# Physical Geometric Algorithms for Structural Molecular Biology[*]

Chris Bailey-Kellogg[†] John J. Kelley, III[†‡] Ryan Lilien[†] Bruce Randall Donald[†‡§¶]

## Abstract

A wealth of interesting computational problems arises in proposed methods for discovering new pharmaceuticals. This paper surveys our recent work in three key areas, using a Physical Geometric Algorithm (PGA) approach to data interpretation, experiment planning, and drug design:

(1) *Data-directed computational protocols for high-throughput protein structure determination.* A key component of structure determination through nuclear magnetic resonance (NMR) is that of assigning spectral peaks. We are developing a novel approach, called JIGSAW, to automated secondary structure determination and main-chain assignment. Jigsaw consists of two main components: graph-based secondary structure pattern identification in unassigned heteronuclear ($^{15}$N-labeled) NMR data, and assignment of spectral peaks by probabilistic alignment of identified secondary structure elements against the primary sequence.

(2) *Experiment planning and data interpretation algorithms for reducing mass degeneracy in mass spectrometry (MS).* MS offers many advantages for high-throughput assays (e.g. small sample size and large mass limits), but it faces the potential problem of mass degeneracy — indistinguishable masses for multiple biopolymer fragments (e.g. from a limited proteolytic digest). We are studying the use of selective isotopic labeling to substantially reduce potential mass degeneracy, especially in the context of structural determination of protein-protein and protein-DNA complexes.

(3) *Computer-aided drug design (CADD).* We are developing new CADD tools and applying them to the design of an inhibitor for the Core-Binding Factor-$\beta$ oncoprotein (CBF$\beta$-MYH11), a fusion protein involved in some forms of Acute Myelomonocytic Leukemia (AMML). Computational-structural studies of CBF help determine the molecular basis for its function and assist in the development of therapeutic strategies. A key issue in such studies is geometric modeling of protein flexibility; our approach attempts to account for flexibility by using an ensemble of structures representing low-energy conformations as determined by solution NMR.

Our long-range goal is the structural and functional understanding of biopolymer interactions in systems of significant biochemical as well as pharmacological interest. The research overviewed here represents a set of important steps towards that goal.

Invited paper at the Special Session on Computational Biology and Chemistry, International Conference on Robotics and Automation (2001), Seoul, Korea.

## 1 Introduction

The field of *Physical Geometric Algorithms (PGA)* studies computational processes that compute or reason about geometric or spatial relationships in the physical world, and their realization in application areas such as robotics and microelectromechanical systems. PGA research pursues the value proposition that, for such systems, predictions of behavior, arguments of correctness, and combinatorial precision devolve to a geometric analysis.

Some of the most challenging and influential opportunities for PGA arise in developing and applying information technology to understand the molecular machinery of the cell. Our recent work (and work by others) shows that many PGA techniques may be fruitfully applied to the challenges of computational molecular biology. PGA research may lead to high-throughput, automated systems that are useful in structural molecular biology.

Concomitantly, a wealth of interesting computational problems arises in proposed methods for discovering new pharmaceuticals. These problems include identifying the low-energy conformations of molecules, interpreting protein NMR (nuclear magnetic resonance) and X-ray data, inferring constraints on the shape of active drug molecules based on measurements of activity of related drug molecules, and docking candidate drug molecules to known protein targets. We survey our research on computer-aided drug design, new techniques for automated NMR data interpretation, and ex-

[*] This work is supported by the following grants to B.R.D.: National Science Foundation grants IIS-9906790, EIA-9901407, EIA-9802068, CDA-9726389, EIA-9818299, CISE/CDA-9805-548, IRI-9896020, and IRI-9530785, U.S. Department of Justice contract 2000-DT-CX-K001, and an equipment grant from Microsoft Research.

[†] Dartmouth Computer Science Department, Hanover, NH 03755, USA.

[‡] Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[§] Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.

[¶] Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: brd@cs.dartmouth.edu

periment planning and data interpretation for reducing mass degeneracy in mass spectrometry.

## 2 Computer-Aided Drug Design

PGA research in computational structural biology can assist in our long-range goal of understanding biopolymer interactions in systems of significant biochemical as well as pharmacological interest. One example is given by our work on Core-Binding Factor (CBF). CBF is a heterodimeric transcription factor involved in hematopoesis. Oncogenic translocations in CBF-$\alpha$ and -$\beta$ are implicated in Acute Myelomonocytic Leukemia (AMML). The oncogenic form of CBF-$\beta$, fusion protein CBF$\beta$-MYH11, oligomerizes with wild-type $\alpha$-subunits to sequester them outside the nucleus, vitiating transcription. The ultimate goal of this research is to design a small-molecule inhibitor to disrupt the complex formed by the wild-type $\alpha$ with the oncoprotein CBF$\beta$-MYH11. As a first step, we designed a ligand *in silico* to disrupt dimerization of the wild-type $\alpha$ and $\beta$ subunits. Our inhibitor could be useful in itself, in that one could potentially disrupt the healthy transcription factor with a small ligand, allowing new *in vivo* studies of AMML. Our inhibitor could also serve as a lead compound to inhibit the oncogenic form of CBF-$\beta$.

