

A Hierarchical Grow-and-Match Algorithm for Backbone Resonance Assignments Given 3D Structure

Fei Xiong and Chris Bailey-Kellogg
Department of Computer Science
Dartmouth College
6211 Sudikoff Laboratory
Hanover, NH 03755, USA
Email: cbk@cs.dartmouth.edu

Abstract—This paper develops an algorithm for NMR backbone resonance assignment given a 3D structure and a set of relatively sparse ^{15}N -edited NMR data, with the through-space ^{15}N -edited NOESY as the primary source of information. Our approach supports high-throughput solution studies of dynamics and interactions (e.g., ligand binding), when the structure has previously been determined by crystallography or modeled computationally. We employ a graph matching approach, identifying correspondence between a given contact graph and a corrupted version representing the NMR data. Our hierarchical grow-and-match algorithm decomposes the contact graph into sequential fragments with relatively dense interactions, and then combines possible assignments for the fragments, searching over the combinations with effective but conservative pruning. Our algorithm is complete, guaranteed to identify all solutions consistent with the data within a likelihood threshold of the optimal solution. It also deals correctly and uniformly with missing edges, which are quite common under this formulation. Tests on a number of experimental datasets and simulations with varying noise and sparsity demonstrate that our algorithm can handle significant data corruption (2.5–6.0 noisy edges per correct one) and sparsity (10–40% of the correct edges missing). In addition to the reference solution, the complete ensembles include a number (up to 30) of alternatives. We use these complete ensembles to characterize confidence in parts of an assignment.

I. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy enables analysis of protein structure, dynamics, and interactions in near-physiological conditions. While much work has focused on the use of NMR in structure determination (including at a genomic scale [1]), applications in studies of dynamics and interactions can be equally important, *even if the structure has already been determined by crystallography or modeled computationally*. NMR-based methods enable rapid and cost-effective screening for binding, and have become a vital tool in drug development [2], [3] as well as characterization of protein-protein interactions [4]. Since NMR does not require crystallization of the sample, conditions can be varied in order to study structure-function relationships [5]. Nuclear spin relaxation provides insights into protein structural dynamics (again, in solution) [6], [7].

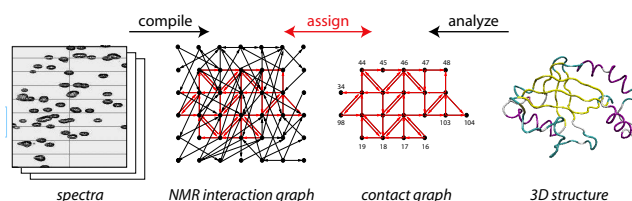


Fig. 1. Assignment given 3D structure. The goal is to find the correspondence between graph representations of the NMR data and the structure.

This paper develops a novel algorithm that exploits an available structure to interpret NMR data, supporting studies of dynamics and interactions; see Fig. 1. In the targeted scenarios, the key bottleneck is to determine the *resonance assignment* mapping NMR data to specific atoms in the protein (e.g., those affected by ligand binding). While backbone resonance assignment (i.e., resonance assignment of the backbone atoms) in support of structure determination has been well-studied (e.g., [8]–[16]), we aim essentially to invert the process, and use the structure in support of assignment. Our approach represents both the NMR data and the structure as graphs (the “NMR interaction graph” and the “contact graph”, respectively); the NMR graph is essentially a corrupted version of the contact graph, with an unknown correspondence between vertices (and thereby edges). Our goal is to find the correspondence (middle of Fig. 1). We note that this problem is different from those addressed by previous graph-based approaches [13], [15], [17], which focused on uncovering patterns in an NMR graph rather than matching it to a specified contact graph. Section IV further elaborates on the general context of our work.

Our method’s key contributions are as follows:

- It is *complete*, guaranteed to determine all solutions consistent with the data (to within a likelihood threshold of the optimal one). This is particularly important for sparse datasets and with somewhat subjective scoring functions, where we must be careful to characterize the similarities and differences among competing high-

quality solutions [14].

- It takes advantage of a *structure* to effectively decompose the solution space, and then efficiently search through it by hierarchically merging partial solutions and eliminating those that provably cannot lead to complete solutions of sufficient quality.
- It is *minimalist*, requiring only 4 spectra from ^{15}N -labeled protein, saving substantial spectrometer time and substantial expense compared to standard triple-resonance-based assignment methods.

We demonstrate the effectiveness of our algorithm in studies of a number of proteins with experimental data and with simulation studies under varying amounts of noise and sparsity.

II. METHODS

a) Graph Representation: We represent a protein structure as a *contact graph* $G^C = (V^C, E^C)$. $V^C = \{v_1^C, v_2^C, \dots\}$ is a set of residue positions. An edge in E^C represents a pair of residues for which a pair of protons is within a specified distance threshold (say, 3, 4, or 5 Å). We likewise compile a *NOESY interaction graph* $G^D = (V^D, E^D)$ from a set of four ^{15}N spectra—HSQC, HNHA, TOCSY, and NOESY (for details see [13], [15], [17]). The key source of information is the NOESY experiment, which captures through-space interactions between protons separated by a relatively close distance of at most 5 Å. $V^D = \{v_1^D, v_2^D, \dots\}$ is a set of *pseudoresidues* whose correspondence to the residues is unknown; determining it is our goal. E^D is a set of edges representing possible explanations for NOESY peaks (i.e., their atoms are interacting). Which edges are correct and which are noisy (due to ambiguity in interpreting the data) is unknown. The graph has a number of properties, following [15], [18]:

- Each vertex is labeled with a *secondary structure type*, either α or β , as determined from HNHA.
- Each vertex is labeled with an *amino acid class*. We use here the classes output by RESCUE [19], which uses two-level neural network to estimate amino acid type from proton chemical shifts. The first level associates a pseudoresidue with one of the ten type classes (IL, A, G, P, T, V, KR, FYWHDNC, EQM, and S) with very high accuracy (avg: 91.9%, min:88.1%); amino acids within a class are treated as indistinguishable.
- Each edge is directed and labeled with an *interaction type* based on the chemical shift ranges. We use only H^{N} and H^{α} , since a structure model’s side-chain atomic coordinates are usually less reliable.
- Each edge has a *match score*, w , evaluating the quality of the edge as an explanation for the peak. Here we compute the log likelihood of an edge being true by assuming the observed chemical shift difference follows a Gaussian noise distribution.

