BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 21 2007, pages 2829–2835
doi:10.1093/bioinformatics/btm406

Structural bioinformatics

Minimizing the overlap problem in protein NMR: a computational framework for precision amino acid labeling

Michael J. Sweredoski[1,3], Kevin J. Donovan[2,3], Bao D. Nguyen[2], A.J. Shaka[2,3] and Pierre Baldi[1,3,*]

[1]Department of Computer Science, [2]Department of Chemistry and [3]Institute for Genomics and Bioinformatics, University of California, Irvine, USA

Received on April 21, 2007; revised on July 6, 2007; accepted on August 6, 2007

Advance Access publication September 25, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Recent advances in cell-free protein expression systems allow specific labeling of proteins with amino acids containing stable isotopes (^{15}N, ^{13}C and ^{2}H), an important feature for protein structure determination by nuclear magnetic resonance (NMR) spectroscopy. Given this labeling ability, we present a mathematical optimization framework for designing a set of protein isotopomers, or *labeling schedules*, to reduce the congestion in the NMR spectra. The labeling schedules, which are derived by the optimization of a cost function, are tailored to a specific protein and NMR experiment.

Results: For 2D ^{15}N-^{1}H HSQC experiments, we can produce an exact solution using a dynamic programming algorithm in under 2 h on a standard desktop machine. Applying the method to a standard benchmark protein, calmodulin, we are able to reduce the number of overlaps in the 500 MHz HSQC spectrum from 10 to 1 using four samples with a true cost function, and 10 to 4 if the cost function is derived from statistical estimates. On a set of 448 curated proteins from the BMRB database, we are able to reduce the relative percent congestion by 84.9% in their HSQC spectra using only four samples. Our method can be applied in a high-throughput manner on a proteomic scale using the server we developed. On a 100-node cluster, optimal schedules can be computed for every protein coded for in the human genome in less than a month.

Availability: A server for creating labeling schedules for ^{15}N-^{1}H HSQC experiments as well as results for each of the individual 448 proteins used in the test set is available at http://nmr.proteomics.ics.uci.edu.

Contact: pfbaldi@ics.uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Protein NMR spectra are often limited by what could be termed *spectral congestion*, the inability to discern individual resonance peaks with either certainty or clarity. For a 2D $[^{15}N,^{1}H]$ heteronuclear single-quantum correlation (HSQC) experiment (Bodenhausen and Ruben, 1980) a large, structured protein will display a spectrum rich with peaks. This crowding arises because there is one peak expected from each NH spin pair in the molecule, and there are many such pairs. Spectral congestion, exacerbated by line broadening from the slower tumbling rates of larger proteins, can produce a spectrum that is impossible to assign.

One solution is to use 3D and 4D spectra to improve the resolution, but these high-dimensional experiments typically incur a penalty in both instrument time and absolute sensitivity. Another avenue is to employ higher magnetic field strengths B_0, at 800 or 900 MHz ^{1}H resonance frequency. However, access to these state-of-the-art instruments is limited and their cost is a barrier to widespread adoption.

An obvious third way to tackle spectral congestion is simply to reduce the number of resonance peaks. There are at least two ways to achieve this. The first is to exploit some property of the spin systems, using a pulse sequence that selects certain 'spin topologies' (Levitt and Ernst, 1985) while rejecting most others. Broadly known as *editing* methods, it is difficult to design a set of edited spectra that (i) are each sufficiently simple; and (ii) preserve the entire information content when considered in aggregate.

A second way to reduce the number of peaks is to reduce the number of NMR-active nuclei. This method has been employed in selective ^{1}H-methyl group labeling (Rosen *et al.*, 1996) of otherwise perdeuterated proteins. Selective ^{1}H-methyl group labeling of the branched chain amino acids alanine, valine, leucine and isoleucine (γ2 only) is possible by overexpression in $^{2}H_2O$ using protonated pyruvate as the sole carbon source.

A more advanced, precise and controlled biochemical manipulation is now possible by dispensing with cell-based recombinant protein expression altogether, assembling just the relevant protein synthesis machinery itself, as a carefully formulated extract, and then conducting the protein synthesis in a test tube, supplying individual amino acids, the gene that codes for the protein of interest, and an energy source (Kudlicki *et al.*, 2007). These *in vitro* coupled transcription-translation protein expression systems have become the focus of increased research attention over the last 5 years, as it has become clear that the methodology has a number of telling advantages compared to expression in *Escherichia coli* or other living systems

*To whom correspondence should be addressed.

(Kigawa *et al*., 2004; Klammt *et al*., 2004; Koglin *et al*., 2006; Morita *et al*., 2004; Shi *et al*., 2004; Staunton *et al*., 2006).

Recently, a particular cell-free expression system has been shown to provide fast, efficient protein expression for use in NMR experiments (Keppetipola *et al*., 2006). The central feature is the ability to supply to the protein synthesis reaction mixture only certain designated subsets of labeled amino acids, for instance, containing ^{15}N, while all other amino acids are ^{14}N isotopomers, and so are not observed in a 2D or 3D NMR experiment. The absence of facile amino acid scrambling, which can defeat attempts to label only a subset of the amino acids in cell-based systems, is one important advantage of cell-free protein expression. Such lessened cross-talk is made possible by the absence of aminotransferase activity found in whole living organisms, allowing specific labeling by amino acid type (Kigawa *et al*., 1995, 1999), although not yet by position.

Much progress has already been achieved with the new power of cell-free protein expression, and spectra have been simplified by focusing on only certain peaks of interest. For example, selectively labeled samples have been used to assign some resonances in a congested spectral region of a protein containing a large number of α-helical regions (Trbovic *et al*., 2005).

Other efforts have used selectively labeled samples for the assignment of peaks from a single residue type (Kainosho and Tsuji, 1982; Yabuki *et al*., 1998).