A key issue was geometric modeling of protein flexibility. In our "Computational Screening Studies for Core-Binding Factor-$\beta$: Use of Multiple Conformations to Model Receptor Flexibility," [19] we present an approach in computer-aided drug design that attempts to account for a target protein's flexibility. Computational techniques were employed in docking a database of 70,000 ligands to an ensemble of structures representing low-energy conformations of CBF-$\beta$ (as determined by solution NMR). Docking algorithms were used for each run and the top binding ligands were consolidated and screened in the wet-lab. Using our protocol, a small molecule inhibitor was designed to prevent dimerization of CBF. Our results — a ligand designed to disrupt the wild-type protein-protein interface — were validated in the wet-lab using electrophoretic mobility shift assays and SAR by NMR ($^{15}$N-HSQC chemical shift perturbation).

## 3 Algorithms for NMR Structural Biology

### 3.1 Introduction

High-throughput, data-directed computational protocols for Structural Genomics (or Proteomics) are required in order to evaluate the protein products of genes for structure and function at rates comparable to current gene-sequencing technology. We are pursuing a PGA approach known as the Jigsaw algorithm, a novel high-throughput, automated approach to protein structure characterization with nuclear magnetic resonance (NMR). For more details on this work, please see our recent papers in *The Journal of Computational Biology*, and *The International Conference on Computational Molecular Biology (RECOMB)* [3, 4, 16].

Jigsaw applies graph algorithms and probabilistic reasoning techniques, enforcing first-principles consistency rules in order to overcome a 5-10% signal-to-noise ratio. Jigsaw utilizes only four NMR experiments, none of which requires $^{13}$C-labeled protein, thus dramatically reducing both the amount and expense of wet lab molecular biology and the total spectrometer time. Results for three test proteins demonstrate that Jigsaw correctly identifies 79-100% of $\alpha$-helical and up to 65% of $\beta$-sheet NOE connectivities, and correctly aligns up to 90% of secondary structure elements. Jigsaw is very fast, running in minutes on a Pentium-class Linux workstation. This approach yields quick and reasonably accurate (as opposed to the traditional slow and extremely accurate) structure calculations. It could be useful for quick structural assays to speed data to the biologist early in an investigation, and could in principle be applied in an automation-like fashion to a large fraction of the proteome.

### 3.2 Algorithmic Approach

Jigsaw consists of two main components: (1) graph-based secondary structure pattern identification in unassigned heteronuclear ($^{15}$N-labeled) NMR data, and (2) assignment of spectral peaks by probabilistic alignment of identified secondary structure elements against the primary sequence. Deferring assignment eliminates the bottleneck faced by traditional approaches, which begin by correlating peaks among dozens of experiments.

The first key idea of Jigsaw (see Figure 1) is that regular protein secondary structure yields stereotypical through-space atom interactions, which are visible in a NOESY spectrum through the Nuclear Overhauser Effect (NOE). We can find such patterns in a spectrum *even if the positions in the primary sequence (assignments) are unknown.*

Jigsaw encodes NOESY data in a graph with nodes representing *unassigned* putative residues and edges representing possible interactions observed in the NOESY spectrum. This graph is very noisy (only about 10% signal) since many residues have approximately the same chemical shift for an interacting proton. However, buried within this graph is a set of edges that look like the canonical $\alpha$-helix and $\beta$-sheet interactions above.

Jigsaw relies on the fact that the noise edges are evenly distributed, and thus that *it is unlikely that large groups of incorrect edges will conspire to form alpha/beta patterns.* Jigsaw imposes a set of constraints derived from the patterns in order to focus a graph search,

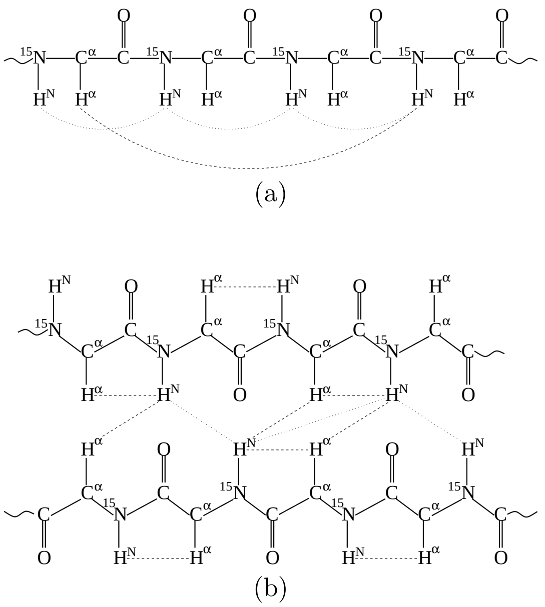


Figure 1: NOESY $H^{\alpha}$-$H^{N}$ (dashed) and $H^{N}$-$H^{N}$ (dotted) interactions in (a) $\alpha$-helices and (b) $\beta$-sheets.

working a "jigsaw puzzle" to find the correct secondary structure (see Figure 2). Of course, this jigsaw puzzle is somewhat different in that it has a very large number of extra pieces (and some missing ones, as well). However, we have shown that empirically the graph constraints serve to focus the search and avoid combinatorial explosion.

