

# SAR by MS for Functional Genomics\*

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

October 4, 1999

## Dartmouth Computer Science Technical Report No. PCS TR-99-359

**Abstract:** Large-scale functional genomics will require fast, high-throughput experimental techniques, coupled with sophisticated computer algorithms for data analysis and experiment planning. In this paper, we introduce a combined experimental-computational protocol called *Structure-Activity Relation by Mass Spectrometry (SAR by MS)*, which can be used to elucidate the function of protein-DNA or protein-protein complexes. We present algorithms for SAR by MS and analyze their complexity. Carefully-designed Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight (MALDI TOF) and Electrospray Ionization (ESI) assays require only femtomolar samples, take only microseconds per spectrum to record, enjoy a resolution of up to one dalton in  $10^6$ , and (in the case of MALDI) can operate on protein complexes up to a megadalton in mass. Hence, the technique is attractive for high-throughput functional genomics.

In SAR by MS, selected residues or nucleosides are  $^2\text{H}$ -  $^{13}\text{C}$ -, and/or  $^{15}\text{N}$ -labeled. Second, the complex is crosslinked. Third, the complex is cleaved with proteases and/or endonucleases. Depending on the binding mode, some cleavage sites will be shielded by the crosslinking. Finally, a mass spectrum of the resulting fragments is obtained and analyzed. The last step is the *Data Analysis* phase, in which the mass signatures are interpreted to obtain constraints on the functional binding mode. *Experiment Planning* entails deciding what labeling strategy and cleaving agents to employ, so as to minimize mass degeneracy and spectral overlap, in order that the constraints derived in data analysis yield a small number of binding hypotheses.

A number of combinatorial and algorithmic questions arise in deriving algorithms for both Experiment Planning and Data Analysis. We explore the complexity of these problems, obtaining upper and lower bounds. Experimental results are reported from an implementation of our algorithms.

---

\*This research is supported by the following grants to B.R.D. from the National Science Foundation: NSF IIS-9906790, NSF EIA-9901407, NSF 9802068, NSF CDA-9726389, NSF EIA-9818299, NSF CISE/CDA-9805548, NSF IRI-9896020, NSF IRI-9530785, and by an equipment grant from Microsoft Research. C.S. is supported by NSF Career Award CCR-9624828 and an Alfred P. Sloan Foundation Fellowship.

<sup>†</sup>Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Email: [brd@cs.dartmouth.edu](mailto:brd@cs.dartmouth.edu)

<sup>‡</sup>Dartmouth Computer Science Department, Hanover, NH 03755, USA.

<sup>§</sup>Dartmouth Chemistry Department, Hanover, NH 03755, USA.

# 1 Introduction

We wish to develop high-throughput algorithms for structural and functional determination of the proteome. We hope that algorithms can be designed that require data measurements of only a few key biophysical parameters, and these will be obtained from fast, minimal, 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. To this end we introduce a new method, 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. Data from MALDI-TOF MS (Matrix-Assisted Laser Desorption/Ionization Time-Of-Flight Mass Spectrometry) and ESI (Electrospray Ionization) experiments are employed. Given *a priori* binding-mode and -region hypotheses, computational methods are used to plan selective  $^{13}\text{C}$ - and  $^{15}\text{N}$ -labeling of proteins and DNA so that subsequent MS (Mass Spectrometry) assays of cleaved complexes may distinguish among the set of binding hypotheses. We investigate the complexity of computationally planning a labeling strategy to yield unique mass signatures for cleaved fragments under each binding hypothesis.

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 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. Thus, depending on the function, we will obtain a different mass spectrum. MALDI-TOF mass spectrometry techniques can distinguish masses to within one dalton in  $10^6$  [18]. These techniques are so sensitive that reduced vs. oxidized states of Cys residues can be distinguished in large proteins, although to obtain this resolution, depletion of the naturally abundant  $^{13}\text{C}$  and  $^{15}\text{N}$  isotopes is often necessary [16]. We can also manipulate the mass by  $^2\text{H}$ -,  $^{13}\text{C}$ -, and  $^{15}\text{N}$ -*enrichment* of oligonucleotide and amino acid sequences. The enrichment can be selective (for example, all Leu and Ala residues in a protein can be labeled using either auxotrophic bacterial strains or cell-free synthesis. DNA can also be selectively labeled [12]). This paper reports on computational methods for analyzing the protein and DNA sequences in order to plan selective labeling of the proteins and DNA so that subsequent MS assays of the cleaved complexes are guaranteed to discriminate among the binding hypotheses. Hence, only certain residue and nucleoside types will be isotopically labeled, and the labeling+cleavage plan should result in no *mass degeneracy* — that is, the mass signature of the MS assay will be distinct, for every pair of functional hypotheses. Such MS assays require only femtomolar sample amounts, and take only microseconds per spectrum to record. At a qualitative level, SAR by MS is similar to traditional molecular biology techniques using restriction enzymes and gels for DNA (or proteases and gels for proteins). Both techniques require expression and purification of the biopolymer. However, the mass-resolution, cycle time, and sample sizes are orders of magnitude more favorable using MALDI/ESI MS. Therefore, these quantitative differences make SAR by MS an attractive method for high-throughput functional genomics [17, 15]. The hope is that, with an appropriate algorithmic framework, the technique could eventually scale to multi-protein complexes with masses up to hundreds of kilodaltons (kDa). This paper presents initial steps in such a framework.