The goal is to determine an optimal correspondence (loosely, an *assignment*) between a contact graph and a corrupted version of it, an NMR interaction graph. The correspondence is scored according to the match scores of the

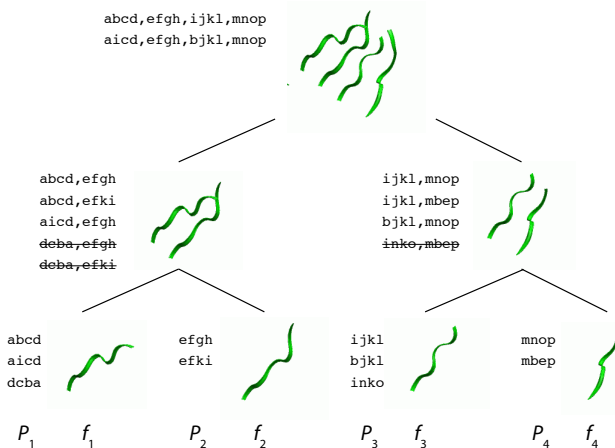


Fig. 2. Intuition for our Hierarchical Grow-and-Match algorithm. The 3D structure is decomposed into fragments; here a β -sheet is decomposed into f_1, \dots, f_4 (in practice, each strand may be composed of multiple fragments). The decomposition accounts for both contact density and the combinatorics of the sets of possible assignments; here P_1, \dots, P_4 are the sets of possible “pseudofragments”, each listing pseudoresidues (indicated by different letters) that could correspond to the fragment’s residues. Pseudofragments are merged according to a hierarchical “merge tree”; those that are inconsistent (use the same pseudoresidue) are immediately pruned and are not shown. A conservative bound (not illustrated) eliminates some combinations that provably cannot lead to a near-optimal solution; these are illustrated by strike-throughs.

NMR edges that correspond to contact edges, with penalties for extras and missings (Eq. 1, below). Matched vertices must have the same secondary structure type and amino acid class, and matched edges the same interaction type. Due to the nature of NMR experiments, there is a great deal of ambiguity in interpreting NOESY peaks; in our studies, we see an average of 2.5–6.0 noise edges for every correct one. Thus we must find the correspondence buried in significant noise.

We develop an algorithm based on two insights: multiple consistent edges are necessary to obtain effective constraint on an assignment, and sequential edges (i.e., residue i to residue $i + 1$) provide an appropriate basis for uncovering a correspondence. Noise edges in the NMR graph result from chemical shift degeneracy, and are not correlated with spatial proximity. Thus we are more confident in a set of NMR edges that consistently match a set of contact edges; it is not likely that many false positives can “conspire” to match properly. Confidence is gained by both the number of edges and their *density* (number of edges divided by number of vertices). While all contacts are important, to construct an assignment, we want to focus on edges likely to match well, and sequential residues reliably are in contact and reliably generate NOESY peaks. Furthermore, the one-dimensional structure of sequential residues makes strings of them relatively easy to manipulate. To incorporate both density and sequentiality, our algorithm first forms fragments based on sequential interactions, and then hierarchically merges them according to the contact density of edges across two fragments (Fig. 2). We now detail this process.

b) *Sequential Fragments*: As discussed above, we construct an assignment based on sequential connections. We define a *sequential fragment* $f = \langle v_{i1}^C, v_{i2}^C, \dots \rangle$ to be a sequence of vertices in V^C for a substring of the primary sequence. A corresponding *pseudofragment* $p = \langle v_{j1}^D, v_{j2}^D, \dots \rangle$ is a sequence of vertices in V^D giving an assignment for a sequential fragment. Note that while there is a natural order to V^C (the primary sequence), there is none to V^D .

Previous work [14], [16] on assignment using sequential fragments raised two important considerations: the fragments should be long enough to provide sufficient constraint, but short enough to keep under control the combinatorial number of corresponding pseudofragments. The same holds here, with an extra twist for our context: the fragments should account for contact density. We thus take advantage of the natural organization of secondary structure elements—the decomposition should “respect” helices and sheets, using them as core fragments (perhaps subdivided, to control pseudofragment combinatorics). We follow the basic previous approach [14], [16] of growing one fragment until there would be too many corresponding pseudofragments (according to a threshold θ), and then starting a new fragment. We also start a new fragment when the secondary structure type changes. If a single-residue fragment results, we merge it into the previous fragment. The result is a set $S = \{f_1, f_2, \dots\}$ of sequential fragments and a set of sets $\mathcal{P} = \{P_1, P_2, \dots\}$, where $P_i = \{p_{i1}, p_{i2}, \dots\}$ is a set of alternative pseudofragments.

c) *Merge Tree*: The decomposition of the graphs into fragments and pseudofragments yields high-quality sequential “building blocks”; we combine these based on contact density. β -sheets provide intuition (see Fig. 2)—there are relatively dense connections between adjacent strands, so it makes sense to merge sequential fragments for the strands into (no longer sequential) fragments for the sheet. More generally, let us define a *merge tree*, such that parents represent the unions of their children, and the leaves are the sequential fragments. We call the unions *fragments* (not necessarily sequential), and thereby extend the set S of sequential fragments to a set F of fragments, each of which is the union of the sequential fragment leaves below a node in the tree. The root thus represents a fragment including all residues, to which we want to assign a pseudofragment including all pseudoresidues.