Jigsaw ranks secondary structures it discovers, based on criteria such as how well spectral peaks match, how many edges are missing, how many of the residues the graphs reach, and so forth. Figure 3 show $\beta$-sheets computed by Jigsaw for one example protein, CBF-$\beta$ (discussed above).

The second step in Jigsaw is to align the $\alpha$-helices and $\beta$-strands to substrings of the primary sequence. This relies on the use of a TOCSY spectrum, which has "fingerprints" of the protons on the side chains of the residues. These fingerprints in turn indicate probable amino acid types, which we locate in the primary sequence. This process is carried out probabilistically, assigning the probability that a strip is a certain AA type based on the results of a point-matching algorithm between observed and expected chemical shifts. Figure 4 shows both canonical proton shifts for the different amino acid types, culled from the BioMagResBank (BMRB), and observed proton shifts for some residues of Human Glutaredoxin. While most residues don't look that similar to the expected fingerprints, enough of them look enough like the expected fingerprints that a long helix or strand can be correctly aligned.

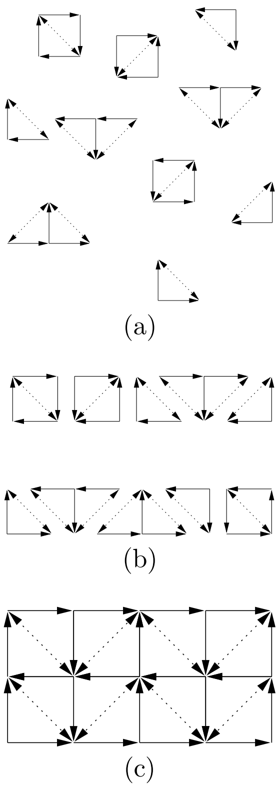


Figure 2: Jigsaw algorithm overview: (a) identify graph fragments, (b) merge them sequentially, and (c) collect them into complete secondary structure graphs. Only correct fragments are shown here. Graphs from experimental data also generate a large number of incorrect fragments, but mutual inconsistencies prevent them from forming either long sequences or large secondary structure graphs.

Given individual AA-type probabilities, the alignment proceeds by computing the joint probability over a secondary structure string (helix or strand), starting at each location in the primary sequence. The best match is taken as the proper alignment.

This alignment process has proved effective in practice. Table 1 provides example results for the helices and strands of CBF-$\beta$. The first set of results uses fingerprints collected from a set of experiments; the second uses fingerprints observed by a single experiment, the 80 ms $^{15}$N TOCSY. Results indicate both the rank of the correct alignment in the list of results, and its relative score — either the ratio of it to the second-best (if it's best) or the ratio of it to the best (if it's not). While the TOCSY alone yields good alignment results, the multi-exeriment results suggest that as TROSY-based pulse sequences improve, the results for a single experiment should get even better. In general, long sequences align better than short ones, although unusually noisy data can disrupt the alignment.