Other efforts in selective labeling have aimed at complete backbone assignment by way of a number of selectively labeled samples, employing selective labeling schemes in conjunction with a specific assignment strategy, including a standard one used by an auto-assignment program (Zimmerman *et al*., 1997). Otting and coworkers devised a combinatorial labeling strategy that uses five separate samples: each residue is labeled in one, two, or three of the samples, and the residue types are assigned based on the presence and absence of peaks (Ozawa *et al*., 2006; Wu *et al*., 2006). Another novel labeling strategy used samples that are partially labeled, the different intensities between residue types then enabling peak assignment based on relative peak intensity (Parker *et al*., 2004). This approach can be used to identify up to 16 residue types from five different samples.

Here, we present an alternative and more comprehensive approach to selective labeling. We use the term *precision labeling* to denote any sample labeled by amino acid type with known amino acids that are nearly 100% enriched at specified positions, jointly or separately; there must be no amino acid scrambling and the sample must not be a mixture of isotopomers. The scheme is completely general, applying to carbon-13, nitrogen-15 and/or doubly labeled samples.

Our approach is unique because it is not wedded to a specific assignment strategy nor does it rely on peak intensity for assignment; rather it can be quite generally applied to try to produce optimum spectra for any *protein*. We will refer to a *labeling schedule* as the set of instructions that describe how many samples to prepare and how to isotopically label amino acids in each sample. Samples prepared according to the optimum labeling schedule should be maximally free and clear of peak overlap and therefore of the most utility for trouble-free assignment. In addition, constraints gleaned from the labeling schedule reduce the set of residues a peak could be mapped to during the assignment process. By creating *decongested* spectra, our approach overcomes one of the major hurdles in resolving the structure of larger molecules at lower magnetic field strengths.

2 METHODS

2.1 Formalization of the scheduling problem as an optimization problem

We will refer to the problem of finding an optimal labeling as Optimal Scheduling for Protein NMR Spectra (OSPNS). We use an integer programming formulation to describe the problem of OSPNS. We assume a fixed number of samples $k \in \mathbb{N}; k \geq 2$. Keep in mind that we will later iterate over various values of k, where $k \leq 20$ to find an appropriate balance between the number of overlaps and the number of samples produced. Additionally, we let $\mathcal{A}$ be the set of all the naturally occurring amino acids. The cost function $\mathcal{C}$

$$\mathcal{C} = \sum_{b\in\mathcal{A}} \sum_{c\in\mathcal{A}} \sum_{d\in\mathcal{A}} \sum_{e\in\mathcal{A}} \sum_{l=1}^{k} \mathcal{O}(b,c,d,e) \cdot C_{b,l} \cdot N_{c,l} \cdot C_{d,l} \cdot N_{e,l} \quad (1)$$

is optimized over the $40 \cdot k$ binary variables $N_{\mathrm{b},l}$ and $C_{b,l}$

$$N_{b,l} = \begin{cases} 1 & \text{if the nitrogens in amino acid } b \text{ are isotopically labeled} \\ & \text{in sample } l \text{ or all nitrogens are uniformly labeled} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$C_{b,l} = \begin{cases} 1 & \text{if the carbons in amino acid } b \text{ are isotopically labeled} \\ & \text{in sample } l \text{ or all carbons uniformly labeled} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The definition of the overlap function $\mathcal{O}$ varies with the NMR experiment so we leave the details for the following section. Finally, we must include a set of constraints that ensure all peaks are observable in at least one spectrum.

$$\sum_{l=1}^{k} N_{b,l} C_{c,l} \geq 1 \quad \forall b, c \in \mathcal{A} \quad (4)$$

It should be noted that there is some degeneracy in the labeling schedules. The sample numbers could be permuted without changing the value of the cost function $\mathcal{C}$.

While the recent advances in cell-free protein expression now allow chemists to select in principle how the amino acids are isotopically labeled, in practice the full palette of isotopic possibilities is not yet routinely available. In addition, by independently labeling the carbon and nitrogen, we are doubling the number of variables in the optimization equation, making the problem of finding a true optimal solution computationally challenging. We will define additional constraints to create two subproblems that work around these limitations. Additional constraints can be created to deal with any additional availability issues.

Subproblem A assumes that one does not need to label any carbons in the samples. We can therefore set $C_{b,l} = 1 \ \forall b \in \mathcal{A}, 1 \leq l \leq k$. We can then rewrite the constraint from Equation (4) as

$$\sum_{l=1}^{k} N_{b,l} \geq 1 \quad \forall b \in \mathcal{A}, 1 \leq l \leq k \quad (5)$$

and minimize the new cost function $\mathcal{C}$

$$\mathcal{C} = \sum_{b\in\mathcal{A}} \sum_{c\in\mathcal{A}} \sum_{l=1}^{k} O(b,c) \cdot N_{b,l} \cdot N_{c,l} \quad (6)$$

With the additional constraints in subproblem A, we create a quadratic programming problem, thus reducing the complexity from the general problem.

Subproblem B allows for the carbons in a particular amino acid class to be labeled, but *only* if the corresponding nitrogen is labeled as well. Subproblem B lets us perform multiple HNCA and HNCO experiments with doubly labeled amino acid samples that are currently available as

off-the-shelf items. The addition of the constraint in Equation (7) to the general problem defines subproblem B.

$$N_{b,l} \geq C_{b,l} \quad \forall b \in \mathcal{A}, 1 \leq l \leq k \tag{7}$$

2.2 Definition of the overlap function $\mathcal{O}$

Now that we have specified each problem, we return to the definition of $\mathcal{O}$, which is specific to each experiment. Our basic motivation is to calculate the number of possible overlapping peaks for any combination of amino acid labeling. To find the value of $\mathcal{O}$, we must first define the following set of variables.

$$\begin{aligned} &seq_i : \text{amino acid class of the } i\text{th residue, } seq_i \in \mathcal{A} \\ &E \in \{H, N, C^{\alpha}, C\} \\ &\delta E_i : \text{chemical shift of the } E \text{ nucleus of the} \\ &\quad i\text{th residue, } \delta E_i \in \mathbb{R} \\ &T_E : \text{overlap threshold between } E \text{ peaks, } T_E \in \mathbb{R} \\ &ov_E(i,j) = \begin{cases} 1 & \text{if } |\delta E_i - \delta E_j| \leq T_E \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

Note that T_E is dependent on several factors including the magnetic field strength and protein size. We empirically measured T_E by looking at various NMR spectra. Throughout this paper, N refers to the backbone nitrogens, H refers to the hydrogens bonded to the backbone nitrogens, and C (without subscripts) refers to the carbonyl carbons in the backbone.