In order to take advantage of contact graph density, we construct a merge tree by clustering fragments hierarchically (average linkage) according to their contacts. For the clustering similarity measure, we count the number of contacts between the residues composing the fragments; again, more contacts provide more constraints and less likelihood of a set of incorrect NMR graph edges appearing to be correct. We break ties according to amino acid composition; those with more common amino acid types are likely to have more conflicts in their matched pseudoresidues, thus enabling earlier detection of inconsistency in partial assignments.

d) *Scoring and Bounding*: An assignment (fragment and corresponding pseudofragment) must satisfy the “hard” constraints of consistency of amino acid type and secondary

structure type, and uniqueness of residues and pseudoresidues. It can then be evaluated for how well it explains the data:

$$s(f, p) = \sum_{(e^C, e^D) \in m(f, p)} w(e^D) + \phi^C(f, p) + \phi^D(f, p) \quad (1)$$

where $m(f, p)$ gives the pairs of corresponding edges induced by the residue/pseudoresidue match between f and p , $w(e^D)$ is the edge’s match score and the ϕ are penalties for missing correspondences. In our current implementation, $\phi^C(f, p)$ penalizes each missing contact edge as if it had actually appeared but with a bad score (probability ≤ 0.05). We penalize unassigned peaks via $\phi^D(f, p)$, adding the $-\log$ of the fraction of peaks that are unassigned, adopting the conservative stance of not penalizing until we can guarantee that there is no possible assignment for a peak.

The score of a partial pseudofragment (i.e., below the root in the merge tree) may let us determine that it is not worth pursuing. To be safe, we must ensure that the pseudofragment’s score, plus the best possible score for remaining fragments/pseudofragments, is not competitive with the score for a complete pseudofragment (say, more than a threshold Δ worse). Suppose that we have remaining a set F' of fragments and a set of sets \mathcal{P}' of possible corresponding pseudofragments. Ultimately, the fragments must be merged to a complete fragment, and we want to bound the score of a corresponding complete pseudofragment. We can decompose the score of such a complete pseudofragment into singleton terms (scores for edges within the individual pseudofragments) and pairwise terms (scores for edges between them). Rather than separately bounding the singleton terms and the pairwise terms (which might be minimized by inconsistent choices of pseudofragments), we “fold” the singleton terms into the pairwise ones:

$$s_2(f_i, f_j; \mathcal{P}') = \min_{p_i \in P_i, p_j \in P_j} \left(\frac{s(f_i, p_i)}{n_i} + \frac{s(f_j, p_j)}{n_j} + s(f_i, f_j, p_i, p_j) \right) \quad (2)$$

where $s(f_i, f_j, p_i, p_j)$ sums match scores of edges between the two fragments (as $s(f, p)$ does within a fragment) and n_i and n_j count the number of fragments with any edge to f_i and f_j , respectively. Thus the singleton scores are divided equally among pairs of interacting fragments, and included in their s_2 scores. Then a bound for the total score adds up all the pairwise scores.

$$b(F', \mathcal{P}') = \sum_{f_i \neq f_j \in F'} s_2(f_i, f_j; \mathcal{P}') \quad (3)$$

We can prove that Eq. 3 is a lower bound on the match score of any complete pseudofragment, and our results below (Fig. 4) also indicate that it is fairly tight.

e) *Search Algorithm*: Based on the merge tree, we can assemble larger and larger pseudofragments; based on the bound, we can eliminate pseudofragments that are guaranteed to be sufficiently suboptimal. Thus we perform Hierarchical-Grow-and-Match as a depth-first search with conservative pruning (i.e., only eliminate partial solutions guaranteed to

```

 $s_* \leftarrow \infty$ 
define HGM( $k, p, \mathcal{P}$ )
  if  $k > |F|$ 
    // complete fragment
    output  $p$ ;  $s_* \leftarrow \min \{s_*, s(f_{k-1}, p)\}$ 
  else
    let  $f_i, f_j$  be the fragments merged to form  $f_k$ 
    foreach  $(p_i, p_j) \in P_i \times P_j$ ,
      sorted by  $s(f_i \cup f_j, p_i \cup p_j)$ 
      if  $p_i \cap p_j = \emptyset$  and  $p \cap (p_i \cup p_j) = \emptyset$ 
        // no shared pseudoresidues
         $p' \leftarrow p \cup p_i \cup p_j$ 
        let  $\mathcal{P}'$  be a copy of  $\mathcal{P}$  with  $P_k$  fixed to  $\{p'\}$ 
        if  $b(F, \mathcal{P}') < s_* + \Delta$ 
          // satisfied bound
          HGM( $k + 1, p', \mathcal{P}'$ )

```

Fig. 3. Hierarchical Grow-and-Match Search. The arguments are k : fragment index; p : pseudofragment; \mathcal{P} : set of sets of alternative pseudofragments. An ensemble is output, and the best score is maintained in s_* .