To begin an investigation of SAR by MS, we defer the problem of planning cleavage strategies, assuming that a fixed library of proteases/endonucleases are employed. Generation of initial or *a priori* binding mode hypotheses is not addressed in this paper, although we envision that docking studies such as [10, 19] can be employed, together with homology searching, DNA footprinting, and

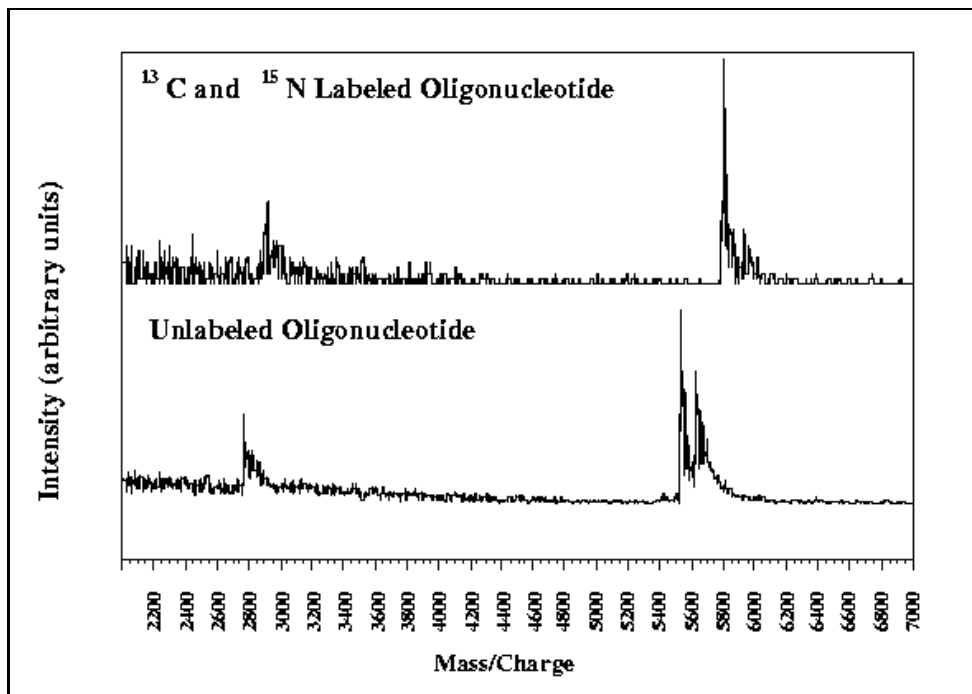


Figure 1: MALDI-TOF mass spectra of an 18bp DNA duplex d(GACATTTGCGGTTAGGTC)\*-d(CTGTAAACGCCAATCCAG). (Top) demonstrates the incorporation of  $^{13}\text{C}$  and  $^{15}\text{N}$  into the deoxyribose and heterocyclic base of each nucleotide in the 18 bp oligonucleotide. (Bottom) displays the lower molecular weight spectrum of an identical 18 bp oligonucleotide with natural isotopic abundance ( $^{12}\text{C}$  and  $^{14}\text{N}$ ). The rightmost peaks ( $> 5400$ ) correspond to the DNA strands. Note the mass shift between the labeled and unlabeled oligos.

mutational analysis. When available, these hypotheses provide priors that restrict the set of fragment interpretations. Assuming an arbitrary (possibly empty) set of initial binding-mode hypotheses, the computational problem of planning the labeling strategies can be viewed as a geometric arrangement problem in a high-dimensional configuration space. A number of interesting combinatorial problems arise, and we derive upper and lower bounds. We explore an optimization strategy for SAR by MS, a randomized approximation algorithm, a data analysis algorithm, and a novel probabilistic framework for quantifying, predicting, and eliminating mass degeneracy through experiment planning. To do this, we first formalize optimal experiment planning in such a way as to minimize mass degeneracy. The optimization version of the problem appears to be difficult, and we prove that, under some fairly natural conditions, an abstraction of the problem is NP-complete. We then consider subclasses of the problem (feasibility vs. optimality) using a randomized approach. We consider tradeoffs between the computational phases of experiment planning and data analysis. Although optimal experiment planning appears intractable, we show how, given a fixed experiment, the technique of spectral differencing yields an output-sensitive polynomial-time algorithm for data analysis. Using spectral differencing, we then derive probabilistic bounds on actual mass degeneracy from an analysis of the statistics of hypothesis degeneracy. Finally, we present experimental studies for labeling and SAR by MS of the protein-protein complex Ubiquitin Carrier Protein UBC9/Ubiquitin-Like Protein UBL1 (SMT3C). The experiments support the theory. Some proofs have been omitted for reasons of space.

## 2 Experiment Planning

### 2.1 Experimental Setup

We now briefly review some aspects of the experiment design.

**Resolution and Mass Range.** MALDI and ESI 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 [17] 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. However, ESI appears to respect weak covalent interactions (such as the hydrogen bonds) [15], whereas complexes for MALDI must be covalently crosslinked.

**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) [11]. After exposure to UV light, the complex is crosslinked. Below 22 kDa, one alternative to crosslinking is to use ESI MS, which seems to respect hydrogen bonds so that bound fragments will have a spectral peak at the sum of their masses [15].

**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  $^{13}\text{C}$  and  $^{15}\text{N}$ -labeled DNA in milligram quantities have the potential to revolutionize the NMR and MS analysis of nucleic acids (see Fig. 1 and [12]).

