (a) Single experi-
ment disambiguating fragment masses for UBL1. (b) Se-
quence of experiments collectively disambiguating frag-
ment masses for UBC9. $\chi = number of remaining am-
biguities. P(interp) is the probability that spectral differ-
cencing can eliminate all incorrect fragments (Eq. (14)).
Finally, given two spectra $S_1$ and $S_2$ with labelings $X_1$ and $X_2$, the set $I(D(S_1, S_2))$ represents the fragment hypotheses consistent with the difference spectrum. We now develop an output-sensitive algorithm for computing the consistent fragment hypotheses $I(D(S_1, S_2))$. Consider a dimeric protein complex $P$ with $n$ residues. Given a cleavage agent $\gamma$, we obtain a crosslinked and cleaved system $P(\gamma)$, containing both 1- and 2-fragments. While the set of possible fragments that could make up $P(\gamma)$ is large ($O(n^4)$), in any particular $P(\gamma)$ we will see only $O(n)$ 1-fragments (see the section on basic combinatorics). A priori, there could be $O(n^2)$ 2-fragments, but we do not expect it is geometrically feasible for every pair of 1-fragments to crosslink. Therefore, we expect to observe only $O(n)$ 2-fragments. Hence, we expect the size $c$ of the crosslinked and cleaved system $P(\gamma)$ to be $O(n)$. For larger proteins, in practice, the mass values are only accurate to some uncertainty bound $\varepsilon$. To cope with this uncertainty, we employ 1D range-searching:

**Claim 5** Suppose we are given two spectra $S_1$ and $S_2$ of a crosslinked and cleaved system $P(\gamma)$ with labelings $X_1$ and $X_2$ (respectively), together with a tolerance $\varepsilon$ representing the resolution of the spectra. Then the fragment hypotheses consistent with the tolerated difference spectrum can be computed in output-sensitive time $O(c^2 \log n)$ where $c$ is the size of $P(\gamma)$, using $O(n^4 \log n)$ preprocessing time.

**Proof:** To compute $I(D(S_1, S_2))$, we can create a binary range tree (Bently 1980) storing for each fragment $f$ the interval $[z(f) - \varepsilon, z(f) + \varepsilon]$ and the datum $f$, where $z(f) = N(f) \cdot (X_2 - X_1)$. While building the range tree, we insert only those fragments for which Eq. (9) holds. This preprocessing requires time $O(n^4 \log n)$. Given a potential mass shift, we perform a tree lookup in time $O(\log n)$. The size of the difference spectrum is bounded above by the size of the Minkowski sum $S_2 \ominus S_1$, which is $O(c^2)$. Thus, we do $O(c^2)$ lookups in time $O(c^2 \log n)$.

**Corollary 6** Spectral differencing under uncertainty can be extended to analyze spectra from $d$ selective labeling schemes, with $O(dc^2 \log n)$ running time and $O(dn^4 \log n)$ preprocessing time.

Although the $O(n^4 \log n)$ preprocessing time is non-trivial, we envision it could be done in parallel with the wetlab molecular biology (selective labeling), which can take on the order of days. After preprocessing, the $O(c^2 \log n)$ computational lookup phase should be very fast, on a similar timescale to MS recording.

Spectral differencing can also be used to compare spectra from single proteins against spectra for a complex of the proteins (see Ex. 3). Space limitations prevent detailed discussion, but the algorithm is similar to that above.

**Probabilistic Mass Degeneracy**

The data analysis techniques discussed in the previous section correlate information among multiple spectra from different labelings, overcoming mass degeneracy by eliminating fragment hypotheses that are not consistent with all spectra. Since there are a large number of fragment hypotheses ($O(s^4)$) but only a small number of observed peaks ($O(s)$), it is likely that many potential ambiguities can be resolved by spectral differencing, given experimental data. The experiment planning sufficient condition (Claim 4) operates without experimental data, assuming the worst case, and thus may be far too strict in practice. This section derives probabilistic measures that approximate the likelihood that spectral differencing will be able to resolve potential ambiguities. In particular, we distinguish correct and incorrect fragment hypotheses as those that respectively do and do not correspond to peptides existing in the sample. We then address the following question: How likely is it that some incorrect fragment hypotheses cannot be eliminated due to mass degeneracy with correct fragment hypotheses?

