Contact replacement for NMR resonance assignment

Fei Xiong1, Gopal Pandurangan2,* and Chris Bailey-Kellogg1,∗
1Department of Computer Science, Dartmouth College, Hanover, NH 03755 and 2Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

ABSTRACT
Motivation: Complementing its traditional role in structural studies of proteins, nuclear magnetic resonance (NMR) spectroscopy is playing an increasingly important role in functional studies. NMR dynamics experiments characterize motions involved in target recognition, ligand binding, etc., while NMR chemical shift perturbation experiments identify and localize protein-protein and protein-ligand interactions. The key bottleneck in these studies is to determine the backbone resonance assignment, which allows spectral peaks to be mapped to specific atoms. This article develops a novel approach to address that bottleneck, exploiting an available X-ray structure or homology model to assign the entire backbone from a set of relatively fast and cheap NMR experiments.

Results: We formulate contact replacement for resonance assignment as the problem of computing correspondences between a contact graph representing the structure and an NMR graph representing the data; the NMR graph is a significantly corrupted, ambiguous version of the contact graph. We first show that by combining connectivity and amino acid type information, and exploiting the random structure of the noise, one can provably determine unique correspondences in polynomial time with high probability, even in the presence of significant noise (a constant number of noisy edges per vertex). We then detail an efficient randomized algorithm and show that, over a variety of experimental and synthetic datasets, it is robust to typical levels of structural variation (1–2 Å), noise (250–600%) and misalignments (10–40%). Our algorithm achieves very good overall assignment accuracy, above 80% in α-helices, 70% in β-sheets and 60% in loop regions.

Availability: Our contact replacement algorithm is implemented in platform-independent Python code. The software can be freely obtained for academic use by request from the authors.
Contact: gopal@cs.purdue.edu; cbk@cs.dartmouth.edu

1 INTRODUCTION
Nuclear magnetic resonance (NMR) spectroscopy is playing an increasingly important role in studies of proteins beyond the determination of their 3D structures. For example, since NMR is performed in solution, it can gather information regarding dynamics (Kay, 1998; Palmer III et al., 1996) and structure-function relationships under varying conditions (Montelione et al., 2000). Similarly, solution NMR is a vital tool in assessing ligand binding for drug development (Hajduk et al., 1997; Shuker et al., 1996) and can also help characterize protein-protein interactions (Chen et al., 1993). These applications of NMR are significant even if the structure has already been determined by X-ray crystallography or a high quality homology model is available.

Unfortunately, the data collected in NMR studies are in terms of the resonance frequencies of the atomic nuclei, which are not readily predictable. “Resonance assignment” determines the previously unknown mapping between the atoms in the protein and the observed resonance frequencies, so that the information about binding, dynamics, etc. can be properly interpreted. Backbone resonance assignment has been well-studied within the context of structure determination (Bartels et al., 1997; Jung and Zweckstetter, 2004; Lin et al., 2002; Mosley and Montelione, 1999; Vitek et al., 2005, 2006; Xu et al., 2000; Zimmerman et al., 1997). However, the standard protocols used in that context require much more (and more expensive) experimentation than is necessary for the dynamics and interaction studies mentioned above. While the standard protocols and assignment approaches could still be employed for those studies, using an available structure offers the potential to reduce the experimental complexity and circumvent traditional barriers to interpretation. Our goal is to develop computational techniques that enable assignment from a minimal set of experiments that require only 15N-labeled sample rather than the much more expensive 13C, 15N-labeling used in standard protocols.

Here we formulate the problem of assignment given a structure and minimalistic NMR data as the contact replacement problem (Fig. 1). A contact graph representing a protein structure has vertices for the individual amino acid residues in the protein and edges between nearby pairs. A particular form of ‘interaction graph’ representing NMR data has vertices for NMR-probed ‘pseudoresidues’ (which correspond via an unknown mapping to the real residues), and edges between pairs that, if they were nearby, would explain the data. The NMR edges are essentially the contact edges, significantly corrupted by experimental noise and ambiguity (around 5 noisy edges per correct one). The contact replacement problem is then to uncover the correspondences between these graphs for a given protein.