## 2.2 Experiment Planning as an Optimization Problem

The problem of planning labeling can be viewed as an optimization problem. We call this problem OMSEP for OPTIMAL MASS SPECTROMETRY EXPERIMENT PLANNING. Consider the simplified problem of determining the binding mode of a protein-protein complex using  $^{13}\text{C}$ - and  $^{15}\text{N}$ -selective labeling followed by MS. For simplicity, we will consider only one, fixed cleavage agent, the protease Trypsin, which cleaves the peptide bond following Lys and Arg residues. Given (any) two fragments  $k$  and  $l$ , we wish to plan a labeling such that their masses are distinct whenever  $k \neq l$ . That is

$$\sum_{i \in R} n_{ki}(m_i + x_i) \neq \sum_{i \in R} n_{li}(m_i + x_i), \quad (1)$$

where  $R$  is the set of residues  $\{\text{Ala, Arg, Asn, Asp, } \dots\}$ ,  $m_i$  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  $x_i \in \{0, \hat{c}_i, \hat{n}_i, \hat{c}_i + \hat{n}_i\}$ , where  $\hat{c}_i$  is the additional mass after labeling residue type  $i$  with  $^{13}\text{C}$  and  $\hat{n}_i$  is the additional mass after labeling residue type  $i$  with  $^{15}\text{N}$ . Thus, for example, for  $i = 2$  (Arginine),  $m_2 = 156$ ,  $\hat{c}_2 = 6$ , and  $\hat{n}_2 = 4$ . Now, let

$$N_{kl} = (n_{k1} - n_{l1}, n_{k2} - n_{l2}, n_{k3} - n_{l3}, \dots), \quad C_{kl} = N_{kl} \cdot (m_1, m_2, \dots), \quad X = (x_1, x_2, \dots). \quad (2)$$

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

$$f_{kl}(X) \neq 0, \quad \text{where } f_{kl}(X) = N_{kl} \cdot X + C_{kl}. \quad (3)$$

We have a constraint of the form Eq. (3) 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). Our goal 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  $f_{kl}(X) = 0$ . 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 constraint (3) can be expressed as a disjunction of inequality relations (that is,  $<$  or  $>$ ). Inequalities can also enforce peak separation in the spectrum. For example, to ensure a peak separation

of at least  $\delta$ , Eq. (3) becomes the disjunction<sup>1</sup>  $f_{kl}(X) > \delta$  or  $f_{kl}(X) < -\delta$ .

Computational experiments that we implemented to search for the minimum-degeneracy labeling suggested that the problem may be difficult to solve efficiently. In particular, the initial heuristic search algorithm we implemented could run for a very long time without finding a labeling. This motivates a careful investigation of the complexity, and the need for better algorithms.

**Basic Combinatorics.** A protein or protein-protein complex is digested by a protease, yielding a set of *fragments*. Due to incomplete digestion or shielding, there may be many more potential fragments than observed fragments. In particular, we consider here the case of *sequential unions*, where two or more sequential fragments remain joined rather than being cleaved at the anticipated cleavage site. *Mass degeneracy* results when the masses of two fragments are indistinguishable with respect to the resolution of a particular spectrum. Finally, a selective *labeling* scheme uses different isotopes in specific amino acids (e.g. Arg with <sup>15</sup>N instead of <sup>14</sup>N) to affect the resulting mass spectrum.

Let  $p$  be the number of 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. (3) is  $O(p^2)$ . Although in theory  $n$  is bounded by a constant of about 20, exhaustive search is not possible, since there are approximately  $4^n$  different labeling schemes ( $8^n$  with <sup>2</sup>H-labeling). We begin by treating  $n$  and  $m$  as parameters that measure the input complexity of the problem. To bound the number of fragments,  $p$ , we consider a 2-protein complex, in which each protein has  $s$  cleavage sites. Each site can (potentially) be shielded from cleavage when it is spatially near the protein-protein interaction site. The regions of the primary sequence between adjacent cleavage sites are called *segments*. Protein *1-fragments* are formed of *sequential unions* of segments.

**Example:** *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). Thus, a protein with  $s$  cleavage sites can have  $O(s^2)$  1-fragments.*

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. We take this peak as evidence that the two constituent 1-fragments are implicated in the active site of the protein-protein complex. In particular, such a 2-fragment is formed by crosslinking the active sites, followed by cleavage on each protein strand. Thus we obtain  $p = O(s^4)$  2-fragments. A *fragment* is defined to be any  $i$ -fragment ( $i = 1, 2$ ). Now, in any MS experiment, we will only see peaks from some of these fragments. These are because the fragments may represent competing (mutually exclusive) hypotheses about binding modes. However, in terms of experiment planning, we must be able to distinguish between any pair of hypothesis. Hence, we have  $m = O(p^2) = O(s^8)$  constraints.

**Example:** *Consider the interaction of the 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  $h$  will bind and shield the cleavage site  $g_1/g_2$ . This hypothesis is encoded as the single 2-fragment  $g_1 \cup g_2 \cup h$ . Let  $m(g_1)$  denote the mass of  $g_1$ , etc. If the hypothesis is false, we should see a mass spectrum with three peaks,  $\{m(g_1), m(g_2), m(h)\}$ . If it is true, we should see a single peak, at  $m(g_1) + m(g_2) + m(h)$ .*

Our goal is to use selective labeling to force the fragment masses to be distinct.

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.