For an HSQC experiment, we define $\mathcal{O}(b,c)$ as the number of times H'-N' cross-peaks from residues in amino acid class b overlap with H' -N' cross-peaks from residues in amino acid class c. Here, R_b is the set of indices of all the residues in amino acid class b.

$$\begin{aligned} R_b &= \{i : seq_i = b\} \\ \mathcal{O}(b,c) &= \begin{cases} \sum_{\substack{i \in R_b \\ j \in R_c}} ov_H(i,j) \cdot ov_N(i,j) & \text{if } b \neq c \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

In both definitions of $\mathcal{O}(b,c,d,e)$ for HNCO and HNCA experiments, we assume that $\mathcal{O}(b,c,d,e) = 0$ if $b = d$ and $c = e$. In addition, $R_{b,c}$ consists of the set of indices of residues in amino acid class c that are immediately preceded by a residue in amino acid class b. $S_{b,c}$ consists of the set of indices of residues in amino acid class c if $c = b$, otherwise $S_{b,c}$ is the empty set.

For an HNCO experiment, we define $\mathcal{O}(b,c,d,e)$ as follows.

$$\begin{aligned} R_{b,c} &= \{i : seq_{i-1} = b \text{ and } seq_i = c\} \\ \mathcal{O}(b,c,d,e) &= \sum_{\substack{i \in R_{b,c} \\ j \in R_{d,e}}} ov_H(i,j) \cdot ov_N(i,j) \cdot ov_C(i-1, j-1) \end{aligned} \tag{10}$$

For an HNCA experiment, we define $\mathcal{O}$ (b,c,d,e) as follows.

$$\begin{aligned} R_{b,c} &= \{i : seq_{i-1} = b \text{ and } seq_i = c\} \\ S_{b,c} &= \begin{cases} \{i : seq_i = b\} & \text{if } b = c \\ \emptyset & \text{otherwise} \end{cases} \\ \mathcal{O}(b,c,d,e) &= \sum_{\substack{i \in R_{b,c} \\ j \in R_{d,e}}} ov_H(i,j) \cdot ov_N(i,j) \cdot ov_{C^{\alpha}}(i-1, j-1) \\ &\quad + \sum_{\substack{i \in R_{b,c} \\ j \in S_{d,e}}} ov_H(i,j) \cdot ov_N(i,j) \cdot ov_{C^{\alpha}}(i-1, j) \\ &\quad + \sum_{\substack{i \in S_{b,c} \\ j \in R_{d,e}}} ov_H(i,j) \cdot ov_N(i,j) \cdot ov_{C^{\alpha}}(i, j-1) \\ &\quad + \sum_{\substack{i \in S_{b,c} \\ j \in S_{d,e}}} ov_H(i,j) \cdot ov_N(i,j) \cdot ov_{C^{\alpha}}(i,j) \end{aligned} \tag{11}$$

2.2.1 Example sequence Suppose we have the following amino acid sequence: GSTYHLDVVS. For an HSQC experiment, we must first calculate the various values of R_b (e.g. $R_G = \{1\}$, $R_V = \{8,9\}$ and $R_S = \{2,10\}$). The HSQC overlap function for some of various amino acid combinations are as follows.

$$\begin{aligned} \mathcal{O}(G,V) &= ov_H(1,8) \cdot ov_N(1,8) + ov_H(1,9) \cdot ov_N(1,9) \\ \mathcal{O}(S,S) &= 0 \end{aligned}$$

For an HNCO experiment, we first calculate the various values of $R_{b,c}$ (e.g. $R_{G,S} = \{2\}$, $R_{V,S} = \{10\}$ and $R_{T,T} = \emptyset$) and then calculate the overlap functions. The following examples illustrate some of the possible combinations.

$$\begin{aligned} \mathcal{O}(G,S,V,S) &= ov_H(2,10) \cdot ov_N(2,10) \cdot ov_C(1,9) \\ \mathcal{O}(T,T,V,S) &= 0 \end{aligned}$$

For an HNCA experiment, we first calculate the various values of $R_{b,c}$ and $S_{b,c}$ (e.g. $R_{G,S} = \{2\}$, $R_{V,S} = \{10\}$, $R_{D,V} = \{8\}$, $R_{V,V} = \{9\}$, $S_{D,V} = \emptyset$ and $S_{V,V} = \{9\}$) and then calculate the overlap functions. The following examples illustrate some of the possible combinations.

$$\begin{aligned} \mathcal{O}(G,S,V,S) &= ov_H(2,10) \cdot ov_N(2,10) \cdot ov_{C^{\alpha}}(1,9) \\ \mathcal{O}(D,V,V,V) &= ov_H(8,9) \cdot ov_N(8,9) \cdot ov_{C^{\alpha}}(7,8) \\ &\quad + ov_H(8,9) \cdot ov_N(8,9) \cdot ov_{C^{\alpha}}(7,9) \end{aligned}$$

Definitions of $\mathcal{O}$ (b,c,d,e) can be made for other NMR experiments following the same logic used for HSQC, HNCO and HNCA experiments. If HNCA experiments are used for constructing a backbone assignment, it would be more beneficial to concentrate our efforts on producing a set of decongested H-N spectra. To achieve this within our framework, one would simply drop the $ov_{C^{\alpha}}$ terms from Equation (11). Notice that in the Equations (9–11) the true number of overlaps are used, but an approximation can also be employed. In the following section, we develop an estimate for the overlap function.