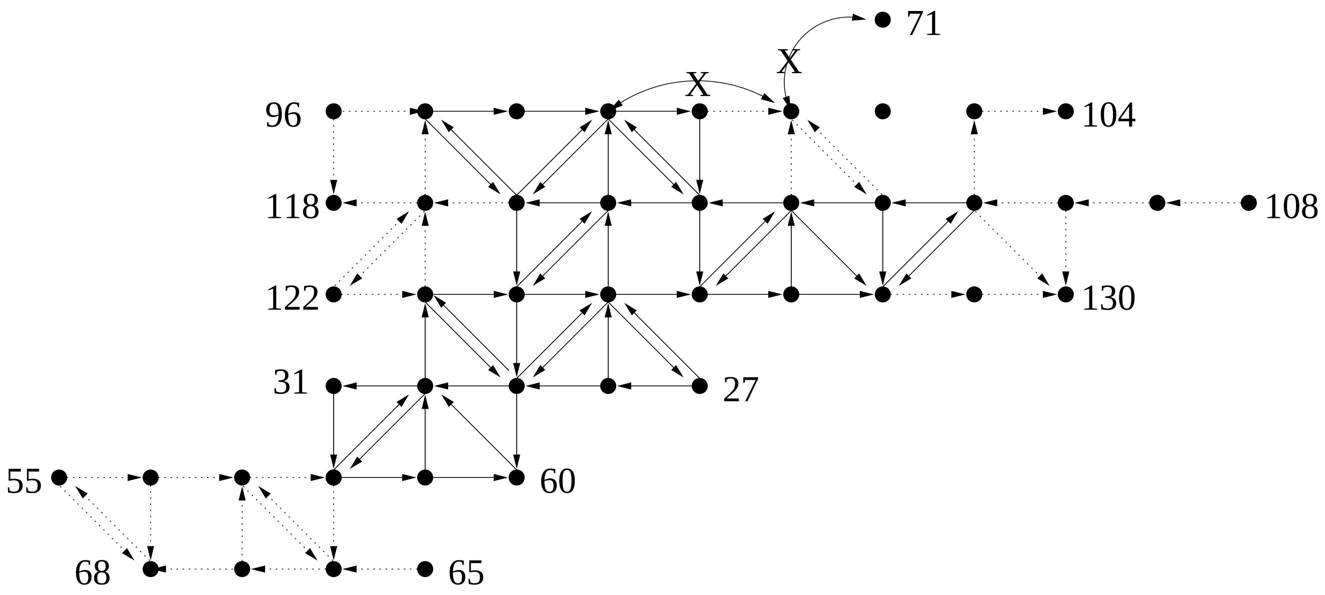


Figure 3: Some $\beta$-sheets of CBF-$\beta$ computed by Jigsaw. Edges: solid=correct; dotted=false negative; X=false positive.

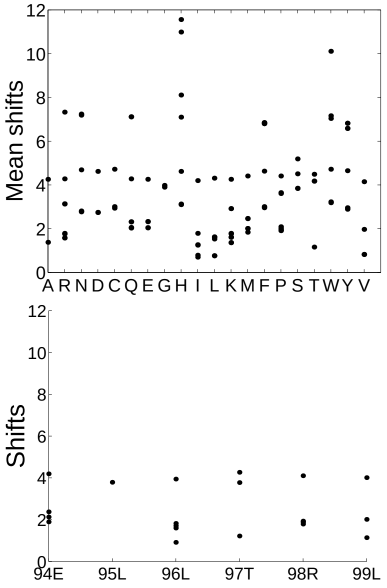


Figure 4: (top) BMRB $^1$H mean chemical shifts over different amino acid types. These shifts define "fingerprints" for the different amino acid types. (bottom) Observed fingerprints for some residues of Human Glutaredoxin. Observed fingerprints don't exactly match expectations (e.g. 95 and 96 are both L), but yield enough information that joint probability across an entire $\alpha$-helix or $\beta$-strand identifies the proper alignment in the primary sequence.

| Sequence | Multi-experiment | | $^{15}$N TOCSY | |
|---|---|---|---|---|
| | Rank | $\rho$ | Rank | $\rho$ |
| $\alpha_1$:10–16 | 1 | $9 \cdot 10^4$ | 1 | $3 \cdot 10^2$ |
| $\alpha_2$:18–23 | 1 | $2 \cdot 10^4$ | 17 | $4 \cdot 10^{-6}$ |
| $\alpha_3$:34–36 | 1 | $4 \cdot 10^1$ | 3 | $7 \cdot 10^{-2}$ |
| $\alpha_4$:43–52 | 1 | $1 \cdot 10^{13}$ | 1 | $2 \cdot 10^4$ |
| $\alpha_5$:131–140 | 1 | $7 \cdot 10^{14}$ | 1 | $1 \cdot 10^{19}$ |
| $\beta_{1,1}$:27–31 | 1 | $4 \cdot 10^3$ | 5 | $3 \cdot 10^{-2}$ |
| $\beta_{1,2}$:55–60 | 1 | $2 \cdot 10^6$ | 1 | $2 \cdot 10^4$ |
| $\beta_{1,3}$:65–68 | 1 | $2 \cdot 10^1$ | 1 | $1 \cdot 10^3$ |
| $\beta_{2,1}$:96–104 | 1 | $2 \cdot 10^1$ | 1 | $7 \cdot 10^2$ |
| $\beta_{2,2}$:108–117 | 1 | $4 \cdot 10^{10}$ | 11 | $3 \cdot 10^{-5}$ |
| $\beta_{2,3}$:122–130 | 1 | $3 \cdot 10^4$ | 5 | $1 \cdot 10^{-1}$ |

Table 1: Fingerprint-based alignment results for $\alpha$-helices and $\beta$-strands of CBF-$\beta$, with fingerprints obtained from a set of experiments or a single 80 ms $^{15}$N TOCSY. $\rho$ indicates the relative score of the alignment — relative to either the best alignment, if the correct one is not best, or else to the second-best alignment.