The name ‘contact replacement’ for our problem is inspired by the names for the analogous problems ‘molecular replacement’ in X-ray crystallography (Rossman and Blow, 1962) and ‘nuclear vector replacement’ in NMR (Langmead and Donald, 2004; Langmead et al., 2004). In molecular replacement, initial data interpretation is aided by matching against available structural information from a related protein. Likewise, in nuclear vector replacement, residual dipolar coupling data are matched against predictions from an available structure (or high-quality model). Contact replacement and nuclear vector replacement are complementary, relying on different types of NMR data with different information content (distances versus orientations). The contact replacement problem is related to threading (sequence–structure alignment), but for threading, residues are in sequential order for both the sequence and the
We first show that by combining connectivity and type, and by structure. The ST2NMR program (Pristovek combinatorics of the problem and the branch-and-bound approach, can handle significant noise and sparsity. However, due to the correspondences between contact graphs and NMR graphs, and and-bound algorithm to find the complete ensemble of consistent structural variation in the contact graph.

Our earlier work on random graph algorithms (Bailey-Kellogg, 2007): PEPMORPH (Erdmann and Rule, 2002) casts assignment given a 3D structure and NMR data as an optimization problem, and uses a Monte Carlo approach to find explanations of the data in terms of distances in the structure. While ST2NMR was shown to be effective for some test data, it requires very specific experimental set-ups and can provide no guarantees or insights into the interaction graph representing the NMR data is essentially a corrupted, ambiguous version of the contact graph representing the structure. The goal is to uncover the correspondences.

Various versions of what we are calling here contact replacement have previously been studied. Our Hierarchical Grow-and-Match (HGM) algorithm (Xiong and Bailey-Kellogg, 2007) uses a branch-and-bound approach to find the complete ensemble of consistent correspondences between contact graphs and NMR graphs, and can handle significant noise and sparsity. However, due to the combinatorics of the problem and the branch-and-bound approach, HGM is effectively restricted to well-defined regions of secondary structure. The ST2NMR program (Pristovek et al., 2002) casts assignment given a 3D structure and NMR data as an optimization problem, and uses a Monte Carlo approach to find explanations of the data in terms of distances in the structure. While ST2NMR was shown to be effective for some test data, it requires very specific experimental set-ups and can provide no guarantees or insights into the interaction content of the data. We tested it on a number of different datasets, and found the accuracy to be fairly low and quite sensitive to the order of the input data (Xiong and Bailey-Kellogg, 2007). PEPMORPH (Erdmann and Rule, 2002) uses graph representations of the structure and data, but augments them with residual dipolar coupling data in order to compute matchings. Our earlier work on graph-based approaches to NMR assignment, Jigsaw (Bailey-Kellogg et al., 2000) and random graph algorithms (Bailey-Kellogg et al., 2000; Kamisetty et al., 2006), were able to effectively uncover secondary structure patterns; our random graph model enabled us to prove that the randomized methods have optimal performance in expected polynomial time. However, these approaches were all restricted to uncovering generic prototypes of secondary structure elements, rather than matching NMR data to an arbitrary 3D structure.

**2 APPROACH**

We first summarize the representations of the input contact graph and NMR interaction graph; for details see Bailey-Kellogg et al. (2000, 2005) and Kamisetty et al. (2006).