**Claim 7** Spectral differencing fails to eliminate all incorrect fragment hypotheses if and only if there exists an incorrect fragment hypothesis $k$, such that, for each labeling $X \in L$, there exists a correct fragment hypothesis $I_X$ such that $f_{1-1}(X) = 0$.

The negation of the condition in Claim 7 indicates when spectral differencing can eliminate all incorrect fragment hypotheses. Note that this does not mean that all peaks will be uniquely assigned, since the correct fragment hypotheses might be mass degenerate. However, it does mean that exactly the correct hypotheses will be identified, which is our objective. This novel approach of identifying correct answers without relying on assignment has also proved useful in NMR data analysis (Bailey-Kellogg et al. 2000).

**Probabilistic Framework**

To compute the likelihood of satisfying Claim 7 with a given set of labelings $L$, first impose a distribution on the a priori probability that a fragment is correct. For simplicity, we assume here that this is uniform: the expected number of correct hypotheses $p^*$ is $E(|F|)$ divided by the number of possible hypotheses $p = |F|$. An upper bound can be derived by setting the expected number of correct hypotheses $p^*$ to the number of fragments in the completely-digested protein. Any available modeling assumptions can be incorporated into this distribution. In the derivation below, let $\rho = 1 - p^*/p$ denote the fraction of incorrect fragment hypotheses, and let $\psi(f)$ denote the mass of fragment $f$ in labeling.

80 BAILEY-KELLOGG
We say a particular incorrect fragment hypothesis \( f \) appears in a particular experiment \( i \) unless all of the fragment hypotheses with which it would be mass degenerate are also incorrect. Let \( C(f, i) = \psi^{-1}(\psi_i(f)) \) denote the conflict set (mass-degenerate fragments) of fragment \( f \) in experiment \( i \) and \( c(f, i) = |C(f, i)| \) be the size of the conflict set. Then

\[
P(\text{appears}(f, i)) = 1 - \prod_{g \in C(f, i)} P(\text{incorrect}(g)) = 1 - \psi^c(f, i).
\] (12)

We say a particular incorrect fragment hypothesis \( f \) is eliminatable unless for all experiments \( i \in L \), \( f \) appears in \( i \).

\[
P(\text{elim}(f, L)) = 1 - \prod_{i \in L} P(\text{appears}(f, i)) = 1 - \prod_{i \in L} (1 - \psi^c(f, i)).
\] (13)

An incorrect fragment hypothesis \( f \) is uneliminatable when it is not eliminatable.

Finally, a set of labelings \( L \) is interpretable (Claim 7 is unsatisfied) if for all fragments \( f \), \( f \) is not both incorrect and uneliminatable.

\[
P(\text{interpretable}(L)) = \prod_{f \in F} (1 - P(\text{incorrect}(f)) \cdot (1 - P(\text{elim}(f, L)))) = \prod_{f \in F} (1 - p \cdot \prod_{i \in L} (1 - \psi^c(f, i))).
\] (14)

Eq. (14) defines an interpretability metric for a set of labelings, indicating how likely it is that spectral differencing will be able to eliminate all incorrect fragment hypotheses.

**Experimental Results**

We have tested the interpretability metric for the proteins previously discussed. Refer back to Table 2: the last column gives the interpretability metric for both the unlabeled protein and the labeled protein. Note that the metric converges to 1.0 with the addition of more labelings distinguishing more mass-degenerate pairs, demonstrating the power of spectral differencing to combine information across experiments. In the extreme case, when the sufficient condition (Claim 4) is satisfied (as with the planned labeling for UBL1), the metric equals 1.0.