Estimation of the overlap function $\mathcal{O}$

We must estimate the number of overlaps between each amino acid class when presented with a new protein for which no NMR data has previously been acquired. One simple approximation would be to assume that each residue has roughly the same probability of overlap with another residue, independent of the amino acid class. However, we know that certain factors such as amino acid chemical structure and secondary protein structure affect the location of chemical shifts in the spectrum. From the corresponding statistics listed at RefDB (Zhang *et al.*, 2003) database, we model the chemical shifts distributions with Gaussian distributions. The Gaussian models accurately describe the locations of the chemical shifts and are the standard distributions used in the RefDB and Biological Magnetic Resonance Data Bank (BMRB) (Seavey *et al.*, 1991) databases.

Given the distributions of the chemical shifts for the designated nuclei and the assumption that the distributions are independent, we can calculate the probability that the chemical shifts will overlap. For any backbone atom E, we can model the E-chemical shift of the ith residue by a Gaussian distribution $X_i \sim N(\mu_i, \sigma_i^2)$ where $\mu_i, \sigma_i^2 \in \mathbb{R}$ can be conditioned on features including amino acid class and secondary structure of residue i. The more information we have about the location of each residues chemical shifts, the greater the accuracy of the overlap function and therefore the fewer the number of overlaps we will observe in the spectra. In our tests, μ_i and σ_i are the mean and SD of the E-chemical shifts of all residues in the RefDB that are in the same amino acid class and secondary structure of residue i.

We can then model the difference between the distribution of E-chemical shifts of residues i and j as $X_{i-j} \sim N(\mu_j - \mu_i, \sigma_j^2 + \sigma_i^2)$. The probability of overlap is therefore stated as follows.

$$ov_E(i,j) = Pr(-T_E \leq X_{j-i} \leq T_E) \tag{12}$$

Optimization of the cost function $\mathcal{C}$

Optimization techniques such as simulated annealing (Kirkpatrick *et al.*, 1983) and genetic algorithms (Holland, 1962, 1975) can find good

solutions to the general problem, but there is no guarantee of optimality. The quality of the solutions for both methods depends on several parameters, including how long they are allowed to run. We attempted to optimize the cost function with off-the-shelf solutions, but we could not get satisfactory results. Finding solutions to subproblem A, which is quadratic in nature, took much longer than what would acceptable.

However, we are able to compute exact optimal solutions to subproblem A using a dynamic programming algorithm in under 2 h on a standard desktop machine. Since subproblem A involves the precision labeling of either the carbons or the nitrogens and the uniform labeling of the other, there are half as many variables over which to optimize. By taking advantage of this reduction in the size of the search space, we develop a dynamic programming algorithm that is optimal and whose run time is independent of the size of the protein. Note that even in the most general case, the run-time for evaluating the cost function is independent of the protein size and number of dimensions assuming the overlap function is precomputed. The details of the dynamic programming algorithm are given in the Appendix.

3 RESULTS

We first demonstrate our approach using the known HSQC spectrum of calmodulin (Qian *et al.*, 1998) to illustrate the expected gains in clarity that can be expected by taking a systematic approach to minimizing peak overlap. Calmodulin is a well-known benchmark protein in the field of protein NMR and it is large enough that the NMR spectrum on lower-field instruments would be challenging to analyze by standard methods. Using our dynamic programming algorithm for subproblem A, we are able to calculate optimal schedules for all possible number of samples in under 2 h.

We generate the labeling schedule using both the *true* and *estimated* overlap function for a HSQC experiment. The regenerated 2D spectra of the four precision-labeled samples generated with both the true and estimated overlap functions are presented in Figures 1 and 2. It should be noted that overlaps in the H-N spectrum are listed in the original assignment because the original experiment collected data in the N, H, C^{α}, C^{β} and C dimensions.

In the Supplementary Material online, we present a graph that compares the number of overlaps as a function of the number of samples for calmodulin. The Rayleigh criteria (cross-peaks overlap if chemical shifts in each dimension are within half a peak width), with nitrogen peak widths of 0.3 p.p.m., hydrogen peak widths of 0.04 p.p.m. and carbon alpha peak widths of 0.25p.p.m., is used to identify overlaps. With calmodulin, we noticed that the schedules calculated using the estimated number of peak overlaps does not match the performance of the optimal schedules calculated using the known true peak locations. This is due to the imprecision in the estimation of the number of overlaps. The degree of imprecision in our estimates is indicated by the error bars in the comparisons of the overlap functions in Figures 3–6.

To determine the number of samples to use, we must take into account both the cost of creating additional samples and the calculated percent congested (number of overlaps/max number of overlaps). In general, we noticed that 4–5 samples were typically enough for our purposes. Given explicit costs for creating samples and costs for peak overlaps in the spectra, the exact number of samples that would minimize the total cost could be derived easily.

We compare the resulting spectra from our schedules to the spectra resulting from the labeling schedule of Ozawa *et al.* (2006) and Wu *et al.* (2006). According to our overlap criteria at 500 MHz, we observe four overlaps in the four spectra simulated with our schedule using an estimated overlap function and six overlaps in the five spectra that would be produced with the schedule of Ozawa *et al.* (2006) and Wu *et al.* (2006).