**Contact graph:** $G^c = (V^c, E^c)$, where $V^c$ is a set of residue positions and $E^c$ is a set of pairs of nearby residue positions. In particular, we place an edge when a pair of protons is within a specified distance threshold (say, 3, 4 or 5 Å). Each vertex $v$ is labeled with its amino acid type, $\alpha(i)$. $NMR$ interaction graph: $G = (V, E)$, where $V$ is a set of *pseudoresidues* of unknown correspondence to the residues and $E$ is a set of pairs of pseudoresidues that may have interacting protons (i.e. an interaction would explain a peak in the NOE$\Sigma$ spectrum). Such a graph can be compiled from a set of four 1$^H$N spectra (HSQC, HNHA, TOCSY and NOESY), and has a number of properties (Kamisetty et al., 2006; Pristovek et al., 2002; Xiong and Bailey-Kellogg, 2007):

- Each vertex is labeled with a *secondary structure type*, either $\alpha$ or $\beta$, as determined from HNHA.
- Each vertex is labeled with a list $\ell$ of *possible amino acid types*. We use here the classes output by RESCUE (Pons and Deluc, 1999), which employs a two-level neural network to estimate amino acid type from proton chemical shifts. The first-level associates a pseudoresidue with one of the 10 type classes....

![Contact replacement](https://example.com/image.png)
is computed as $G^* = (V^*, E^*)$ and an NMR interaction graph $G = (V, E)$ data is substantially corrupted from and maximizes the score of the edges in $E$ that correspond to edges in Hamiltonian graphs with degree at most three (Garey, 1997) compare absolute or squared difference in chemical shift, except for noise (reasonably modeled as Gaussian), the corresponding Hamiltonian path should match exactly and have the best score. Here we score edges by error probability, i.e. how likely it is that an edge could be generated by noise. In this way, missing edges are naturally penalized since they contribute a score of zero. The score is thus $-\log \left( 1 - \exp \left( \frac{-\Delta(\epsilon)}{2n} \right) \right)$ where $\Delta(\epsilon)$ is the chemical shift difference for edge $e$, and $\sigma$ is the SD of chemical shift difference distribution.

Due to the nature of the $^{15}$N NOESY ($^{1}$H$^{1}$N for one vertex and $^{1}$H for the other), the NMR interaction graph is directed. For consistency, we adopt the same convention for the contact graph.

We assume for simplicity that the contact graph is ‘correct’—it represents exactly those interactions that are physically present (thus its designation as $G^*$ = $(V^*, E^*)$), and all the errors are in the NMR graph. The NMR interaction graph $G$ constructed from NMR data is substantially corrupted from $G^*$, and has an anomalous vertex correspondence. We now formalize our problem in its cleanest form.

**Problem 1.** (Contact replacement). We are given a contact graph $G^* = (V^*, E^*)$ and an NMR interaction graph $G = (V, E)$. The goal is to find a bijection $m$ from $V^*$ to $V$ that matches amino acid classes and maximizes the score of the edges in $E$ that correspond to edges in $E^*$. Formally, if $m(u^*) = v$, then we must have $a(u^*) = \bar{m}(v)$ for all $(u^*, v^*) \in E^*$. The score is computed as $\sum_{(u^*, v^*) \in E^*} \delta(\epsilon)$, where the mapping $c$ between $E^*$ and $E$ is induced by $m$ as $c = \{(a(u^*), (u^*, v^*))(a(u), (u, v)) \in E^*, (u, v) \in E, m(u^*) = u, m(v^*) = v\}$.

One feature of proteins particularly relevant here is that they are made of chains of amino acids. Thus the contact graph has an embedded Hamiltonian path from N terminus to C terminus (in addition to numerous through-space edges connecting residues at any sequential distance). Ignoring missing edges, the NMR graph has a corresponding Hamiltonian path. Our analysis and randomized algorithm both make use of this property, by focusing on finding the Hamiltonian path while ‘bringing along’ the additional edges for scoring purposes.

We note that the contact replacement problem is NP-hard in general, since it contains as a special case the following NP-hard problem: Given a unweighted Hamiltonian graph (undirected or directed) $H$ find a Hamiltonian path in $H$ (i.e. assuming that there are no constraints on vertex labels and all edge scores are the same). We note that the above problem remains NP-hard even when restricted to sparse Hamiltonian graphs, e.g. directed Hamiltonian graphs with maximum out-degree two (Plesnik, 1979) or undirected Hamiltonian graphs with degree at most three (Garey et al., 1976).