---

<sup>1</sup>In practice, mass degeneracy is given in parts per thousand, not as constant. We can encode this by making  $\delta$  dependent on  $k$  and  $l$ , and rewriting this equation as  $f_{kl}(X) > \delta_{kl}$  or  $f_{kl}(X) < -\delta_{kl}$ .

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.

**Lower Bounds.** Experimentally, OMSEP appears difficult to solve efficiently. OMSEP is an instance of the NP-complete problem MINIMUM UNSATISFYING LINEAR SUBSYSTEM (MULS) [5, 13, 7, 6, 8, 14, 3]. We show that a variant of OMSEP is NP-complete: (the proof is in the appendix)

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

## 2.3 Satisficing Instead of Optimizing

Since the optimization problem OMSEP in Sec. 2.2 is intractable, we pursue a different approach. Instead of using only one labeling, we investigate experiment plans with 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.

### 2.3.1 A Necessary Condition

In the Necessary Condition approach, we label the proteins in several different ways, to produce several samples. MALDI or ESI 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.<sup>2</sup>

Let  $L$  be a set of labelings.  $L$  may be represented by a set  $L = \{\ell_1, \ell_2, \dots\}$  where each  $\ell_i$  is a point of the form  $X$  in Eq. (2). For a pair of fragments  $k$  and  $l$ , and a labeling  $\ell \in L$ , we can ask whether their masses are distinct under labeling  $\ell$ . That is:  $f_{kl}(\ell) \neq 0$ ? The constraint  $f_{kl}$  is given in Eq. (3). Hence, our necessary condition is:

**Feasibility Condition:** *Find a set of labelings  $L = \{\ell_1, \ell_2, \dots\}$  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. Let  $|L|$  be the size of  $L$  (the number of labelings required):

**Optimization:** *Minimize the size  $|L|$  of the Feasible Set of Labelings  $L$ .*

It follows from Lemma 1 that the optimization version of this problem is NP-hard. Therefore, we explored how feasibility (without optimality) could be computed (i.e., to obtain a “small” number of unsatisfied constraints) using the primary sequences for Ubiquitin Carrier Protein (UBC9) and Ubiquitin-Like Protein (UBL1) under trypsin cleavage, with the following algorithm:

**Randomized Algorithm.** First, a random labeling is chosen. If two trypsin-cleaved fragments are mass-degenerate under this labeling, then another labeling is randomly chosen. This process is repeated until, for every pair of distinct fragments, there exists at least one labeling in which the pair has distinct masses.

The randomized algorithm merely checks the necessary condition. Somewhat remarkably, in practice, this results in satisfying much stronger conditions (see below). One of our goals is to elucidate why this is so. We believe that such an algorithm can yield efficient labeling strategies. In Sec. 2.4

---

<sup>2</sup>Note that fragments whose primary sequences are permutations of one another cannot be distinguished by labeling+MS.

we report on an implementation using this idea. Using this algorithm, we discovered that UBL1 requires only one labeling. The minimum  $|L|$  for UBC9 is not known, but 5 labelings suffice to solve the feasibility problem for this protein. Hence  $|L| \leq 5$  suffices for UBC9, and  $|L| = 1$  suffices for UBL1.

### 2.3.2 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 in Sec. 2.3.1 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  $\psi_i(k)$  denote the mass of fragment  $k$  in labeling scheme  $X_i$ . Suppose  $\psi_1(k) = \psi_1(g_1)$ ,  $\psi_1(k) \neq \psi_1(g_2)$ ,  $\psi_2(k) = \psi_2(g_2)$ , and  $\psi_2(k) \neq \psi_2(g_1)$ . 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.  $\square$

**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  $g \neq k$ ,  $f_{kg}(X_k) \neq 0$ .*

In practice, the sufficient condition in Claim 4 is much stronger than we need. We give some intuition as to why. First, 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*. Since roughly  $s^4 - s$  fragments will *not* occur in any experiment, we expect to find a great deal of negative evidence. In Sec. 3, we incorporate negative evidence into the data analysis phase.

Second, the necessary condition (Feasibility in Sec. 2.3.1) 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  $O(s^8)$  clauses of the necessary condition, that mass degeneracy is rare. The randomized experiment planning algorithm described above can be viewed as “satisficing 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  $\mathcal{F}^*$  that is selected “randomly” from a larger,  $O(s^4)$ -sized set  $\mathcal{F}$  of fragment hypotheses, given statistics on the mass degeneracy in  $\mathcal{F}$ . Below, and in Sec. 4, we quantitate these observations by modeling the statistical properties of mass degeneracy.

**Statistics of Mass Degeneracy.** A variety of statistical measures of mass degeneracy are possible. For example, let  $k \in \mathcal{F}$  be a fragment and  $\psi_i : \mathcal{F} \rightarrow \mathbb{N}$  be as above. We define  $c(k, i) = |\psi_i^{-1}(\psi_i(k))|$ , to be the number of fragment hypotheses potentially confusable with  $k$  in experiment  $X_i$ . Then given a set of fragment hypotheses  $\mathcal{F}$ , we say that a labeling  $L$  is  $(\beta, \eta)$ -good if the set  $\{k \in \mathcal{F} \mid \forall X_i \in L, c(k, i) > \beta\}$  has fewer than  $\eta$  elements. Hence, in a  $(\beta, \eta)$ -good labeling plan, the number of fragment hypotheses that are mass degenerate more than  $\beta$  is bounded above by  $\eta$ . The Sufficient Condition of Sec. 2.3.2 is equivalent to  $(1, 1)$ -goodness. When a labeling plan is  $(\beta, \eta)$ -good we can give the experimentalist a guaranteed upper bound on the amount of mass degeneracy she will encounter. This worst-case measure is still too strong in practice. In Sec. 4, we explore a weaker measure, in which probabilistic bounds on actual mass degeneracy (in  $\mathcal{F}^*$ ) are derived from an analysis

of hypothesis degeneracy (using the statistics of  $\mathcal{F}$ ). We also describe computational experiments that support the theory. Derivation of this data-driven necessary and sufficient condition for probabilistic mass degeneracy in Sec. 4 depends on the data analysis technique of *spectral differencing*, which we discuss in Sec. 3.