be suboptimal). The search (Fig. 3) follows the structure established by the merge tree: to merge a pair of fragments, branch on the possible pairs of pseudofragments, ordered by score. (We assume binary trees, but the generalization is straightforward.) The search proceeds bottom-up, left-to-right, through the tree; the initial invocation is for the first non-leaf node, with an empty pseudofragment. We assume that the fragments are numbered accordingly— S is the set of sequential fragments and F has S followed by merged fragments in order. Thus to assemble fragment f_k , we merge fragments f_i and f_j , choosing one pseudofragment each from sets P_i and P_j . Pseudoresidues already used in earlier pseudofragments cannot be reused later in the same search branch. Upon merging the selected pseudofragments (setting P_k to the merged result), we verify that the bound is satisfied and then recurse to the next fragment in the tree (f_{k+1}), with the pseudofragment assembled so far (p') and the choices of pseudofragments for the remaining fragments (\mathcal{P}'). The algorithm will find the optimal assignment, and by setting threshold Δ to be greater than zero, it will also find a complete ensemble of nearly-optimal solutions. Variations of this depth-first approach are straightforward; e.g., we also used a beam-search-like approach that propagates several choices simultaneously, in order to more rapidly identify a solution.

III. RESULTS

Tab. I summarizes the datasets, both experimental and synthetic, that we used to validate HGM. The proteins are of moderate size for typical NMR studies. Since HGM separates residues and pseudoresidues by secondary structure type, we present here results for separate assignments of α -helices and β -sheets. For all datasets, we generated pseudofragments using a threshold (see Sec. IIb) of $\theta = 1000$. The likelihood threshold (see Sec. IIe) Δ was used to collect alternative assignments that are at most 100 times worse *a posteriori* than the optimal one.

A. Experimental datasets

We used four experimental datasets: three from previous contact-based assignment work [15], including Human Glutaredoxin (PDB ID: 1JHB), Core Binding Factor β (PDB ID: 2JHB), and the catalytic domain of GCN5 histone acetyltransferase (PDB ID: 5GCN); the fourth from the ST2NMR paper [18], *Paracoccus denitrificans* cytochrome C_{552} (PDB ID: 1QL4). For brevity, and since assignment is based on structure, we refer to each protein by its PDB ID. We constructed NMR interaction graphs from the already compiled pseudoresidues, adding vertex labels as described in the Methods. We generated edges from the the ^{15}N -edited NOE peak lists with conservatively large match tolerances: 0.05 for ^1H , 0.015 for H^{N} , and 0.35 for ^{15}N . We used a proton-proton contact threshold of 4 Å, and selected the most representative contact graph from each deposited ensemble. Overall each correct edge had a mean of 2.5–6.0 noise edges, and the average missing rate is slightly above 30%, except for 1QL4 which has a significant missing rate of 62.5%.

Tab. II summarizes assignment results by HGM for the experimental datasets. Note that the search is rather efficient, explicitly testing only a small fraction of states (i.e., number of recursive calls of the HGM algorithm). For example, if we multiply the number of pseudofragments for each fragment, there are about 1.24×10^{10} possible states for 1JHB’s β -sheets, of which 2716 are visited, and 1.93×10^{24} for 1QL4, of which 29 million are visited. The difference in number of visited states for different datasets is due to a combination of number of residues and pseudoresidues, ambiguity in amino acid type, number of edges, and noise and missing rates. For instance, 1JHB’s 43 α -helix residues were represented in the contact graph by 160 edges, of which 52 are missing in the NMR graph. This leads to a very sparse graph to match and requires HGM to search deeper before being able to safely prune a sub-optimal solution.

For all the test cases, HGM took from a few minutes to a few days (e.g. 1QL4) to conduct a complete search. The variety of its performance is highly dependent on the quality of input NMR data, as we discussed above. However, the advantage of searching according to the merge tree is clearly apparent. For example, for 1JHB’s β -sheets, the number of nodes visited in successive merge steps decreases exponentially from 2603 to 107 to 6. Even though naively, the combinatorics should increase, effective pruning eliminates most solutions. Similarly, for 2JHB’s β -sheets, this number decreases from 1.3×10^6 (max, second step) to 8 (min, final step). The same trend has been observed for other experimental datasets.

Fig. 4 illustrates the lower bound estimated by Eq. 3 relative to the optimal match score. In general, the bound is tight — the score difference is less than 8%.

Deposited solutions, determined by expert spectroscopists, serve as ‘reference’ assignments. For each dataset, HGM found the reference assignment, but also a number of alternative assignments that were about as good (or perhaps even better), under our scoring model. The largest ensemble (30 solutions)

TABLE I
DATASETS (TOP 4 EXPERIMENTAL; BOTTOM 8 SYNTHETIC)

PDB ID	BMRB Entry	α					β				
		#	# res	# edges	noise (\times)	miss (%)	#	# res	# edges	noise (\times)	miss (%)
1JHB	N/A	5	43	160	5.4	32.5	4	18	49	2.5	38.7
2JHB	4092	5	36	138	3.5	33.3	6	42	99	5.2	18.2
5GCN	4321	4	56	245	4.9	32.7	7	52	115	4.6	28.7
1QL4	4777	6	47	168	6.0	62.5	—	—	—	—	—
1KA5	2030	3	40	162	3.2, 2.8	10, 30; 21, 41	4	23	56	3.8, 2.8	10, 30; 21, 41
1EGO	2152	3	39	165	3.0, 2.2	10, 30; 21, 41	4	19	42	3.2, 2.7	10, 30; 21, 41
2NBT	1675	—	—	—	—	—	3	16	36	1.6, 1.7	10, 30; 21, 41
2FB7	7084	—	—	—	—	—	5	32	74	2.7, 3.0	10, 30; 21, 41
1G6J	5387	2	18	75	1.4, 1.4	10, 30; 21, 41	5	22	47	3.9, 3.1	10, 30; 21, 41
1P4W	5615	5	66	253	4.1, 3.8	10, 30; 21, 41	—	—	—	—	—
1SGO	6052	4	47	199	4.0, 2.9	10, 30; 21, 41	6	26	68	3.8, 3.4	10, 30; 21, 41
1RYJ	5106	—	—	—	—	—	5	27	55	3.9, 4.3	10, 30; 21, 41

Columns for each secondary structure type give number of secondary structure elements, number of residues, number of contact graph edges, average number of noisy NMR edges per contact edge and percentage of missing contact edges. For synthetic data, we specify two different sets of missing rates; listed are the corresponding noise rates. ‘—’ indicates none of that secondary structure (or only one trivial element, in the case of 1RYJ).