The problem has been shown hard to approximate in directed graphs: it is not possible to find paths even of superpolynomial length in constant out-degree Hamiltonian graphs unless Satisfiability can be solved in subexponential time (Bjorklund et al., 2003). For undirected Hamiltonian graphs, the best known algorithms give longer paths (e.g. length $\sqrt{2n}/\log n$) in Hamiltonian graphs in polynomial time (Feder and Motwani, 2005; Feder et al., 2002; Gapow, 2004). We note that the above algorithmic results do not apply to our problem because we have additional information (amino acid classes) for the vertices. More importantly, NMR interaction graphs are not arbitrary graphs and indeed have a special structure as captured by the random graph model described next.

### 2.1 Random graph model for NMR interaction graph

In order to develop and analyze effective algorithms, we must consider and model the nature of the relationship between the ideal contact graph $G^*$ and the observed NMR interaction graph $G$. We note that traditional $G(n, p)$ random graph models (Bollobas, 2001) essentially add noisy edges randomly and independently. However, the noisy edges in an NMR interaction graph are not arbitrarily distributed. Instead, chemical shift degeneracy is the key source of noise in these graphs, imposing a particular correlation structure among noise edges. We have developed a random graph model that properly captures the noise in NMR interaction graphs (Bailey-Kellogg et al., 2005; Kamisetty et al., 2006).

**Definition 1.** (Model $M(G^*, w)$ random graph). The model $M(G^*, w)$ ‘generates’ a random graph from the (correct) graph $G^*$, where $w$ is a parameter that determines the number of noisy edges generated per correct edge. Let $\pi$ be a random permutation of $V^*$. Denote by $\pi(v)$ the index of $v \in V^*$ in the permutation. We then consider an ambiguous all vertices within a ‘window’ of size $w$ around a particular vertex. For each edge of $G^*$, additional edges are generated as follows. Consider an edge $(u^*, v^*) \in E^*$. Then for each $u$ in the window of width $w$ around $\pi(u^*)$, i.e. $|\pi(u^*) - \pi(u)| \leq w$ we add the edge $(u, v^*)$ to the random graph.

This model captures the way in which uncertainty in the data leads directly to ambiguity in the edges posited in an NMR graph. In particular, NMR spectra reflect interactions between atoms as peaks in $R^2$ or $R^3$, where each dimension indicates the coordinates (resonance frequencies, in units called ‘chemical shifts’) of one of the interacting atoms. Uncertainty in the measured chemical shifts of the protons thus leads to ambiguity in matches, and the construction of noise edges. When two vertices have atoms that are similar in chemical shift, they will tend to share edges—each edge for the one will also appear for the other. Since there is no systematic, global correlation between chemical shifts and positions of atoms in the primary sequence or in space, we simply model chemical shift similarity according to a random permutation. The model can be extended in order to generate synthetic data (e.g. incorporating edge scores, accounting for missing edges, etc.); see Section 5 for our actual simulation testbed. We use this basic model in the next section to analyze the contact replacement problem.

For amino acid types, we assume a simple independent model for the purpose of analysis. (Of course, in practice we know the actual amino acid types for the contact graph.) In particular, let $A$ be the set of amino acid types and $D$ be some fixed probability distribution over $A$ (e.g. 1/20 for each, or using empirically observed frequencies). Assume without loss of generality that for all $a \in A$, the probability that $a$ is chosen, $Pr(a)$, is greater than $q$, where $q > 0$ is...
some fixed constant. We assume that a vertex is labeled by sampling independently at random from $D$.