## 2.4 Experimental Results

The randomized algorithm (Sec. 2.3.1) quickly identified isotopically-labeled mass spectrometry experiments to disambiguate fragments in two example proteins. 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 for the proteins UBL1 [2] and UBC9 [1] are given in Table 1. Fragments of UBL1 can be disambiguated with one correctly-chosen isotopic labeling. Fragments of UBC9, however, require a set of labelings—the first labeling leaves 18 ambiguous pairs, of which only 10 are ambiguous with respect to the second labeling, and so forth. In Sec. 4 we calculate a probabilistic measure of how well these planned experiments are expected to eliminate mass degeneracy (see  $P(\text{interp})$  in Table 1).

For the UBL1-UBC9 complex, the program identifies 120 fragments for UBL1 and 276 fragments for UBC9, and thus 33516 fragments for the cross product. It then identifies 434241 mass-degenerate pairs in this set of fragments. This is far too many pairs for a small set of experiments to disambiguate, underscoring the importance of computational modeling and prediction of feasible fragments in the complex. A reasonable set of priors would restrict the number of functional hypotheses to a few hundred. Our experiments are evidence that SAR by MS can discriminate among hundreds of hypotheses, which should be sufficient for many complexes of interest.

## 3 Data Analysis

Optimal experiment planning (Sec. 2.2), attempts to carefully design the experiments so that the data analysis devolves to a table-lookup. The process is designed to minimize ambiguity in fragment hypothesis interpretation. Without experiment planning to minimize mass degeneracy, the data analysis may yield ambiguous results (i.e., competing fragment and binding-mode hypotheses). Since optimal experiment planning appears difficult, in this section, we investigate an alternative approach, obtaining polynomial-time algorithms when some potential ambiguity can be tolerated.

A continuum of design tradeoffs is possible between planning and analysis. To explore this idea, we picked a point near the other end of the design spectrum, in which we assume that the experiment plan (labeling+cleavage) is given *a priori*, and the data analysis algorithm reports on the hypotheses than can be inferred from the collected spectra. The hypotheses will typically not be unique, since the experiment was not optimally planned. Our algorithm investigates the power of *spectral differencing*. We show that, for a fixed plan, spectral differencing analysis can be done efficiently (polynomial time).

### 3.1 Spectral Differencing

Trained spectroscopists interpret mass spectra using a technique called *spectral differencing*, in which two spectra from different labelings of a complex (but using the same cleavage agents) are compared. For example, a peak in an unlabeled (natural isotopic abundance) mass spectrum will shift to a higher mass in a selectively  $^{15}\text{N}$ -labeled spectrum (cf. Fig. 1). When peaks can be tracked across spectra, the corresponding mass shifts can be used to infer which fragment generated the peak. In this vein, we now consider the simpler problem of data interpretation given a labeling and cleavage plan (for example, the plan may have been selected in a randomized strategy, such as in Sec. 2.3.1).

Given a complex and a fixed cleavage agent, let  $S_i$  be a mass spectrum, represented as a set of masses (at observed peaks)  $\{s_1, s_2, \dots\}$ , under labeling scheme  $X_i$ .  $X_i = \{x_1, x_2, \dots\}$  is a vector of

labels as in Eq. (2). Let  $\phi_i(s)$  be the set of fragments which could have produced peak  $s$ :  $\phi_i(s) = \{k \in \mathcal{F} \mid s \approx \psi_i(k)\}$ , where  $\psi_i(k)$  is the mass of fragment  $k$  under  $X_i$ . Spectral differencing then identifies pairs of peaks in two different spectra  $S_1$  and  $S_2$  such that the same fragment could have caused both peaks. Let  $\mathcal{F}$  be the set of all possible fragments. 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  $\phi_1(s_1) \cap \phi_2(s_2)$ . Due to mass degeneracy,  $s_1$  in spectrum  $S_1$  could have multiple explaining fragments  $k \in \phi_1(s)$ . However, each such  $k$  must also have a peak  $s_2$  in spectrum  $S_2$  with  $k \in \phi_2(s_2)$  in order to be consistent with the spectral difference. This approach uses negative evidence to rapidly prune the number of fragments being considered.