TABLE II
HGM RESULTS FOR EXPERIMENTAL DATA

PDB ID	2°	seq. frags	seq. pseudofrags	visited	ens. size	ref. rank
1JHB	4β	4	120–1440 (570)	2716	3	1
1JHB	5α	13	10–720 (287)	1.59×10^6	9	1
2JHB	6β	10	20–4320 (952)	2.35×10^6	4	1
2JHB	5α	11	110–1232 (693)	10434	2	1
5GCN	7β	14	16–2016 (379)	1.87×10^7	10	1
5GCN	4α	18	88–1760 (591)	1.24×10^6	1	1
1QL4	6α	10	20–672 (301)	2.94×10^7	30	2

For each PDB ID and secondary structure, columns give number of sequential fragments, min–max (mean) number of pseudofragments for each sequential fragment, number of recursive calls, size of the result ensemble and rank within the ensemble of the reference solution.

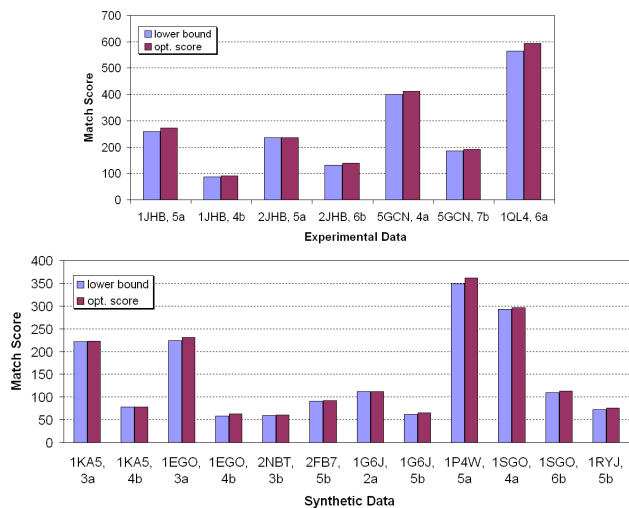


Fig. 4. Comparison of HGM-identified lower bound on match score with the score of optimal assignment for experimental (top) and synthetic (bottom), 10, 30% missing datasets.

was found for 1QL4, due to the extreme sparsity of the dataset. Even in that case, the reference solution is ranked as second best, with a score difference of 0.44 due to one difference in the assignment (a swap of the pseudoresidues mapped to residues 5 and 52).

The differences among the high-quality assignments were typically confined to swaps between a few ‘equivalent’ pseudoresidues. For example, in the α -helices of 1JHB, pseudoresidues assigned to positions 58, 90 and 91 are exchangeable and the top four solutions of the ensemble provided all the possibilities. By computing a complete ensemble of feasible assignments, HGM allows us to carefully evaluate the remaining ambiguity, as Fig. 5 illustrates. Overall, the average number of possible assignments for each residue position has been reduced from 7.9 to 1.2. The most significant ambiguity was caused by significant numbers of missing edges in particular secondary structure elements (e.g., in the first and the third helices of 1QL4). Fig. 5 doesn’t include 5GCN 4α , since it has only one solution returned by HGM. The mean number of assignments before HGM for 5GCN 4α is 11.86.

We evaluated the effect of the contact distance threshold on the performance of HGM. For example, we found that for the 5 α -helices of 2JHB, as we varied the threshold from 4 to 4.5 to 5 to 5.5 Å, the ensemble size increased from 2 to 6 to 16 to 52. The number of visited states jumped to 724719 at a threshold of 5.5 Å. Results for other proteins (not shown) were similar. With a larger threshold, more edges are included in the contact graph, more of which are missing in the NMR graph (as discussed at the start of this section). Consequently, the ‘wild-card’ scores from missing edges start to dominate the match scores of existing edges.