### 3.3 Summary

Jigsaw offers a novel approach to the automated assignment of NMR data and the determination of protein secondary structure. Since Jigsaw uses only four spectra and $^{15}$N-labeled protein, it is applicable in a much higher throughput fashion than traditional techniques, and could be useful for applications such as quick structural assays and Structure-Activity Relation (SAR) by NMR. It demonstrates the large amount of information available in a few key spectra. Finally, Jigsaw formalizes NMR spectral interpretation in terms of graph algorithms and probabilistic reasoning techniques, laying the groundwork for theoretical analysis of spectral information. Jigsaw has been successfully applied to NMR data sets from (1) the glutaredoxin family of proteins, which play an important role in maintenance of the redox state of the cell as well as in DNA biosynthesis and (2) Core-Binding Factor (described above). These are just first steps in developing new PGA for NMR structural biology. Future work will extend the Jigsaw formalism to apply to larger proteins, develop faster and more accurate algorithms, and collaborate with NMR structural biologists to develop a useful suite of high-throughput tools. PGA will be important for spectral intrerpretation, conformational search, pattern recognition, kinematics, dynamics, and modelling. Computational approaches adapted from robotics and machine vision can be useful in solving key problems in NMR structural biology. New algorithms are required that can quickly extract significantly more structural information from sparse experimental data. For example, in [16], a novel approach to multidimensional NMR analysis is proposed in which the data are interpreted in the time-frequency domain, as opposed to the traditional frequency domain. Time-frequency analysis exposes behavior orthogonal to the magnetic coherence transfer pathways, thus affording new avenues of NMR discovery. In particular, we demonstrate the heretofore unknown presence of through-space inter-atomic distance information within $^{15}$N-edited heteronuclear single-quantum coherence ($^{15}$N-HSQC) data. A biophysical model explains these results, and is supported by further experiments on simulated spectra.

## 4 Algorithms for Structure from Mass Spectrometry

### 4.1 Introduction

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). In structural mass spec, mass peaks must be uniquely assignable in order to distinguish hypotheses. We are studying the PGA tasks of planning and interpreting MS experiments that use selective isotopic labeling, thereby substantially reducing potential mass degeneracy. 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 (see Figure 5). For more details on this work, please see our recent papers in *The Journal of Computational Biology* and *The International Conference on Intelligent Systems for Molecular Biology (ISMB)* [1, 2].

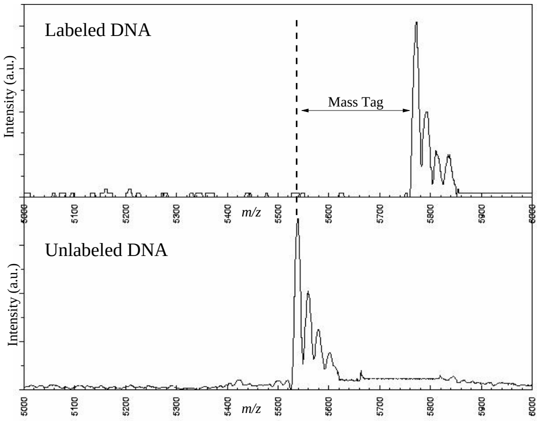


Figure 5: MALDI-TOF mass spectra of an 18 bp DNA oligonucleotide d(GACATTTGCGGTTAGGTC): (top) $^{13}$C-,$^{15}$N-labeled 18-mer; (bottom) $^{12}$C,$^{14}$N-labeled 18-mer. The $m/z$ difference between the two peaks is called the mass tag.

We have developed algorithms to 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 (see Figure 6). 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 interface/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, *provided that peaks are uniquely assignable.*

We have explored the PGA problem of eliminating mass degeneracy in SAR by MS, developing a computational experiment planning framework that seeks to maximize the expected information content of an SAR by MS experiment, and an efficient data analysis algorithm that interprets the resulting data.

### 4.2 Algorithmic Approach

For ease of exposition, we address proteins and protein-protein complexes here. A protein or protein-protein complex is digested by a protease, yielding a set of $s$ possible *segments.* Any digestion site might be shielded,

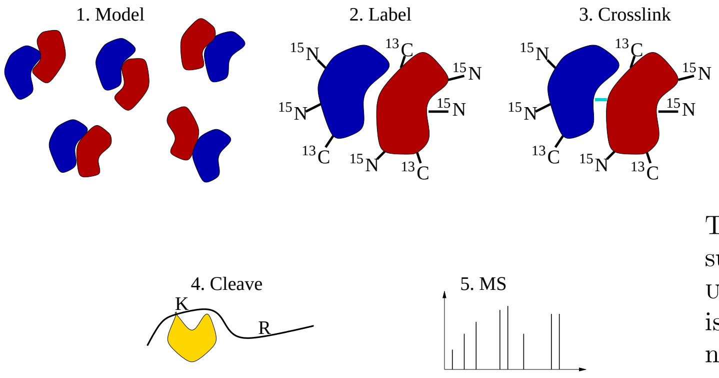


Figure 6: SAR by MS protocol overview. Mass peaks implicate residues in interaction site and nearby in space, assuming that mass peaks can be uniquely assigned. Isotopic labeling is employed to reduce mass degeneracy. Several different labelings may be used for 'multidimensional' MS (resulting in several different spectra, which must be correlated during data analysis).

| $^{13}$C-labeled | $^{15}$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 2: Isotopically-labeled experiment planning results from the randomized algorithm for the protein UBC9. $\chi$ = number of remaining ambiguities. P(interp) is the probability that spectral differencing can eliminate all incorrect fragments.

yielding a set of $O(s^2)$ possible *1-fragments* in the sequential union of segments. Finally, any pair of these might be cross-linked, yielding a set of $O(s^4)$ possible *2-fragments* in the cross product of 1-fragments.