We have also studied the ability of random labeling sets to satisfy the interpretability condition. Fig. 3 shows histograms of the metric for sets of 1, 2, and 5 labelings, with 100 samples generating each histogram. As these plots illustrate, the interpretability metric provides a concrete indication that UBL1 is easier to disambiguate than UBC9. Randomization is able to effectively sample the space of labelings, and our planning algorithm can find sets of labelings that, with high probability, spectral differencing will be able to interpret. Fig. 3 shows empirical evidence that the Randomized Algorithm (Table 1) and the interpretability metric (Eq. (14)) are mutually beneficial, and may be combined in a package for experiment planning to probabilistically eliminate mass degeneracy.

**Conclusion**

MALDI MS is a fast experimental technique requiring subpicomolar sample sizes. It is therefore attractive for high-throughput functional genomics studies. However, the information extracted is rather minimal compared to NMR or Xray crystallography, so a large burden is placed on the algorithmic problems of experiment planning and data analysis. In this paper, we explored the problem of eliminating mass degeneracy in SAR by MS, developing an experiment planning framework that seeks to maximize the information content of an SAR by MS experiment, and an efficient data analysis algorithm that interprets the resulting data. We investigated optimal experiment planning (OMSEP) where the objective is to minimize mass degeneracy, and showed that, under fairly natural conditions, a \(^{13}\text{C}\)-only variant of this problem is NP-complete. We then explored more tractable subclasses, tradeoffs, and implementation experiments. We developed a randomized algorithm that processes across spectra to eliminate mass degeneracy. While this technique appears to be efficient, it does not minimize the number of experiments. We implemented and tested the algorithm in a study of the protein-protein complex Ubiquitin Carrier Protein/Ubiquitin-Like Protein (SMT3C).

On the other hand, if we are given an *a priori* experiment plan, we can use the information content in the difference spectra to track mass shifts. This more sophisticated data analysis can be done efficiently, and we provide an output-sensitive, polynomial time algorithm for the spectral-differencing data analysis. Using spectral differencing, we then derived probabilistic bounds on actual mass degeneracy using an anal-
analysis of the statistics of hypothesis degeneracy. This let us quantitate the effectiveness of the randomized algorithm. Computational experiments on the SMT3C system support our construction of a data-driven necessary and sufficient condition (Eq. (14)) for probabilistic mass degeneracy.

The algorithms and bounds we explored represent first steps in a computational framework for SAR by MS. We believe this will be a dynamic and fruitful area for future research.

References


Scaloni, A.; Miraglia, N.; Orrb, S.; Amodeo, P.; Motta, A.; Maroni, G.; and Pucci, P. 1998. Topol-


Acknowledgements. We would like to thank Xian Chen of the Life Sciences Division of Los Alamos National Labs, and Ryan Lilien, Chris Langmead and all members of Donald Lab for helpful discussions and suggestions.

Appendix

Lower Bounds (Proof of Lemma 1)

We wish to show that OMSEP is a difficult problem, by showing that it is NP-complete. There are several difficulties in proving a real biological or biochemical problem to be NP-hard. First, the number of amino acids is fixed at 20 and the maximum "reasonable" size of a protein is also fixed by nature, so in a complexity-theoretic sense all problems can be solved in constant time. Of course this doesn't capture the observed complexity of these problems. Thus, we will allow the number of amino acids and the length of the protein to be variables. In the case of protein size, this is a standard abstraction that has been used elsewhere. It is less standard for the number of amino acid types, but we believe the combinatorial argument in the problem definition section justifies this abstraction.

There is another way in which an NP-completeness proof may fail to capture true biochemical problems. A biochemical problem may have restrictions on the possible input parameters that don't arise in other types of problems. For example, to show that a problem with a non-negative input parameter x is NP-hard, it is sufficient to show that it is NP-hard when x is restricted to be 0 or 1. However, this might not be sufficient for a biochemical problem in which x is a physical parameter, such as mass, and restricting it to be 0 or 1 leaves a set of problems that are not physically realizable or interesting. Thus the challenge, roughly, is to show that the set of instances which are hard has a non-empty intersection with the set of problems that arise biochemically.