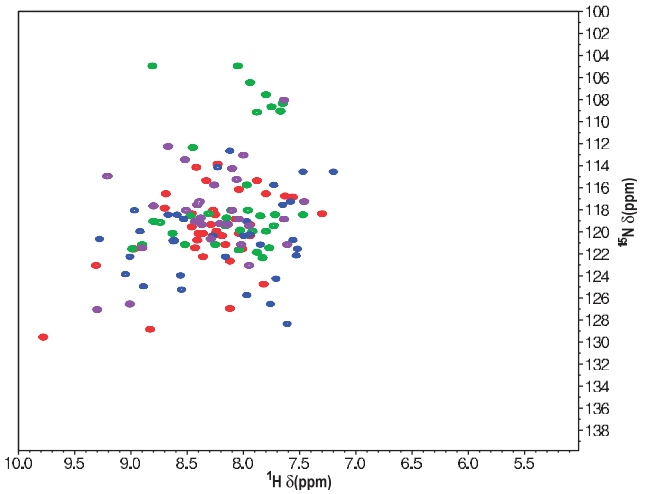

Fig. 1. ^{15}N-^{1}H HSQC spectra of calmodulin (Qian *et al.*, 1998) generated using our dynamic programming algorithm with the true cost function $\mathcal{C}$. The red sample includes A, R, N and D. The green sample includes C, Q, E, G and H. The blue sample includes I, L, K, F and S. The purple sample includes M, T, W, Y and V. The overlaps according to our criteria are between: E6 (8.99, 121.7) and D118 (8.97, 121.7), R30 (7.94, 120.5) and K77 (7.94, 120.4), L32 (8.62, 121.0) and L105 (8.62, 120.9), T34 (8.10, 118.2) and E87 (8.11, 118.2), S38 (8.14, 119.4) and M124 (8.14, 119.5), A46 (8.27, 120.5) and L48 (8.27, 120.5), A73 (8.52, 121.3) and E82 (8.52, 121.3), M76 (7.95, 119.6) and A127 (7.94, 119.5), H107 (8.16, 119.6) and M124 (8.14, 119.5), R126 (8.42, 119.2) and V142 (8.43, 119.2).

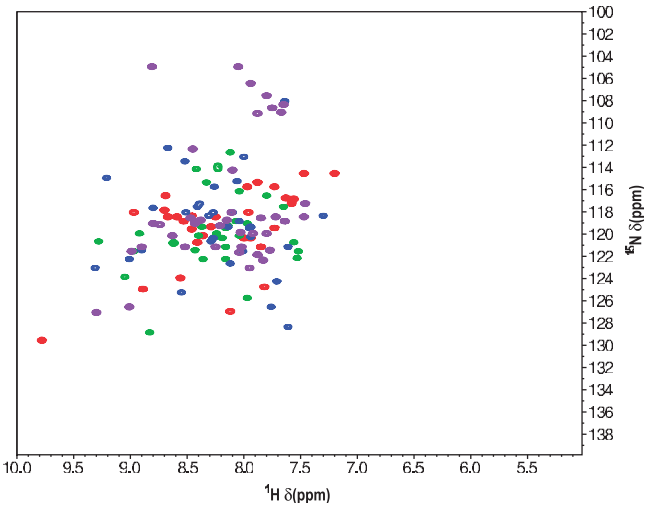

Fig. 2. ^{15}N-^{1}H HSQC spectra of calmodulin (Qian *et al.*, 1998) generated using our dynamic programming algorithm with an estimated cost function $\mathcal{C}$. The red sample includes R, N, Q, I and F. The green sample includes D, H, L, S and Y. The blue sample includes A, C, K, M and T. The purple sample includes E, G, W and V.

In all fairness, the goal of the schedule in Wu *et al*. is just to allow identification of the amino acids, and not necessarily to decongest the spectrum. Additionally, there are no performance criteria cited in Ozawa *et al*. (2006) or Wu *et al*. (2006) relevant to our problem. Note that the schedule in Wu *et al*. is developed for the average amino acid composition of proteins and not tailored to specific proteins. It should be noted that since our method is the first to directly tackle the problem of spectral congestion, there are no fair comparisons to other methods.

With only 10 overlaps in the HSQC spectrum of calmodulin, one could argue that a computer is not needed to calculate a schedule if we already know which peaks overlap. However, for cases where there are many overlaps, or we are using an estimated overlap function with real numbers, rather than integers, the problem would be nearly impossible solve by hand.

To further address these issues we have also computed schedules for a set of 448 proteins found in the BMRB (Seavey *et al*., 1991). The proteins in the test set have 50 or more residues, have N, H and C^{α} chemical shifts for 90% of its residues and have been rereferenced in the RefDB (Zhang *et al*., 2003). For each of these 448 proteins, we computed schedules using (a) the true overlap function; (b) an estimated function using secondary structure predicted by the SSpro software (Pollastri *et al*., 2002) in the SCRATCH server suite (Cheng *et al*., 2005) and (c) an overlap function that assumes the probability of overlap between any two residues is constant. Additionally, 10 random

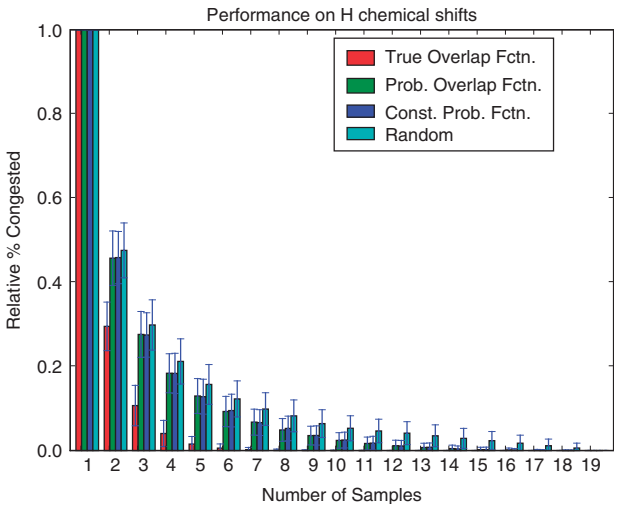

Fig. 3. Relative percent congestion in the H spectrum for the set of 448 proteins. Schedules are optimized with a true overlap function, an estimated overlap function, an equal number of residues per sample and an equal number of amino acids per sample.