Our goal is to ensure that no pair of fragments has the same mass. The entails enforcing a system of linear inequalities of the form $f_{kl}(X) \neq 0$, where $k$ and $l$ are fragments and $X$ is a labeling (a $\{0,1\}^2$ vector indicating whether or not each different amino acid type is $^{13}$C and/or $^{15}$N labeled). There are a quadratic number of such constraints, so a complex with $s$ cleavage sites per protein has $O(s^8)$ such constraints to satisfy. However, 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. In general, the set of possible binding modes can be constrained by a variety of techniques, for example by docking studies, chemical shift mapping for protein-protein complexes, and docking algorithms, together with homology searching, DNA footprinting, and mutational analysis. When available, this information restricts the set of *a priori* fragment interpretations.

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 $f_{kl}(X) = 0$. We have shown this problem to be NP-complete, even if restricted to $^{13}$C labeling.

Even if we could solve single-experiment planning, the resulting labeling might have too much mass degeneracy. Therefore, we pursue a different approach, allowing experiment plans to use several different labelings. A necessary condition is that every pair of fragments be distinguishable in some labeling (else we could never determine which of the pair is present). However, this isn't sufficient if there are multiple experiments, since a fragment $k$ could be mass degenerate with $g_1$ in experiment 1 and $g_2$ in experiment 2 (and thus distinguishable from $g_2$ in experiment 1 and from $g_1$ in experiment 2, satisfying the necessary condition), making it impossible to know whether $k$ actually exists. A sufficient condition is that for every fragment, there is at least one labling in which it is distinguishable from every other fragment.

The sufficient condition is too strong in practice, since there are more potential than observed fragments, and (as discusssed below), we can leverage negative evidence. We have implemented a randomized algorithm to plan a set of labelings that satisfies the necessary condition. Table 2 demonstrates its effectiveness: the number of degenerate pairs goes to 0 after a small number of experiments, and the probability of interpretation (discussed below) converges to 1.

Given a set of mass spectra, we can leverage negative evidence to eliminate fragments not supported by a peak in each spectrum. An efficient (polynomial-time) algorithm for testing the existence of fragments builds a range tree for the fragments, with keys representing intervals around the predicted masses. This preprocessing step can be performed in parallel with the molecular biology. Then, given a set of spectra, simply look up each peak to find fragment explanations and intersect the sets.

A given experiment plan can be analyzed in a probabilistic framework that predicts how likely it is that the interpretation algorithm will be able to resolve all ambiguities. The key ideas are outlined below, with intuition in Figure 7.

1. Determine the *a priori* probability $\wp$ that a fragment hypothesis is incorrect (e.g. uniform $\wp = 1 - p^*/p$, or based on model). *Correctness* is a function of the biological ground truth.

2. A fragment $f$ *appears* in an experiment $i$ when something it is degenerate with (set $C(f,i)$ of size $c(f,i)$) is correct. *Appearance* is a function of our observations.

$$\begin{aligned} P(\text{appears}(f,i)) &= 1 - \prod_{g \in C(f,i)} P(\text{incorrect}(g)) \\ &= 1 - \wp^{c(f,i)} \end{aligned}$$

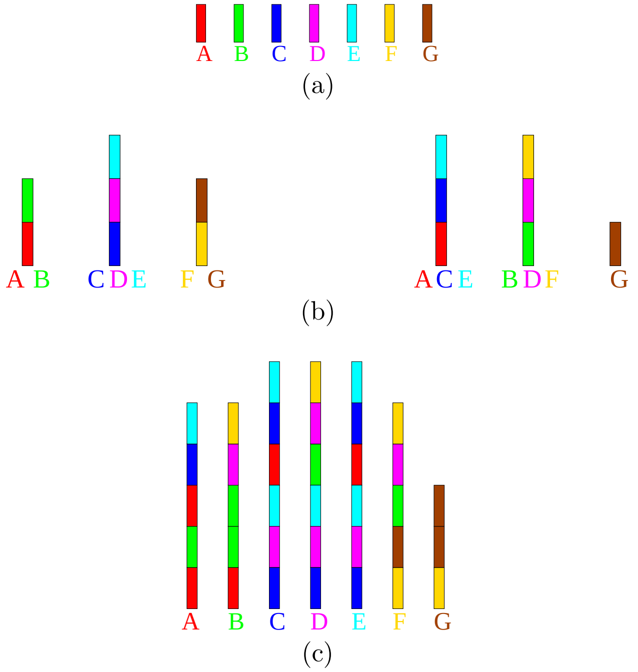
3. Spectral differencing can *eliminate* fragment $f$ unless it appears in all experiments.