We now develop a fast algorithm for spectral differencing. The *difference spectrum* of  $S_1$  and  $S_2$  is obtained from the Minkowski difference  $S_2 \ominus S_1 = \{s_2 - s_1 \mid s_2 \in S_2, s_1 \in S_1\}$  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 example above,  $X_1 = \mathbf{0}$  (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 maximum value  $t$  that depends on the primary sequence (for example, if all Arginine residues are labeled with  $^{15}\text{N}$ , 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  $\hat{n}_2 = 4$ ). There will be a lower bound  $l$  as 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

$$D(S_1, S_2) = \{(s_1, s_2) \in S_1 \times S_2 \mid s_2 - s_1 \in [l(s_1), t(s_1)]\}. \quad (4)$$

Now, suppose a peak  $s \in S_1$  is caused by a fragment  $f_k$ . Following Eq. (2), we get

$$s = N(f_k) \cdot (M + X_1). \quad (5)$$

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,  $s = N(f_2) \cdot (M + X_1)$ ,  $s = N(f_3) \cdot (M + X_1)$ , as well. Suppose  $(s, r) \in D(S_1, S_2)$ , i.e.,  $r$  is a candidate match for  $s$  across spectra. We say a fragment  $f_k$  *explains the mass shift*  $(s, r)$  when Eq. (5) holds and

$$r - s = N(f_k) \cdot (X_2 - X_1). \quad (6)$$

Let  $\mathcal{F}$  be the set of all possible fragments. We define the *set of interpretations of the mass shift*  $(s, r)$  as  $I(s, r) = \{f_k \in \mathcal{F} \mid \text{Eqs. (5) and (6) hold}\}$ . 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  $\mathcal{P}$  with  $n$  residues. Given a cleavage agent  $\gamma$ , we obtain a crosslinked and cleaved system  $\mathcal{P}(\gamma)$ , containing both 1- and 2-fragments. While the set of *possible* fragments that could make up  $\mathcal{P}(\gamma)$  is large ( $O(n^4)$ ), in any particular  $\mathcal{P}(\gamma)$  we will see only  $O(n)$  1-fragments (see Sec 2.2). *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  $\mathcal{P}(\gamma)$  to be  $O(n)$ .

For larger proteins, we find that 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  $\mathcal{P}(\gamma)$  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 toleranced difference spectrum can be computed in output-sensitive time  $O(c^2 \log n)$  where  $c$  is the size of  $\mathcal{P}(\gamma)$ , using  $O(n^4 \log n)$  preprocessing time.*

**Proof:** To compute  $I(D(S_1, S_2))$ , we store, for each fragment  $f$ , the interval  $[z(f) - \varepsilon, z(f) + \varepsilon]$  and the datum  $f$  in a binary range tree [9], where  $z(f) = N(f) \cdot (X_2 - X_1)$ . This preprocessing requires

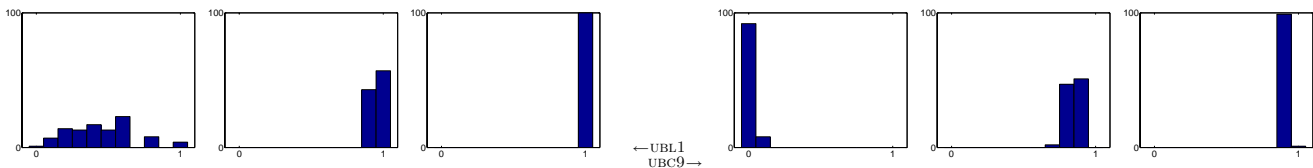


Figure 2: Interpretability of randomly planned sets of 1, 2, and 5 labelings (left to right), for UBL1 (graphs 1-3) and UBC9 (graphs 4-6). Each bar indicates how many sets, out of 100, have the given probability of interpretability.

time  $O(n^4 \log n)$ . Given a potential mass shift, we perform a lookup in the tree 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  $O(c^2 \log n)$ . For each non-empty lookup, we also check in  $O(1)$  time that Eq. (5) holds.  $\square$

**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 nontrivial, we envision it could be done in parallel with the wetlab molecular biology (selective labeling), which can take on the order of days. Wetlab expression and purification times will be similar in MS, NMR, Xray, or gel studies. However, taking the MALDI or ESI mass spectrum will be orders of magnitude faster than the the post-expression phases of NMR, Xray, or gel methods. After preprocessing, the  $O(c^2 \log n)$  computational lookup phase should be very fast, on a similar timescale to MS recording.

## 4 Probabilistic Mass Degeneracy

The data analysis techniques discussed in Sec. 3 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:

**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  $l_X$  such that  $f_{kl_X}(X) = 0$ .*

Negating the condition in Claim 7, we learn 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 satisfy our objective of eliminating incorrect hypotheses.

### 4.1 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  $p^* = E(|\mathcal{F}^*|)$  divided by the number of possible hypotheses  $p = |\mathcal{F}|$ . 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  $\varphi = p^*/p$ .

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  $C(f, i) =$

<sup>13</sup> C-labeled	<sup>15</sup> N-labeled	P(interp)
Unlabeled	Unlabeled	0.43
ARCEGILKSWV	NDQEHILSWV	1.0

<sup>13</sup> C-labeled	<sup>15</sup> N-labeled	$\chi$	P(interp)
unlabeled	unlabeled	27	0.021
NDQEHILKSTWV	RCQHKMSTWYV	18	0.88
QGISWV	ACQEGIKPY	10	0.99
ANDCEGHILS	RCQGILMFPSWY	3	0.9998
ARNQEHKMSV	ACQGLMWY	1	0.99999
DCQEILSW	ANEGLKMFTWY	0	0.9999997

Table 1: Isotopically-labeled experiment planning results from the randomized algorithm. (Left) Single experiment disambiguating fragment masses for UBL1. (Right) Sequence of experiments collectively disambiguating fragment masses for UBC9.  $\chi$  = number of remaining ambiguities. P(interp) is the probability that spectral differencing can eliminate all incorrect fragments (Eq. (7)).