3 THEORETICAL ANALYSIS AND IMPLICATIONS

We present a theoretical analysis to show that the contact replacement problem can be solved with high probability in polynomial time. For the analysis, we assume that the NMR interaction graph $G = M(G^*, w)$ is generated from the correct contact graph $G^* = (V^*, E^*)$ which is a Hamiltonian path of length $n$ ($= 28$) of amino acids. Now, for $w = O(1)$, we know that there exists a subgraph $H$ of $G$, denoted as $H^*$, such that for every $v_1, v_2$ such that $v_1 \in V_1, v_2 \in V_2$, $(m(v_1)) \in (v_1)$ and there is an edge $e = (u_1, v_1) \in E_1$ if there is an edge $(m(u_1), m(v_1)) \in E_2$.

In the following, ‘with high probability (whp)’ means with probability at least $1 - 1/n^2(1)$ where $n$ is the number of amino acids in the protein.

**Theorem 1.** Under our $M(G^*, w)$ random graph model, if $w = O(1)$, then the contact replacement problem can be solved in polynomial time whp.

**Proof.** Without loss of generality, we will assume that $|A| = 2$ (A is the set of amino acid types). The proof can easily be made to work without this restriction. The proof hinges on the following claims. The first term is the number of ways of choosing vertices from which the noisy edges emanate. The third term bounds the probability that the noisy edges form a path between them with the amino acid labels matching those of the corresponding vertices in $P$. The last term is the probability that the amino acid labels for the correct vertices match. We can bound the sum as follows (note that we take $\infty = 1$):

$$\Pr[H' = P] \leq \sum_{k=0}^{k-1} n^k q^k (\log n)^k$$

Plugging $k = c \log n$, the above sum is bounded by

$$\Pr[H' = P] \leq n \sum_{k=0}^{\log n} \binom{n}{k} q^k \log^k n$$

$$\leq n \sum_{k=0}^{c \log n} \binom{n}{k} q^k \log^k n$$

There are some subgraph $Q$ of $G^*$ such that $P = Q$.

The above analysis shows that the contact replacement problem can be solved in polynomial time if $w = O(1)$, i.e. there is at most
a constant number of noisy edges per vertex. This is significant for two reasons. First, in practice, typically the number of noisy edges per vertex is a constant (around 5). Second, if there is no amino acid information, the randomized algorithm of Bailey-Kellogg et al., 2005; Pandurangan, 2005) can find long paths (of length at least G(n, log n)) in polynomial time only if the number of noisy edges per vertex is at most one. Our analysis here shows that this threshold barrier can be surmounted by using amino acid type information. Our experimental results validate this theoretical prediction.

4 METHODS

In practice, the simplified model and algorithm used in the analysis may not be fully applicable, in particular because some edges may be missing and some amino acid type information may be erroneous (the correct type for a contact graph vertex not included in the class for the corresponding NMR vertex). Such errors result in ‘breaks’ in the correspondence between a contact graph and NMR graph. Thus we seek to find a set of disjoint paths (‘fragments’) in the NMR graph that together match the Hamiltonian path in the contact graph. Given such an equivalence, we score all corresponding edges, including the non-sequential ones. By biasing our algorithm on paths, we take advantage of our long-path result from the previous section, while by including all edges in the score, we take advantage of all available information to better control the search.

A key insight of our algorithm is that in searching for good matchings, the best ones tend to share a lot of substructure. (Our results below on assignment ambiguity, Fig. 5, illustrate.) In branching-based searches, such shared substructure can appear on many different branches, making exhaustive search very inefficient and causing backtracking to perform wasteful undoing and redoing. In contrast, we use more efficient local fixes to resolve inconsistencies and continue searching with most of the structure still intact. Figure 2 gives the pseudocode for our algorithm. The algorithm maintains (and fixes up) a single set F of fragments, with a mapping m to the contact graph that is always consistent (i.e. fragments do not overlap). Some fragments may not be mapped, meaning that under the current matching, they are considered noise. On each iteration, the algorithm sequentially extends one fragment, adding an NMR vertex that will correspond to the next residue position in the sequence. Several things could happen upon growing to that vertex; see Figure 3. In the simplest case, the algorithm picks up an unmatched NMR vertex (and its fragment) and simply extends the matching. However, it may run into a conflict and need to fix up the current matching. If a fragment wants to grow to a vertex in the middle of another fragment, then the other fragment is split at the point of conflict to allow its suffix to be taken away. If the growth results in a mismatch of amino acid type or of alignment, then a realignment is attempted. Matching the fragment somewhere else in the contact graph may result in a consistent matching, or may produce another conflict, potentially fixed by replacing part of the conflicting fragment with the new fragment. To keep each step simple enough, we only recursively handle the conflict at this point if it’s simple enough to fix. The algorithm repeats until convergence. In practice, we run a fixed number of iterations, and keep track of m through the iterations in order to analyze the distribution of good solutions.