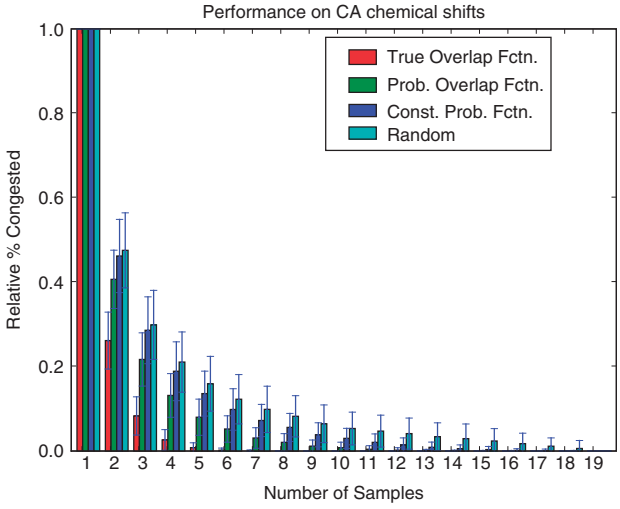

Fig. 5. Relative percent congestion in the C^{α} spectrum for the set of 448 proteins. Schedules are optimized with a true overlap function, an estimated overlap function, an equal number of residues per sample and an equal number of amino acids per sample.

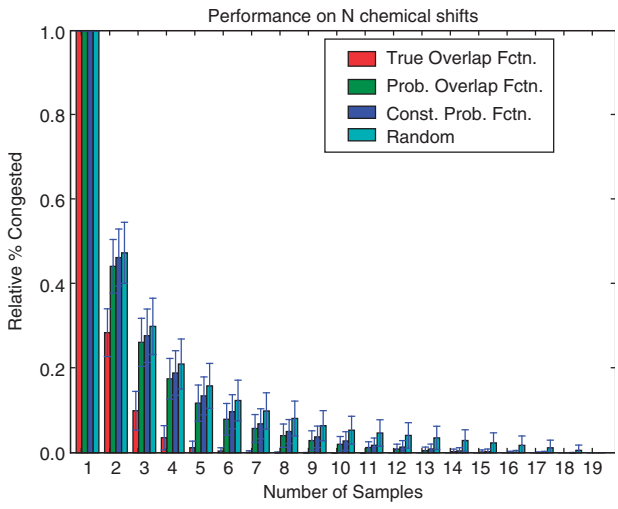

Fig. 4. Relative percent congestion in the N spectrum for the set of 448 proteins. Schedules are optimized with a true overlap function, an estimated overlap function, an equal number of residues per sample and an equal number of amino acids per sample.

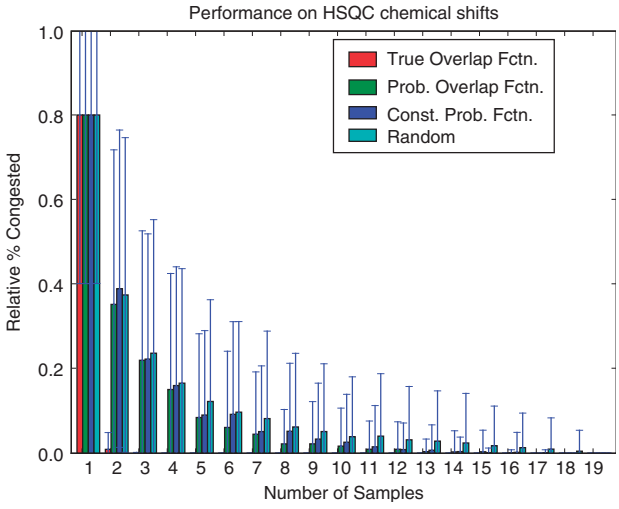

Fig. 6. Relative percent congestion in the HSQC spectra for the set of 448 proteins. Schedules are optimized with a true overlap function, an estimated overlap function, an equal number of residues per sample and an equal number of amino acids per sample.

schedules are generated for each protein with approximately the same number of amino acid classes labeled in each sample. Because there is not an exact mapping between the proteins in the BMRB and the PDB at the residue level, we could not prepare schedules using the true secondary structure for all the proteins in our set. On a subset where a reasonable mapping could be created, the results using schedules developed with true secondary structure knowledge were nearly identical to the schedules developed with the predicted secondary structure.

After computing the optimal schedules according to the cost functions we evaluate the schedule derived from true cost function using the relative percent congested [i.e. (number of overlaps − min number of overlaps)/(max number of overlaps − min number of overlaps)]. In Figures 3–5 we look at the performance of the various schedules to resolve overlaps in the H spectrum, the N spectrum and the C^{α} spectrum averaged over the set of 448 proteins. The error bars represent the SD in the performance.

In the H spectrum, we see that there is little gain using an estimated overlap over a schedule that assumes constant probability of overlap between residues. This is to be expected because the H chemical shifts are mostly independent of amino acid class and secondary structure. In the N spectrum and C^{α} spectrum, we see more benefit from using an estimated overlap function. We are able to achieve 17.5% relative congestion in the N spectrum and 13.2% relative congestion in the C^{α} spectrum using only four samples. This is a significant improvement over the performance of both the random schedules (20.9 and 21.0%, respectively) and constant probability of overlap schedules (18.9 and 18.9%, respectively).

In addition to looking at the spectra of individual backbone atoms, we also look at the performance of the schedules on resolving multidimensional spectra. However, issues arise when trying to measure the performance of the labeling schedules on high-dimensional spectra. Unless selective labeling is used, the chemical shifts in the BMRB that would compose a high-dimensional spectrum, such as a HNCA experiment, must already be resolved or they would not have been listed in the database. With this in mind, we must limit our evaluations to the HSQC spectra of the set of 448 proteins. The results from the HSQC spectra as well as the results from the C^{α} spectra provide proof that our method can work in higher dimensions.

For the HSQC spectra, we use the same evaluation methods as with the individual backbone atom spectrum and we observe roughly the same trends in Figure 6. One peculiarity of this figure is that with one sample, the proteins are ~80% relatively congested. This can be explained by the proteins in the test set that are already as decongested as they can be with respect to the HSQC spectrum. On the set of 448 proteins, we observe an average reduction of the relative percent congestion to 15.1% in four samples using our estimated overlap function. This is in comparison to the 56.1% relative congestion in the schedule developed by Wu *et al.* achieved in five samples.