$\psi_i^{-1}(\psi_i(f))$  denote the *conflict set* (mass-degenerate fragments) of fragment  $f$  in experiment  $i$ , and  $c(f, i) = |C(f, i)|$ . Then  $P(\text{appears}(f, i)) = 1 - \prod_{g \in C(f, i)} P(\text{incorrect}(g)) = 1 - (1 - \varphi)^{c(f, i)}$ .

We say a particular incorrect fragment hypothesis  $f$  is *eliminatable* unless for all experiments  $i \in L$ ,  $f$  appears in  $i$ . Hence,  $P(\text{eliminatable}(f, L)) = 1 - \prod_{i \in L} P(\text{appears}(f, i)) = 1 - \prod_{i \in L} (1 - (1 - \varphi)^{c(f, i)})$ .

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.

Since  $P(\text{interpretable}(L)) = \prod_{f \in \mathcal{F}} 1 - (P(\text{incorrect}(f)) \cdot (1 - P(\text{eliminatable}(f, L))))$ ,

$$P(\text{interpretable}(L)) = \prod_{f \in \mathcal{F}} 1 - \left( (1 - \varphi) \cdot \prod_{i \in L} (1 - (1 - \varphi)^{c(f, i)}) \right). \quad (7)$$

Eq. (7) 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.

## 4.2 Experimental Results

We have tested the interpretability metric for the proteins in Section 2.4. Table 1 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 interpretability probability equals 1.0.

We have also studied the ability of random labeling sets to satisfy the interpretability condition. Figure 2 shows histograms of the metric for sets of 1, 2, and 5 random labeling sets, 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. 2 shows empirical evidence that the Randomized Algorithm (Sec. 2.3.1) and the interpretability metric (Eq. (7)) are mutually beneficial, and may be combined in a package for experiment planning to probabilistically eliminate mass degeneracy.

## 5 Conclusions

MALDI and ESI MS are fast experimental techniques requiring subpicomolar sample sizes. They are therefore attractive for high-throughput functional genomics studies. However, while much faster, the information extracted is rather minimalist compared to NMR or Xray crystallography. Therefore, a large burden is placed on the algorithmic problems of experiment planning and data analysis

for SAR. In this paper, we explored the complexity of SAR by MS. 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. If we require only feasibility instead of optimality, we can develop 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 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. (7)) 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

- [1] UBC9 (or human UBC1), accession number p50550/q15698.
- [2] UBL1 (or human SM33), accession number p55856/q93068.
- [3] E. Amaldi and V. Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical Comput. Sci.*, 147:181–210, 1995.
- [4] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Comput. Sci.*, 209:237–260, 1998.
- [5] S. Arora, L Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *J. Comput. System Sci*, 54:317–331, 1997.
- [6] S. Arora, C. Lund, R. Matwani, M. Sudan, and M. Szegedy. Proof verification and intractability of approximation problems. In *Proc. IEEE FOCS*, pages 12–33, 1992.
- [7] S. Arora and S. Safra. Probabilistic checking of proofs: a new characterization of NP. In *Proc. IEEE FOCS*, pages 2–13, 1992.
- [8] G. Ausiello et al. *Complexity and Approximation: Combinatorial Optimization Problems and their Approximability Properties*. Springer-Verlag, 1999.
- [9] J. L. Bentley. Multidimensional divide and conquer. *Commun. ACM*, 23:214–229, 1980.
- [10] H. J. Bohm and G. Klebe. What can we learn from molecular recognition in protein-ligand complexes for the design of new drugs? *Angew. Chem. Int. Ed. Engl.*, 35:2588–2614, 1996.
- [11] Y. J. Cao et al. Photoaffinity labeling analysis of the interaction of pituitary adenylate-cyclase-activating polypeptide (PACAP) with the PACAP type I receptor. *Euro. J. Biochem.*, 224(2):400–406, 1997.

- [12] Xian Chen, S.V. Santhana Mariappan, John J. Kelley III, John H. Bushweller, E. Morton Bradbury, and Goutam Gupta. A PCR-based method for large scale synthesis of uniformly  $^{13}\text{C}/^{15}\text{N}$ -labeled DNA duplexes. *Federation of European Biochemical Societies (FEBS) Letters*, 436:372–376, 1999.
- [13] U. Feige, S. Goldwasser, L. Lovasz, S. Safra, and M. Szegedy. Approximating clique is almost NP-complete. *Proc. IEEE FOCS*, pages 2–12, 1992.
- [14] M. M. Halldorsson. Approximation via partitioning. *Technical Report IS-RR-95-0003F, School of Information Science, Japan Advanced Institute of Science and Technology, Hokuriku*, 1995.
- [15] J. A. Loo. Studying noncovalent protein complexes by electrospray ionization mass spectroscopy. *Mass Spectrometry Reviews*, 16:1–23, 1997.
- [16] Alan G. Marshall et al. Protein molecular mass to 1 da by  $^{13}\text{C}$ ,  $^{15}\text{N}$  double-depletion and FT-ICR mass spectrometry. *Journal of the American Chemical Society*, 119(2):443–434, 1997.
- [17] M. Scalf et al. Controlling the charge states of large ions. *Science*, 283:194–197, 1999.
- [18] T. Solouki et al. High-resolution multistage MS, MS2, and MS3 matrix-assisted laser desorption/ionization FT-ICR mass spectra of peptides from a single laser shot. *Analytical Chemistry*, 68(21):3718–3725, 1996.
- [19] M. Sridharan, R. Lilien, and B. R. Donald. Computational binding prediction studies for a library of ligands to inhibit Core Binding Factor- $\beta$  (CBF- $\beta$ ) binding to CBF- $\alpha$ . *In preparation*, 1999.