$$\begin{aligned} P(\text{elim}(f,L)) &= 1 - \prod_{i \in L} P(\text{appears}(f,i)) \\ &= 1 - \prod_{i \in L} (1 - \wp^{c(f,i)}). \end{aligned}$$

4. An experiment plan $L$ is *interpretable* if all fragments are either correct or eliminatable.

$$\begin{aligned} &P(\text{interpretable}(L)) \\ &\quad = \prod_{f \in \mathcal{F}} (1 - P(\text{incorrect}(f)) \cdot (1 - P(\text{elim}(f,L)))) \\ &\quad = \prod_{f \in \mathcal{F}} (1 - \wp \cdot \prod_{i \in L} (1 - \wp^{c(f,i)})). \end{aligned}$$

P(interp) in Table 2 above shows that the probability of interpretability converges to 1 for an example protein. Fig 8 provides another example: how likely it is that randomly planned sets of labelings are correct. With 5 labelings, it is quite likely (practically guaranteed for UBL1) that the experiments will be interpretable.



(a)

(b)

(c)

Figure 7: *This figure should be viewed in color. Postscript is available at* `http://www.cs.dartmouth.edu/~cbk/papers/icra01.ps.gz`. Log-space intuition for probablistic framework. (a) *A priori* fragment probabilities. (b) Probability of appearance depends on equivalence classes of mass-degenerate fragments in each experiments. (c) Probability of elimination depends on appearance in all experiments.

## 4.3 Summary

We have tested our high-throughput techniques for obtaining structure from mass spec on the UBL1/UBC9 protein-protein complex. Yeast ubiquitin conjugating enzyme UBC9 has a functional human homolog UBEI2, which is critical for regulating the cell cycle. It complexes with UBL1 (a human ubiquitin-like protein) and associates with the RAD51/RAD52 proteins in their double-stranded DNA repair pathway. UBEI2/UBC9 is also involved in DNA recombination, and is essential for cell-cycle progression. These are just first steps in developing new PGA for structural MS. Future work will extend our algorithmic approach to SAR by MS, developing more efficient approximation algorithms, and generalizing our method to larger complexes by incorporating prior information into the probabilistic framework. In general, the set of possible binding modes can be constrained by a variety of techniques, for example by protein docking algorithms and NMR chemical shift mapping for protein-protein complexes, together with homology searching, DNA footprinting, and mutational analysis. When available, this information restricts the set of *a priori* fragment interpretations. In turn, this 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 will be made on the computer, and labeling+MS would be performed as a way of screening these hypotheses to test which are correct.

## References

[1] C. Bailey-Kellogg, J. J. Kelley III, C. Stein, and B. R. Donald. Reducing mass degeneracy in SAR by MS by stable isotopic labeling. In *The 8th Int'l Conf. on Intelligent Sys. for Mol. Bio. (ISMB-2000)*, pages 13–24, August 2000.

[2] C. Bailey-Kellogg, J. J. Kelley III, C. Stein, and B. R. Donald. Reducing mass degeneracy in SAR by MS by stable isotopic labeling. *J. Comp. Bio.*, 8(1):19–36, 2001. In press.

[3] C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. In *The 4th Int'l Conf. on Comp. Mol. Bio. (RECOMB-2000)*, pages 33–44, April 2000.

[4] C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY Jigsaw:

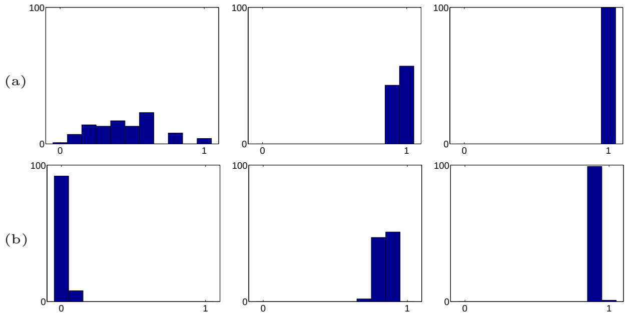


Figure 8: Interpretability of randomly planned sets of 1, 2, and 5 labelings (left to right), for (a) UBL1 and (b) UBC9. Each bar indicates how many sets, out of 100, have the given probability of interpretability.

Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comp. Bio.*, 7:537–558, 2000.

[5] 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.

[6] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, and N.J. Skelton. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press Inc., 1996.

[7] X. Chen, Z. Fei, L. M. Smith, E. M. Bradbury, and V. Majidi. Stable isotope assisted MALDI-TOF mass spectrometry allows accurate determination of nucleotide compositions of PCR products. *Anal. Chem.*, 71:3118–3125, 1999.

[8] X. Chen, S. V. Santhana Mariappan, J. J. Kelley III, J. H. Bushweller, E. M. Bradbury, and G. Gupta. A PCR-based method for large scale synthesis of uniformly $^{13}$C/$^{15}$N-labeled DNA duplexes. *Federation of European Biochemical Societies (FEBS) Letters*, 436:372–376, 1999.