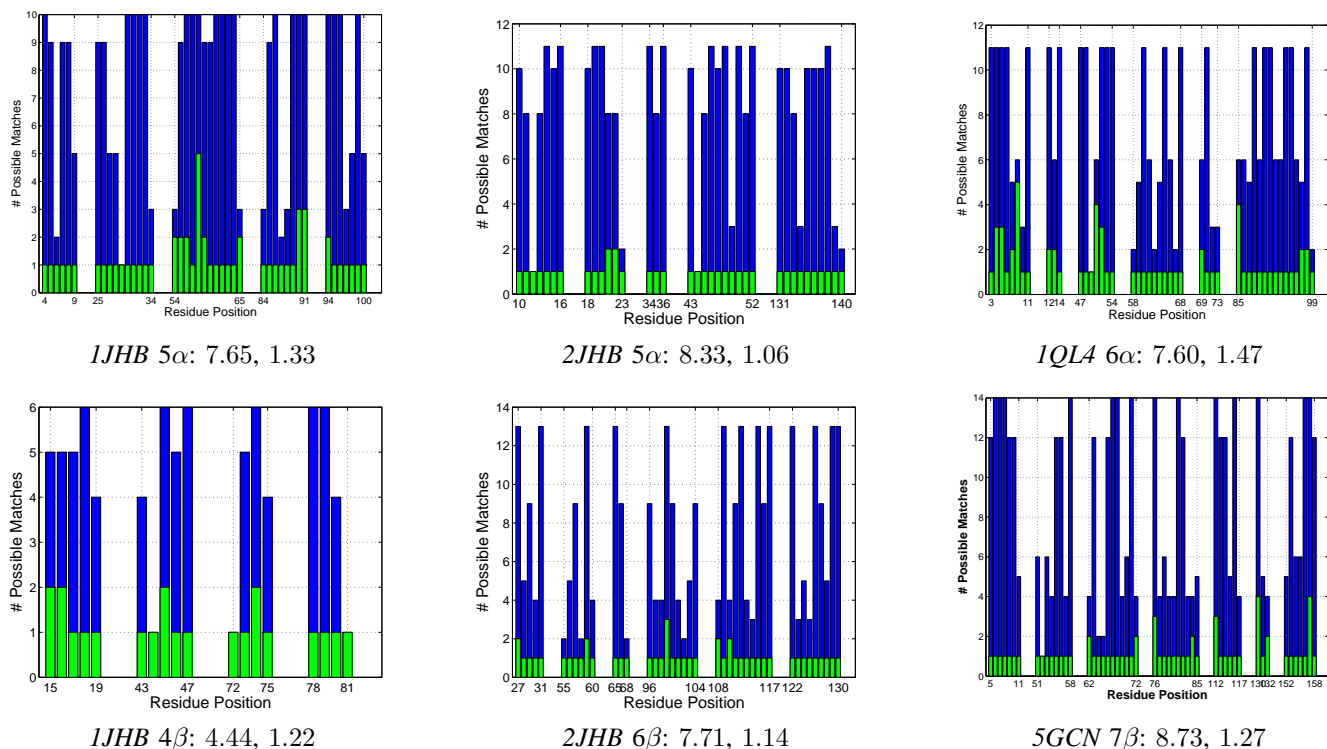


Fig. 5. Assignment ambiguity before (blue) and after (green) HGM. The bars indicate how many pseudoresidues can be mapped to each residue *a priori* and within the HGM ensemble. The means before and after HGM are listed after their PDB ids.

TABLE III
HGM RESULTS FOR SYNTHETIC DATA AT TWO DIFFERENT MISSING RATES.

PDB ID	2°	seq. frags	seq. pseudofrags	missing 10% (≤ 3 Å) and 30% (3–4 Å)			missing 21% (≤ 3 Å) and 41% (3–4 Å)		
				visited	ens. size	ref. rank	visited	ens. size	ref. rank
1KA5	4 β	5	28–2520 (937)	76	2	1	428	2	1
1KA5	3 α	11	30–720 (316)	96	1	1	236	3	1
1EGO	4 β	5	18–600 (250)	2,965	4	1	2,658	4	1
1EGO	3 α	12	10–840 (359)	618	2	1	1,291	2	1
2NBT	3 β	3	24–144 (104)	33	8	2	65	8	2
2FB7	5 β	9	47–660 (288)	1,512	4	2	140,844	8	1
1G6J	5 β	5	9–756 (455)	188	2	1	176	7	1
1G6J	2 α	3	90–540 (270)	9	4	2	6	3	1
1P4W	5 α	22	18–3960 (869)	1.75×10^6	2	1	8.95×10^6	2	1
1SGO	6 β	8	12–1080 (322)	79,542	2	1	2.52×10^6	21	1
1SGO	4 α	13	270–4050 (931)	15,472	3	1	24,982	1	1
1RYJ	5 β	8	28–1800 (400)	81,630	7	1	2.08×10^6	17	1

Columns as described in Tab. II.

To the best of our knowledge, only one algorithm, ST2NMR [18], has been developed for backbone assignment based on a 3D structure and NOESY data. ST2NMR uses a Monte Carlo approach to optimize an assignment, explaining NOESY peaks in terms of distances in the structure. It was shown to be effective for some example test data, but requires specific experimental set-ups and can provide no guarantees or insights into the information content of the data. For a comparison, we tested ST2NMR on our 3 different experimental datasets (the publication already provided results for 1QL4), but found the resulting assignment to be highly dependent on the order of the input pseudoresidues. For example, over a

set of 10 runs with different random pseudoresidue order for 1JHB, the assignment accuracy of ST2NMR varied from 28% to 77%, with a median of 56%. The results were confirmed by communication with the authors, who pointed out that ST2NMR depends critically on the combination of good structure and NOESY spectra, with good matches between them, and that ST2NMR works better with 2D homonuclear NOESY rather than 3D ^{15}N -edited NOESY. This result further emphasizes the need for complete search algorithms, like HGM, to be able to evaluate the reliability of an assignment in a situation with sparse, noisy data.