**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

## A Lower Bounds (Proof of Lemma 1)

This proof has been relegated to the appendix for reasons of space.

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 is 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 “ $4^{20}$ ” combinatorial argument in Sec. 2.2 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 some restrictions on the possible input parameters that don’t arise in other types of problems. For example, if one is trying 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  $x$  might be some physical parameter, such as mass, and restricting it to be say, 0 or 1, leaves you with a set of problems that are not physically realizable or interesting. Thus the challenge, roughly, is to show that 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 [3, 4]:

**Problem name:** BIN FLS  $\neq$

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

Problem definition: Does there exist  $x_j \in \{0, 1\}$ ,  $j = 1, \dots, m$  such that

$$\sum_{j=1}^m a_{ij}x_j \neq b_i, i = 1, \dots, n \quad (8)$$

See [5, 13, 7, 6, 8, 14] 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,  $i = 1, \dots, n$ , there is an equivalent instance with  $n + m$  variables and  $2n$  inequalities, in which  $n$  of the right hand sides are  $r_i$ ,  $i = 1, \dots, n$ , and  $n$  are 0.*

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

$$\sum_{j=1}^m a_{ij}x_j + (r_i - b_i)y_i \neq r_i, i = 1, \dots, n$$
$$y_i \neq 0, i = 1, \dots, n$$

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

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  $^{13}\text{C}$  selective labeling is permitted. We call this problem  $^{13}\text{C}$ -OMSEP-SAT:

**Problem name:**  $^{13}\text{C}$ -OMSEP-SAT

Input:  $m$  amino acids  $z_1, \dots, z_m$ , 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  $(h_{i1}, h_{i2}, \dots, h_{im})$  is the “difference vector”  $N_{kl}$  in Eq. (2) ( $h_{ij}$  the  $j^{\text{th}}$  element of the vector  $N_{kl}$ , corresponding to the difference in the number of residues of amino acid type  $j$ ).

Problem definition: Each of the  $n$  constraints can be written as

$$\sum_{j=1}^m h_{ij}(c_j x_j + m_j) \neq 0 \quad (9)$$

where  $x_j \in \{0, 1\}$ . Can we simultaneously satisfy all the constraints?

**Claim 9**  $^{13}\text{C}$ -OMSEP-SAT is NP-hard.

**Proof.** The proof is by reduction from BIN FLS  $\neq$ . Assume WLOG that  $a_{ij} \in \mathbb{Z}$  ( $i = 1, \dots, n, j = 1, \dots, m$ ) (if not, multiply both sides of Eq. (8) 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

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

Given such an instance of BIN FLS  $\neq$ , we create an instance of  $^{13}\text{C}$ -OMSEP-SAT. Note that all  $m_j$  and  $c_j$   $j = 1, \dots, m$ , are chosen by nature. For each  $j = 1, \dots, m; i = 1, \dots, n$ , we set

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

Now let's look at our system of inequalities:

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

Making the substitutions from the mapping, we get

$$\sum_{j=1}^m a_{ij} \left( \prod_{k \neq j} c_k \right) (c_j x_j) + \sum_{j=1}^m a_{ij} \left( \prod_{k \neq j} c_k \right) m_j \neq 0 \quad i = 1, \dots, n.$$

or

$$\sum_{j=1}^m a_{ij} \left( \prod_{k \neq j} c_k \right) (c_j x_j) \neq - \sum_{j=1}^m a_{ij} \left( \prod_{k \neq j} c_k \right) m_j \quad i = 1, \dots, n.$$

But

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

so we can rewrite the inequalities as

$$\left(\prod_k c_k\right) \sum_{j=1}^m a_{ij} x_j \neq - \left(\prod_k c_k\right) \sum_{j=1}^m \frac{a_{ij} m_j}{c_j}$$

so this system is just the system (8) scaled by  $(\prod_k c_k)$ , and so is satisfiable if and only if (8) 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.  $\square$

**Problem name:**  $^{13}\text{C-OMSEP}$

Input: Identical to  $^{13}\text{C-OMSEP-SAT}$ . The constraints are again given in the form of Eq. (9).

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

**Lemma 1**  $^{13}\text{C-OMSEP}$  is NP-complete.

**Proof:** NP-hardness follows directly from Claim 9.  $^{13}\text{C-OMSEP}$  is in NP because it is an instance of the NP-problem MINIMUM UNSATISFYING LINEAR SUBSYSTEM (MULS) [5, 13, 7, 6, 8, 14, 3].  $\square$

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 thus 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 Sec. 1, we assume we are given *a priori* binding-mode and -region hypotheses, the situation is different. In this case we allow that we may not want to consider all pairs of fragments, due to other information. Therefore we can construct a protein corresponding to the set of constraints. To do so, for each constraint, we generate two fragments that will have the appropriate differences and will generate the appropriate constraint. We then assume, in experiment planning, that none of these fragments will interact with any other fragments except for designated pairs. Specifically, this requires adding a pair of new fragments (and cleavage sites), for each  $i$ . The fragments are given by the difference vector  $N_{kl}$  in Eq. (2), namely  $(h_{i1}, h_{i2}, \dots, h_{im})$ . Furthermore, this construction requires that our input set of *a priori* hypotheses to the experiment planner ignore all pairwise fragment-fragment constraints except for the correct ones in the reduction.

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.