The following problem BIN FLS \(\neq\), (Feasible Linear System with \{0, 1\} variables and \(\neq\) constraints), is known to be NP-complete (Amaldi & Kann 1995; 1998):

**Problem name:**BIN FLS \(\neq\)

Input: \(a_{ij} \in \mathbb{Q}, \ i = 1, \ldots, n, j = 1, \ldots, m\) and \(b_i \in \mathbb{Q}, \ i = 1, \ldots, n\).

Problem definition: Does there exist \(x_j \in \{0, 1\}, \ j = 1, \ldots, m\) such that

\[
\sum_{j=1}^{m} a_{ij} x_j \neq b_i, \ i = 1, \ldots, n
\]

See (Arora et al. 1997; Feige et al. 1992; Arora & Safra 1992; Arora et al. 1992; Ausiello & others 1999; Halldorsson 1995) for other related work on BIN FLS.

**Lemma 8** For every instance of BIN FLS \(\neq\), and any set of \(r_i\) with the size of each \(r_i\) bounded by a polynomial in the original input size, \(i = 1, \ldots, n\), there is an equivalent instance with \(n + m\) variables and \(2n\) inequalities, in which \(n\) of the right hand sides are \(r_i\), \(i = 1, \ldots, n\), and \(n = 0\).

**Proof:** Let the \(n\) additional binary variables be called \(y_1, \ldots, y_n\). Then we form the following system of \(2n\) inequalities. Consider the following modified problem:

\[
\sum_{j=1}^{m} a_{ij} x_j + (r_i - b_i) y_i \neq r_i, \ i = 1, \ldots, n
\]

Since in any satisfying assignment, all the \(y_i\)'s must be 1, this instance is algebraically equivalent to the BIN FLS \(\neq\) one.

Lemma 8 tells us we have the freedom to choose any rational right hand sides; in particular we can choose them as functions of biochemical parameters and still have an NP-complete problem.

We now introduce an variant of OMSEP, in which only \(^{13}\)C selective labeling is permitted. We call this problem \(^{13}\)C-omsep-sat:

**Problem name:** \(^{13}\)C-omsep-sat

Input: \(m\) amino acids \(z_1, \ldots, z_m\), each with \(c_j\) carbons and mass \(m_j\) \((c_j > 0\) and \(m_j > 0\) for proteins\). \(n\) constraints, where a constraint \(i\) can be specified by \(m\) coefficients \(h_{ij}\) where \((h_{i1}, h_{i2}, \ldots, h_{im})\) is the "difference vector" \(N_{ki}\) in Eq. (3) \((h_{ij},\) the \(j\)th element of the vector \(N_{ki}\), corresponds to the difference in the number of residues of amino acid type \(j)\).

**Problem definition:** Each of the \(n\) constraints can be written as

\[
\sum_{j=1}^{m} h_{ij} (c_j x_j + m_j) \neq 0
\]

where \(z_j \in \{0, 1\}\). Can we simultaneously satisfy all the constraints?
Claim 9 13C-omsep-sat is NP-hard.

Proof. The proof is by reduction from BIN FLS \( \neq \). Assume WLOG that \( a_{ij} \in \mathbb{Z} \) (\( i = 1, \ldots, n, j = 1, \ldots, m \)) (if not, multiply both sides of Eq. (15) by \( 1/q \) where \( q \) is the LCM of the denominators of the \( a_{ij} \)). By Lemma 8, we know we can create an instance in which we specify the right hand sides. We will set

\[
b_i = -\sum_{j=1}^{m} \frac{a_{ij} m_j}{c_j}.
\]

(19)

Given such an instance of BIN FLS \( \neq \), we create an instance of 13C-omsep-sat. Note that all \( m_j \) and \( c_j, j = 1, \ldots, m \), are chosen by nature. For each \( j = 1, \ldots, m; i = 1, \ldots, n \), we set

\[
h_{ij} = a_{ij} \prod_{k \neq j} c_k.
\]

(20)

Now let's look at our system of inequalities:

\[
\sum_{j=1}^{m} h_{ij} (c_j x_j + m_j) \neq 0 \quad i = 1, \ldots, n.
\]

(21)