In order to study our algorithm’s performance under varying noise and sparsity levels, we also generated synthetic datasets using chemical shift data deposited in the BMRB [20]. We chose a random set of eight moderate-sized proteins previously tested with RESCUE [19], with varying α -helix and β -sheet content (refer again to Tab. I). To construct the NMR interaction graph, we first simulated NOE peaks for pairs of interresidue backbone protons within a distance ≤ 4 Å. We likewise restricted the contact graph to 4 Å, thereby essentially ignoring the longer-distance NOEs, since we found in the experimental data that they are missing at a significant frequency and thus the information they provide is not worth the additional computational complexity they require. We randomly deleted peaks according to observed statistics correlating the missing probability with the interatomic distance [21], and tested two different missing rates at different distances: $d \leq 3$ Å, missing either 10% or 21%; $3 < d \leq 4$ Å, missing either 30% or 41%. Note that these rate ranges cover the rates observed in the experimental datasets, except for the extremely sparse IQL4. We generated an “observed” interresidue proton chemical shift for each remaining peak by adding Gaussian noise with variance 0.02 (corresponding to the 0.05 ^1H match tolerance). We added an edge to the NMR graph for each proton whose chemical shift matches the noisy value within the 0.05 threshold, yielding an average of 1.4–4.3 noise edges per correct edge. As discussed [15], all the noise is on the interresidue side of the interaction, since the intraresidue side has two chemical shifts (H^{N} , ^{15}N) by which to resolve chemical shift ambiguity. The simulated noise rates are somewhat smaller than the experimental ones, but we did not consider it realistic to increase the chemical shift tolerance in order to artificially inflate them.

Tab. III summarizes the results of HGM on the synthetic datasets. Here reference solutions indicate the original BMRB assignments. As with experimental data, for all test cases the reference assignment is included with an optimal or near-optimal score, and only a few assignment swaps differentiate the best solutions. Fig. 4 illustrates that the lower bound remains quite tight. In general, the HGM algorithm performed very well and only explicitly tested a tiny fraction of the search space before finding the complete solution set. The synthetic tests further demonstrate the effect observed before: an increase in the missing rate yields greater search complexity and ensemble size (the last two sets of columns). These results also indicate that, given a uniform missing rate, α -helices tend to have a better tolerance for missings than do β -sheets since their tertiary structures are usually more compact and thus generate more edge constraints. Also, assignment of β -sheets benefits from the hierarchical merge order which naturally utilizes the spatial proximity of β -strands, even when sequentially separated. In such cases, a poor local assignment, e.g., many missing edges in a strand, can be effectively overcome by way of connections with neighboring strands.

Our work differs from most traditional approaches to backbone resonance assignment in that it is structure based. As we showed in the Results, our guarantee of completeness appears to be very valuable, as ST2NMR (the only other existing approach using structure+NOESY) did not perform well on our sparse, noisy datasets. Our work is complementary to structure-based assignment approaches that reply primarily on experiments other than NOESY. For example, the NVR work by Langmead and Donald [22], [23] uses residual dipolar coupling (RDC) data as global orientational restraints, with only unambiguous NOEs to help prune. It remains interesting future work to fully integrate RDC with NOESY data and thereby perhaps overcome their individual limitations.

There are other techniques for assignment based on the NOESY, but they do not use information from an available 3D structure. The Main-Chain Directed approach represents an early approach to backbone assignment based on the NOESY [24], [25], although that work was developed for homonuclear spectra, only partially automated, and applied to experimental data for only one small protein. The automated Jigsaw approach [17] was successfully applied to uncover α -helix and β -sheet patterns in NOESY data (and thereby assign those regions). More recent work developed an algorithmic basis for the Jigsaw-style approach, with a randomized algorithm that gives optimal performance in expected polynomial time for the special case of uncovering secondary structures in corrupted NMR graphs [13], [15]. We note that we are focusing here on backbone assignment based on the NOESY. Work on NOE assignment (e.g., [26]), including side-chain interactions, is certainly related but typically is addressed only once backbone resonances have been assigned by standard techniques. An algorithm combining these two aspects for simultaneous backbone and side-chain assignment would represent an interesting, and significant, advance.

In our focus on matching graph representations of protein structures, our work is somewhat like sequence-structure alignment (threading); e.g., Xu and co-workers developed a divide-and-conquer approach that uses hierarchical combinations of sub-alignments [27] and can incorporate *assigned* NOE data to constrain them [10]. A significant difference is that for threading, residues are in sequential order for both the sequence and the structure, while in contact-based assignment, the pseudoresidues come with no explicit order.

V. CONCLUSION

To reduce the time and expense of NMR-based studies of protein interactions and dynamics, we develop an algorithm to find *all* feasible mappings between a contact graph encoding the structure and a corrupted version encoding the NMR data, limiting the combinatorial explosion by hierarchically decomposing the structure and effectively pruning partial solutions. Tests on both experimental and synthetic data show that the algorithm handles significant noise and sparsity in assigning relatively contact-dense regions (α -helices and β -sheets).

An important step for practical utility of HGM is to characterize its performance on x-ray structures and homology models, the natural inputs for structure-based assignment. We must model and account for any systematic differences between these models (generating contact graph edges) and the native solution state probed by NMR (generating NMR graph edges). Since HGM focuses on high-contact-density regions and the shortest edges, we believe that it will be fairly robust to modest structural differences. Variable correspondence and lack of correspondence in different subgraphs may also provide evidence to help select among models, as was previously done for RDC data by Langmead and Donald [28].

Preliminary tests also indicate that our method can naturally extend from secondary structure to connected loops, and thereby assign larger portions of the structure. An interesting aspect of NOESY-based assignment is that the data are inherently local, thereby providing the possibility of assigning different portions of the structure to different confidence levels.

Source code of HGM is freely available for academic use upon request to the contact author.

ACKNOWLEDGMENT

We thank members of the CBK Lab for helpful discussions on this work. We also thank Dr. Pristovek for his clarification of the usage of ST2NMR. This work was supported in part by an NSF CAREER award to CBK (IIS-0444544).