While the results for randomly splitting the amino acids into two or three equal sized groups nearly matches that of our algorithm using the estimated overlap function, we see a more dramatic increase in performance when more than four samples are used.

Finally, our methods have been implemented in a web server located at http://nmr.proteomics.ics.uci.edu for user to produce schedules for minimally congested HSQC spectra. The user supplies the primary sequence and optionally the secondary structure. If the user does not supply the secondary structure, a secondary structure prediction made with SSpro is used. The optimized labeling schedules are then emailed back to the user.

4 CONCLUSION

In this article, we have provided a systematic way to design labeling schedules to decongest NMR spectra using precision-labeled samples. The method is a prime example of how to leverage the amino acid labeling that is now possible via cell-free protein expression.

By finding an optimal schedule using our computational framework, we have shown a dramatic increase in spectral resolution for not only the benchmark protein calmodulin (Qian *et al.*, 1998), but also on a curated set of 448 proteins listed in the BMRB. Using four samples on a 500 MHz HSQC spectra of calmodulin, the number of peak overlaps drops from 10 to 1 if the true cost function is utilized and from 10 to 4 if the cost function is estimated. In addition, our method is able to reduce the relative percent congestion by 84.9% in four samples using our estimated overlap function.

While this article has mainly focused on HSQC experiments, the approach is quite general and can be applied to any of the usual NMR experiments for backbone assignment. Future work will include adding the option to compute schedules for additional NMR experiments using our web server. Additionally, we will study how to incorporate the collected spectra into an integrated backbone assignment method.

The framework we have developed in this article provides a high-throughput method for assigning the backbone chemical shifts of proteins on a proteomic scale by allowing lower field, less expensive NMR spectrometers to run in parallel on larger structures. To help facilitate the high-throughput methods, we provide a web server that implements our method to produce decongested ^{15}N-^{1}H HSQC spectra. The simplified and clarified decongested spectra can be combined with the additional constraints gleaned from the labeling schedule in automated assignment programs thus streamlining the backbone assignment process. It is our hope and belief that our method for developing labeling schedules in conjunction with advances in cell-free protein expression can help usher in a new generation of high-throughput NMR spectroscopy studies.

ACKNOWLEDGEMENTS

Work supported by an NIH grant (GM-66763) and a UC Discovery Grant bio05-10533 to A.J.S., and a Laurel Wilkening Faculty Innovation award, a Microsoft Faculty Research Award, an NIH Biomedical Informatics Training grant (LM-07443-01) and an NSF MRI grant (EIA-0321390) to P.B.

Conflict of Interest: none declared.

REFERENCES

Bodenhausen,G. and Ruben,D.J. (1980) Natural abundance N-15 NMR by enhanced heteronuclear spectroscopy. *Chem. Phys. Lett.*, **69**, 185–189.

Cheng,J. *et al.* (2005) Scratch: a protein structure and structural feature prediction server. *Nuc. Acids Res.*, **33**, 72–76.

Holland,J. (1962) Outline for a logical theory of adaptive systems. *J. Assoc. Comput. Mach.*, **9**, 297–314.

Holland,J. (1975) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.
Kainosho,M. and Tsuji,T. (1982) Assignment of the three methionyl carbonyl carbon resonances in streptomyces subtilisin inhibitor by a carbon-13 and nitrogen- 15 double-labeling technique. A new strategy for structural studies of proteins in solution. *Biochemistry*, **21**, 6273–6279.
Keppetipola,S. *et al*. (2006) From gene to HSQC in under five hours: high-throughput NMR proteomics. *J. Am. Chem. Soc*., **128**, 4508–4509.
Kigawa,T. *et al*. (1995) Cell-free synthesis and amino acidselective stable isotope labeling of proteins for NMR analysis. *J. Biomol. NMR*, **6**, 129–134.
Kigawa,T. *et al*. (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett*., **442**, 15–19.
Kigawa,T. *et al*. (2004) Preparation of escherichia coli cell extract for highly prodcutive cell-free protein expression. *J. Struct. Funct. Genomics*, **5**, 63–68.
Kirkpatrick,S. *et al*. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.
Klammt,C. *et al*. (2004) High level cell-free expression and specific labeling of integral membrane proteins. *Eur. J. Biochem*., **271**, 568–580.
Koglin,A. *et al*. (2006) Combination of cell-free expression and NMR spectroscopy as a new approach for structural investigation of membrane proteins. *Magn. Reson. Chem*., **44**, S17–S23.
Kudlicki,W. *et al*. (2007) *Cell-Free Protein Expression*. Landes Bioscience, Austin, TX.
Levitt,M.H. and Ernst,RR (1985) Multiple-quantum excitation and spin topology filtration in high-resolution NMR. *J. Chem. Phys*., **83**, 3297–3310.
Morita,E. *et al*. (2004) A novel way of amino acid-specific assignment in ^{1}H-^{15}N HSQC spectra with a wheat germ cell-free protein synthesis system. *J. Biomol. NMR*, **30**, 37–45.
Ozawa,K. *et al*. (2006) 15N-labelled proteins by cellfree protein synthesis: strategies for high-throughput nmr studies of proteins and protein-ligand complexes. *FEBS J*., **273**, 4154–4159.
Parker,M. *et al*. (2004) A combinatorial selective labeling method for the assignment of backbone amide NMR resonances. *J. Am. Chem. Soc*., **126**, 5020–5021.
Pollastri,G. *et al*. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.
Qian,H. *et al*. (1998) Sequential assignment of ^{1}H, ^{15}N, ^{13}C resonances and secondary structure of human calmodulin-like protein determined by NMR spectroscopy. *Protein Sci*., **7**, 2421–2430.
Rosen,M.K. *et al*. (1996) Selective methyl group protonation of perdeuterated proteins. *J. Mol. Biol*., **263**, 627–636.
Seavey,B. *et al*. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*., **1**, 217–236.
Shi,J. *et al*. (2004) Protein signal assignments using specific labeling and cell-free synthesis. *J. Biomol. NMR*., **28**, 235–247.
Staunton,D. *et al*. (2006) Cellfree expression and selective isotope labelling in protein NMR. *Magn. Reson. Chem*., **44**, S2–S9.
Trbovic,N. *et al*. (2005) Strategy for the rapid backbone assignment of membrane proteins. *J. Am. Chem. Soc*., 13504–13505.
Wu,P. *et al*. (2006) Amino-acid type identification in 15N-HSQC spectra by combinatorial selective 15N-labeling. *J Biomol NMR*, **34**, 13–21.
Yabuki,T. *et al*. (1998) Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. *J. Biomol. NMR*, **11**, 295–306.
Zhang,H. *et al*. (2003) RefDB: A datbase of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, **25**, 173–195.
Zimmerman,D. *et al*. (1997) Automated analysis of proteinnmr assignments using methods from artificial intelegence. *J. Mol. Biol*., **269**, 592–610.