At several places in the algorithm, we choose an option ‘with probability according to its score.’ In general, the score refers to the total score of NMR edges matched to contact graph edges (refer again to the graph definition for our scoring function). Since we are using discrete amino acid classes, we require that the matched contact amino acid type be a member of the NMR amino acid class. In choosing an edge from u, we only consider the edges along the current path, while in choosing an alignment or whether or not to splice, we consider the total of all edges before versus after the possible change.

\begin{verbatim}
\begin{algorithm}
  \caption{Randomized algorithm for contact replacement: given a contact graph \( G' = (\mathcal{V}', \mathcal{E}') \) and NMR graph \( G = (\mathcal{V}, \mathcal{E}) \), determine the matching \( m \).
  \end{algorithm}
\end{verbatim}

5 RESULTS

Table 1 summarizes the datasets, both experimental and synthetic, that we used to validate our algorithm. The proteins are of moderate size for typical NMR studies, and this collection has representative structural diversity and assignment difficulty. We used three experimental datasets from previous contact-based assignment work (Kamisetty et al., 2006; Xiong and Bailey-Kellogg, 2007), including human glutaredoxin (PDB ID: 1JHB), core binding factor \( \beta \) (PDB ID: 2JHB) and the catalytic domain of GCN5 histone acetyltranferase (PDB ID: 5GCN). For brevity, and since assignment is based on structure, we refer to each protein by its PDB ID. The noise rate (average number of noisy NMR edges per contact edge) is as high as 5.4 (1JHB \( \alpha \)-helices) and the missing rate as high as 51.8\% (5GCN loops). Since such complete experimental datasets are a rare commodity, in order to more broadly test our approach, we also used a set of previously generated synthetic datasets (Xiong and Bailey-Kellogg, 2007) based on chemical shift data deposited in the BMRB. These synthetic datasets include noise edges according to Gaussian noise with variance 0.02 (corresponding to a standard 0.05 \( \pi \) match tolerance) and missing edges according to observed statistics correlating the missing probability with the interatomic distance (Doreleijers et al., 1999): \( d \leq 3 \, \text{Å}, \text{missing \% 21; } 3 < d \leq 4 \, \text{Å}, \text{missing \% 41.} \)
Datasets (top 3 experimental; bottom 9 synthetic)

Table 1.