[9] X. Chen, L. M. Smith, and E. M. Bradbury. Site-specific mass tagging with stable isotopes in proteins for accurate and efficient peptide identification. *Anal. Chem.*, 2000. In press.

[10] H. A. Gabb, R. M. Jackson, and M. J. Sternberg. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Bio.*, 272:106–120, 1997.

[11] P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Bio.*, 273(1):283–298, October 1997.

[12] B.J. Hare and G. Wagner. Application of automated NOE assignment to three-dimensional structure refinement of a 28 kD single-chain T cell receptor. *J. Biomol. NMR*, 15:103–113, 1999.

[13] X. Huang, N.A. Speck, and J.H. Bushweller. Complete heteronuclear NMR resonance assignments and secondary structure of core binding factor $\beta$ (1-141). *J. Biomol. NMR*, 12:459–460, 1998.

[14] J. J. Kelley III. *Glutaredoxins and CBF: The backbone dynamics, resonance assignments, secondary structure, and isotopic labeling of DNA and proteins.* PhD thesis, Dartmouth College, 1999.

[15] J.J. Kelley III and J.H. Bushweller. $^{1}$H, $^{13}$C, and $^{15}$N NMR resonance assignments of vaccinia glutaredoxin-1 in the fully reduced form. *J. Biomol. NMR*, 12:353–355, 1998.

[16] C. Langmead and B. R. Donald. Extracting structural information using time-frequency analysis of protein NMR data. In *The 5th Int'l Conf. on Comp. Mol. Bio. (RECOMB-2001)*, April 2001. Accepted; in press.

[17] C. J. Langmead and B. R. Donald. Time-frequency analysis of protein NMR data. Poster, The 8th Int'l Conf. on Intelligent Sys. for Mol. Bio. (ISMB-2000), August 2000.

[18] R. H. Lathrop and T. F. Smith. Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Bio.*, 255:651–665, 1996.

[19] R. Lilien, M. Sridharan, X. Huang, J. H. Bushweller, and B. R. Donald. Computational screening studies for core binding factor beta: Use of multiple conformations to model receptor flexibility. Poster, The 8th Int'l Conf. on Intelligent Sys. for Mol. Bio. (ISMB-2000), August 2000.

[20] J. A. Loo. Studying noncovelent protein complexes by electrospray ionization mass spectroscopy. *Mass Spectrometry Reviews*, 16:1–23, 1997.

[21] A. G. Marshall et al. Protein molecular mass to 1 da by $^{13}$C, $^{15}$N double-depletion and FT-ICR mass spectrometry. *J. American Chem. Soc.*, 119(2):443–434, 1997.

[22] C. Mumenthaler and W. Braun. Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Bio.*, 254:465–480, 1995.

[23] R. Norel, D. Petrey, H. J. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. *Proteins: Structure, Function, and Genetics*, 36:307–317, 1999.

[24] A. Scaloni, N. Miraglia, S. Orrù, P. Amodeo, A. Motta, G. Maroni, and P. Pucci. Topology of the calmodulin-melittin complex. *J. Mol. Bio.*, 277:945–958, 1998.

[25] S.B. Shuker, P.J. Hajduk, R.P. Meadows, and S.W. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274:1531–1534, 1996.

[26] 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. *Anal. Chem.*, 68(21):3718–3725, 1996.

[27] C. Sun, A. Holmgren, and J. Bushweller. Complete $^{1}$H, $^{13}$C, and $^{15}$N NMR resonance assignments and secondary structure of human glutaredoxin in the fully reduced form. *Protein Sci.*, 6:383–390, 1997.

[28] H. Takahashi, T. Nakanishi, K. Kami, Y. Arata, and I. Shimada. A novel NMR method for determining the interfaces of large protein-protein complexes. *Nature Structural Biology*, 7:220–223, 2000.

[29] R. A. Venters, W. J. Metzler, L. D. Spicer, L. Mueller, and B. T. Farmer. Use of $H^1_N$-$H^1_N$ NOEs to determine protein global folds in predeuterated proteins. *J. American Chemical Society*, 117(37):9592–9593, 1995.

[30] K. Wüthrich. *NMR of Proteins and Nucleic Acids.* John Wiley & Sons, 1986.

[31] M. M. Young, N. Tang, J. C. Hempel, C. M. Oshiro, E. W. Taylor, I. D. Kuntz, B. W. Gibson, and G. Dollinger. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *PNAS*, 97:5802–5806, 2000.

[32] F. Zappacosta, A. Pessi, E. Bianchi, S. Venturini, M. Sollazzo, A. Tramontano, G. Marino, and P. Pucci. Probing the tertiary structure of proteins by limited proteolysis and mass spectrometry: the case of minibody. *Protein Sci.*, 5:802–813, 1996.

[33] D.E. Zimmerman, C.A. Kulikowsi, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J. Mol. Bio.*, 269:592–610, 1997.