APPENDIX: EXACT SOLUTION USING DYNAMIC PROGRAMMING

The formulation of the dynamic programming algorithm requires that we augment the cost function $\mathcal{C}$ in Equation (6) with several additional variables. The new cost function $\mathcal{C}_{\mathrm{S,l,m}}$ operates on the subset of amino acids $\mathcal{S}\subseteq\mathcal{A}$ and range of samples specified by the integer variables l, m such that $1 \le l \le m \le k$. We can then define $\mathcal{C}_{\mathcal{S},l,m}$ as

$$\mathcal{C}_{\mathcal{S},l,m} = \sum_{b\in\mathcal{S}}\sum_{c\in\mathcal{S}}\sum_{i=l}^{m}\mathcal{O}(b,c)\cdot N_{b,i}\cdot N_{c,i} \tag{13}$$

Note that $\mathcal{C}_{\mathcal{A},1,k}$ is the same as the cost function $\mathcal{C}$ in Equation (6). In addition, we can add an offset variable $w \in \mathcal{N}$ to $\mathcal{C}_{\mathcal{S},l+w,m+w}$ without changing the cost so long as $l + w \ge 1$ and $m + w \le k$.

The following equality, which shows that the minimum of $\mathcal{C}_{\mathcal{S},l,m}$ is equal to the minimum of two subproblems, allows us to make use of a dynamic programming algorithm.

$$\min \mathcal{C}'_{\mathcal{S},l,m} = \min_{\mathcal{T}\subseteq\mathcal{S}} \mathcal{C}'_{\mathcal{T},l+w,\lfloor\frac{m+l}{2}\rfloor} + w + \mathcal{C}'_{\mathcal{S}/\mathcal{T},\lceil\frac{m+l}{2}\rceil+w,m+w} \tag{14}$$

Finding an optimal solution requires the computation of a cost matrix $P \in \mathbb{R}^{k\times 2^{|\mathcal{A}|}}$, where $\mathcal{A}_e$ is the eth subset of $\mathcal{A}$ and $P[i,e] = \min C'_{\mathcal{A}_e,1,i}$. The base row of our matrix, when $l = m = 1$, can be computed easily as

$$P[1,e] = \mathcal{C}_{\mathcal{A}_e,1,1} = \sum_{b\in\mathcal{A}_e}\sum_{c\in\mathcal{A}_e}\mathcal{O}(b,c)\ \forall\ 1 \le e \le 2^{|\mathcal{A}|} \tag{15}$$

Recursive computation of the second row of the matrix uses the costs from the first row.

$$P[2,e] = \min_{\mathcal{A}_f\subseteq\mathcal{A}_e} P[1,f] + P[1,g] \text{ where } \mathcal{A}_g = \mathcal{A}_e/\mathcal{A}_f \tag{16}$$

We can then recursively compute successive rows of the matrix using the costs from the prior rows.

$$P[i,e] = \min_{\mathcal{A}_f\subseteq\mathcal{A}_e} P[\lfloor\frac{i}{2}\rfloor,f] + P[\lceil\frac{i}{2}\rceil,g] \text{ where } \mathcal{A}_g = \mathcal{A}_e/\mathcal{A}_f \tag{17}$$

The sets $\mathcal{A}_f$ and $\mathcal{A}_g$ should be saved for each e and i for reconstruction of the optimal schedule using backtracking.

Once we have computed the entire matrix P, we must reconstruct the labeling schedule that gave us the optimal cost computed in $P[k,2^{|\mathcal{A}|}]$. We start with the assumption that each amino acid is only labeled once (i.e., $\sum_{i=1}^{k} N_{b,i} = 1\ \forall b \in A$). To construct our optimal schedule we will set $N_{b,i} = 1$ for some $1 \le i \le \lfloor\frac{k}{2}\rfloor$ and all $b \in \mathcal{A}_f$ and set $N_{c,j} = 1$ for some $\lceil\frac{k}{2}\rceil \le j < k$ and all $c \in \mathcal{A}_g$ where $\mathcal{A}_f$ and $\mathcal{A}_g$ were the sets that gave us the optimal cost for $P[k,2^{|\mathcal{A}|}]$. We then proceed to split the amino acids in $\mathcal{A}_f$ into the two sets that gave us the optimal cost for $P[\lfloor\frac{k}{2}\rfloor,f]$, knowing that we will label one set of the amino acids in the first half of the samples allotted and the other set in the second half of the samples allotted.

We continue our construction of the optimal schedule by recursively partitioning the amino acids and samples until we reach the base case where there is only one sample allotted for labeling a set of amino acids. At this point, we have each amino acid labeled in one sample and we have constructed a labeling schedule with an optimal cost.

The run time of this algorithm is independent of the protein length and is only dependent on the number of different amino acids, which is fixed at 20 if we only include the naturally occurring amino acids. The run time is therefore constant assuming the overlap function is precomputed.