| PDB ID | RMRB Entry | α/β/loop | No. of elements | No. of residues | No. of edges | Noise(σ) | Missing(%) | RMSD(Å) |
|--------|------------|-----------|-----------------|----------------|-------------|-----------|------------|----------|
| 1JBB   | N/A        | /4/10     | 43/18/44        | 160/498/81     | 5.4/2/3.0   | 32.5/38/73.3 | 1.30/81.6 |
| 1JBB   | 4092       | 5/6/11    | 36/42/64        | 138/99/141     | 3.5/5.2/3.7 | 33.3/18/241.1 | 1.50/92.6 |
| 5GCN   | 4321       | 4/12      | 56/52/58        | 245/115/110    | 4.0/4/2.2   | 32.7/28/75.1 | 1.51/60.5 |
| 1KAS   | 2030       | 3/4/8     | 40/23/52        | 162/56/58      | 3.2/2/8.19  | 23, 41     | 0.80/70.8 |
| 1EGO   | 2152       | 3/4/8     | 39/19/27        | 165/42/49      | 2.2/2/7.26  | 23, 41     | 2.1/14/5.6 |
| 2FB7   | 7084       | – /3/6   | – /32/63        | – /74/96       | – /0.2/4.4  | 23, 41     | – /1.6/7.7 |
| 1G6J   | 5387       | 2/58      | 18/22/36        | 75/47/71       | 1.4/3/1.3   | 23, 41     | 1.01/1.2/3 |
| 1PAW   | 5615       | 5/ – /4   | 66/ – /33       | 253/ – /29     | 3.8/ – /2.7 | 23, 41     | 1.2/ – /3.5 |
| 1SGO   | 6052       | 4/6/9     | 47/26/64        | 199/68/131     | 2.93/4/6.3  | 23, 41     | 2.8/1.39/6 |
| 1RYJ   | 5106       | 1/5/7     | 9/2/37          | 31/55/51       | 1.04/3/3.4  | 23, 41     | 1.3/1.42/6 |
| 2FRT   | 1675       | / – /3   | / – /5/10       | / – /36/108    | / – /1.02/9 | 23, 41     | / – /1.54/5 |
| 1YYC   | 6515       | 2/9/11    | 36/72/66        | 149/165/153    | 1.24/7/2.6  | 23, 41     | 2.0/1.76/2 |

Columns give number of secondary structure elements, number of residues, number of contact graph edges, average number of noisy NMR edges per contact edge, percentage of missing contact edges and average RMSD to the reference model among structures in the deposited ensemble. Each column is broken into statistics for α-helices, β-sheets and loop regions, separated by slashes. ‘–’ indicates no instance of that secondary structure.

For each dataset, we ran our algorithm 100 times, each for 10 000 iterations. For each run, we kept the top-scoring assignment over the 10 000 iterations. We then took as our solution ensemble the top 10 assignments over the 100 runs. For validation purposes, we used deposited solutions, which were determined by expert spectroscopists, as ‘reference’ assignments.

For all test cases, the randomized algorithm took from 20 min to a few hours for the assignment of a whole protein. The time required depends on the quality of the input NMR data and of the structure—noisier datasets and less-representative structures take longer, as the search space is not well constrained.

Figure 4 illustrates some examples of the convergence of the algorithm; other runs and other datasets had similar behavior. In general, the score increases rapidly over the initial iterations (a few hundred steps). During this phase, pseudoresidues are being organized into various ‘short’ paths aligned to the primary sequence, naturally increasing the score. With successive iterations, the short paths will start to grow into each other and conflicts occur, requiring fix-up moves to remove the conflicts. While moves are made so as to prefer increased score, locally bad moves are occasionally made in order to escape local optima. In many cases, the score converges to a value near that of the reference solution. As we will see below, the variation tends to produce only minor ambiguity in the resulting correspondence, and over the ensemble of solutions the correct assignments tend to be found.

Figure 5 illustrates the assignment results for the experimental datasets. Notice that we can assign the whole protein, and that for most of the positions, the reference assignments are included in the top-ranked solutions. Exceptions tend to be from areas with many missing edges (e.g. 1JBB 51–57) or residues close to a Proline (e.g. 5GCN 34–35), which necessarily induces a break. The results also show that the high-scoring solutions tend largely to agree. For 1JBB, there are on average 1.7 matches for each residue in α-helices, and 1.2 in β-sheets and loops. For 2JHB the ambiguity level is 1.3 for α-helices, 2.5 for β-sheets and 2.4 for loops, and for 5GCN we have 1.3, 2.6 and 3. These numbers can be compared to the expected number of matches a priori, which is simply the number of residues in the protein within the same ambiguous amino acid class, anywhere from 2 to 14.) In general, β-sheets and loops are more ambiguous than α-helices since their tertiary structures generate fewer edge constraints. For the nine synthetic datasets, the average ambiguity is as low as 1 for α-helices (1G6J), β-sheets (1KA5) and loops (1EGO); with a maximum of 2.8 (1SGO), 3.6 (1YYC), 9.1 (1SGO) and median of 1.7 (1KA5), 1.5 (1G6J), 2.1 (1G6J) for the three types, respectively. The most ambiguous case is 1SGO loops since it has both the highest noise ratio (6.3) and the largest RMSD (9.6 Å).