REFERENCES

- [1] J. Prestegard, H. Valafar, J. Glushka, and F. Tian, "Nuclear magnetic resonance in the era of structural genomics," *Biochemistry*, vol. 40, pp. 8677–8685, 2001.
- [2] S. Shuker, P. Hajduk, R. Meadows, and S. Fesik, "Discovering high-affinity ligands for proteins: SAR by NMR," *Science*, vol. 274, pp. 1531–1534, 1996.
- [3] P. Hajduk, R. Meadows, and S. Fesik, "Drug design: Discovering high-affinity ligands for proteins," *Science*, vol. 278, pp. 497–499, 1997.
- [4] Y. Chen, J. Reizer, M. Saier Jr., W. Fairbrother, and P. E. Wright, "Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc," *Biochemistry*, vol. 32, pp. 32–37, 1993.
- [5] G. Montelione, D. Zheng, Y. Huang, K. Gunsalus, and T. Szyperski, "Protein NMR spectroscopy in structural genomics," *Nat. Struct. Biol.*, vol. 7 Suppl, pp. 982–985, 2000.
- [6] A. Palmer III, J. Williams, and A. McDermott, "Nuclear magnetic resonance studies of biopolymer dynamics," *J. Phys. Chem.*, vol. 100, pp. 13 293–13 310, 1996.
- [7] L. Kay, "Protein dynamics from NMR," *Nat. Struct. Biol.*, vol. 5 Suppl, pp. 513–517, 1998.
- [8] D. Zimmerman, C. Kulikowski, 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. Biol.*, vol. 269, pp. 592–610, 1997.
- [9] H. Moseley and G. Montelione, "Automated analysis of NMR assignments and structures for proteins," *Curr. Opin. Struct. Biol.*, vol. 9, pp. 635–642, 1999.
- [10] Y. Xu, D. Xu, O. Crawford, J. Einstein, and E. Serpersu, "Protein structure determination using protein threading and sparse NMR data," in *Proc. RECOMB*, 2000, pp. 299–307.
- [11] G. Lin, D. Xu, Z.-Z. Chen, T. Jiang, and Y. Xu, "A branch-and-bound algorithm for assignment of protein backbone NMR peaks," in *Proc. CSB*, 2002, pp. 165–174.
- [12] J. Jung and M. Zweckstetter, "MARS - robust automatic backbone assignment of proteins," *J. Biomol. NMR*, vol. 30, pp. 11–32, 2004.
- [13] C. Bailey-Kellogg, S. Chainraj, and G. Pandurangan, "A random graph approach to NMR sequential assignment," *J. Comp. Bio.*, vol. 12, pp. 569–583, 2005.
- [14] O. Vitek, C. Bailey-Kellogg, B. Craig, P. Kuliniewicz, and J. Vitek, "Reconsidering complete search algorithms for protein backbone NMR Assignment," *Bioinformatics*, vol. 21, pp. ii230–236, 2005.
- [15] H. Kamisetty, C. Bailey-Kellogg, and G. Pandurangan, "An efficient randomized algorithm for contact-based nmr backbone resonance assignment," *Bioinformatics*, vol. 22, pp. 172–180, 2006.
- [16] O. Vitek, C. Bailey-Kellogg, B. Craig, and J. Vitek, "Inferential backbone assignment for sparse data," *J. Biomol. NMR*, vol. 35, pp. 187–208, 2006.
- [17] C. Bailey-Kellogg, A. Widge, J. Kelley III, M. Berardi, J. Bushweller, and B. Donald, "The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data," *J. Comp. Bio.*, vol. 7, pp. 537–558, 2000.
- [18] P. Pristovek, H. Ruterjans, and R. Jerala, "Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—the program st2nmr," *J. Comp. Chem.*, vol. 23, pp. 335–340, 2002.
- [19] J. Pons and M. Delsuc, "RESCUE: An artificial neural network tool for the NMR spectral assignment of proteins," *J. Biomol. NMR*, vol. 15, pp. 15–26, 1999.
- [20] B. Seavey, E. Farr, W. Westler, and J. Markley, "A relational database for sequence-specific protein NMR data," *J. Biomol. NMR*, vol. 1, pp. 217–236, 1991.
- [21] J. Doreleijers, M. Raves, T. Rullmann, and R. Kaptein, "Completeness of NOEs in protein structures: A statistical analysis of NMR data," *J. Biomol. NMR*, vol. 14, pp. 123–132, 1999.
- [22] C. Langmead and B. Donald, "An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments," *J. Biomol. NMR*, vol. 29, pp. 111–138, 2004.
- [23] C. Langmead, A. Yan, R. Lilien, L. Wang, and B. Donald, "A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments," *J. Comp. Bio.*, vol. 11, pp. 277–298, 2004.
- [24] D. D. Stefano and A. Wand, "Two-dimensional ^1H NMR study of human ubiquitin: a main-chain directed assignment and structure analysis," *Biochemistry*, vol. 26, pp. 7272–7281, 1987.
- [25] S. Nelson, D. Schneider, and A. Wand, "Implementation of the main chain directed assignment strategy," *Biophys. J.*, vol. 59, pp. 1113–1122, 1991.
- [26] L. Wang and B. Donald, "An efficient and accurate algorithm for assigning nuclear overhauser effect restraints using a rotamer library ensemble and residual dipolar couplings," in *Proc. CSB*, 2005, pp. 189–202.
- [27] Y. Xu and D. Xu, "Protein threading using PROSPECT: Design and evaluation," *Proteins*, vol. 40, pp. 343–354, 2000.
- [28] C. Langmead and B. Donald, "High-throughput 3D structural homology detection via NMR resonance assignment," in *Proc. CSB*, 2004, pp. 278–289.