Making the substitutions from the mapping, we get, for \( i = 1, \ldots, n \):

\[
\sum_{j=1}^{m} a_{ij} \left( \prod_{k \neq j} c_k \right) (c_j x_j) + \sum_{j=1}^{m} a_{ij} \left( \prod_{k \neq j} c_k \right) m_j \neq 0
\]

(22)

or

\[
\sum_{j=1}^{m} a_{ij} \left( \prod_{k \neq j} c_k \right) (c_j x_j) - \sum_{j=1}^{m} a_{ij} \left( \prod_{k \neq j} c_k \right) m_j.
\]

(23)

But

\[
\sum_{j=1}^{m} a_{ij} \left( \prod_{k \neq j} c_k \right) (c_j x_j) = \sum_{j=1}^{m} a_{ij} \left( \prod_{k \neq j} c_k \right) x_j
\]

\[
= \left( \prod_{k} c_k \right) \sum_{j=1}^{m} a_{ij} x_j,
\]

(24)

so we can rewrite the inequalities as

\[
\left( \prod_{k} c_k \right) \sum_{j=1}^{m} a_{ij} x_j \neq \left( \prod_{k} c_k \right) \sum_{j=1}^{m} \frac{a_{ij} m_j}{c_j}
\]

(25)

so this system is just the system (15) scaled by \( \left( \prod_{k} c_k \right) \), and so is satisfiable if and only if (15) is. Note that we can add a set of dummy variables and set them to one to obtain the exact form of Lemma 8. If any rational coefficient \( r_i - b_i \) is non-integral, we can clear denominators by multiplying by one over the LCM as described above.

If we let the largest number in the input be \( D \), then the input to BIN FLS \( \neq \) is of size \( O(nm \log D) \). In our problem, the largest number can be as large as \( n!D \), which means that the input is of size \( O(nm(n \log n + \log D)) \), which is just a polynomial blowup.

Problem name: 13C-omsep

Input: Identical to 13C-omsep-sat. The constraints are again given in the form of Eq. (18).

Problem definition: Can we find a set of assignments \( x_j \in \{0, 1\}, (j = 1, \ldots, m) \) that minimizes the number of unsatisfied constraints?

Lemma 1 13C-omsep is NP-complete.

Proof: NP-hardness follows directly from Claim 9. 13C-omsep is in NP because it is an instance of the NP-problem Minimum Unsatisfying Linear Subsystem (MULS) (Arora et al. 1997; Feige et al. 1992; Arora et al. 1992; Ausiello & others 1999; Halldorsson 1995; Amaldi & Kann 1995). \( \square \)

We have thus shown that the problem of determining whether a set of mass degeneracy constraints is simultaneously satisfiable is NP-hard. Recall that each constraint is generated by a pair of fragment hypotheses, and each fragment participates in many constraints. It is therefore natural to ask whether there exists a real protein that could actually generate exactly the constraints that arise in our reductions. If we take the view that all pairs of fragments potentially interact, and we don't know a priori which ones will interact, then we cannot answer this question. If, on the other hand, as discussed in the introduction, we use a priori binding-mode and -region hypotheses to limit the constraints that the planner must address, the situation is different. Then we can construct a protein for a set of constraints by generating, for each constraint \( i = 1, \ldots, n \), a pair of fragments with the appropriate mass difference given by the difference vector \( N_{kl} \) in Eq. (3), namely \( (h_{11}, h_{12}, \ldots, h_{im}) \). We then focus experiment planning on only the designated pairs by using an input set of a priori hypotheses eliminating from consideration other pairwise fragment-fragment constraints.

It is worth asking whether such a reduction is biologically relevant. It may be unlikely that such a protein will be expressed naturally in the proteome of an organism. However, making such a protein is certainly within the capability of standard biotechnology (where, given any de novo, designed primary sequence, the techniques of standard recombinant DNA, protein overexpression, and purification can be used to produce a sample). Until a distribution of "hard" vs. "easy" naturally occurring proteins can be obtained, we feel the result of Lemma 1, which is realizable biotechnologically, provides insight into the empirically observed combinatorial difficulty of the problem.