We compared these results to the Corresponding ones of (Bailey-Kellogg et al., 2005) (limited to α-helices), and found that our algorithm performs much better. Considering each position separately, we can evaluate how frequently the majority of the solution ensemble identifies the correct match. In our results, that is true for 90% of the positions, whereas it holds for <70% of the positions under the earlier method.

For both the experimental datasets and the synthetic ones, we studied the sensitivity of our algorithm to structural variation. For each dataset, an ensemble of NMR-determined structures had been deposited. We generated a contact graph for each different member of the ensemble, and studied how well the original data could be

Fig. 3. Reuse-based growing and aligning. Contact graph and NMR residues in the same column are matched. There are two amino acid types (empty squares and filled circles), which must match. (a) Growing from a matched fragment ending in u to an unmatched fragment with v in the middle leaves behind the prefix of the unmatched fragment in order to append and match the suffix following v. (b) Growing from u to v requires a realignment of the joined fragment.

The joined fragment displaces the suffix starting at w of another fragment.

The joined fragment displaces the suffix starting at w of another fragment.
assigned under the varying structures. The average RMSDs of the ensemble members (all to the reference model) are given as the far right column in Table 1, and are representative of the extent of structural uncertainty one might expect when assigning NMR data using an X-ray structure or high-quality homology model.

Figure 6 illustrates the effect of structural variation on the performance of our algorithm for each secondary structure type. For experimental data, we observe that for α-helices, there is no obvious change in the assignment accuracy when reference structures have a moderate difference (RMSD ≤ 2 Å). However, for β-sheets and loops, the assignment accuracy degrades when RMSD increases beyond about 1.25 Å for β-sheets and 3 Å for loops. Similar results can be observed in the synthetic dataset—α-helices are very tolerant to structural uncertainty, while β-sheets are best for RMSDs under around 1.5 Å, and loops are best up to around 3.5 Å.

Figure 7 summarizes the performance of our algorithm for each dataset under different structure models. These results suggest that, overall, we achieve good accuracy in assignment, above 80% for α-helices, 70% for β-sheets and 60% for loops. Since contacts are discrete, one might expect more effects from structural variation. However, recall that our method focuses on matching paths and uses non-sequential edges for scoring. While the score degrades with the loss of non-sequential edges, path connectivity is fairly well maintained regardless of the 3D coordinates.

6 CONCLUSION

NMR spectroscopy provides scientists with the ability to collect detailed information regarding protein dynamics and interactions in solution. However, in order to interpret the dynamics and interaction experiments, it is necessary to first obtain a resonance assignment so that the observed spectral peaks may be matched to atoms in the protein (e.g. to localize which atoms are affected by binding). In order to increase the throughput and decrease the expense of performing resonance assignment, this article develops a new approach, contact replacement. Contact replacement exploits information from an available 3D structure (from X-ray crystallography or homology modelling) to drive the assignment process, replacing the typical more extensive and expensive set of experiments with a minimalist set. Once contact replacement has been performed, the available assignments can be used to interpret dynamics or perturbation experiments. We note that those are separate experiments not included in the assignment process, and it is an interesting question (regardless of the assignment approach) to propagate uncertainty from assignment to uncertainty in dynamics or interactions.

Contact replacement poses interesting algorithmic problems in matching corrupted graphs, along with basic questions regarding the information content in connectivity and in vertex labels. In this article we presented the first efficient algorithm to solve this problem for entire proteins. We used a random-graph theoretic framework to derive a theoretical justification for why our approach works well in practice. Even with a large number of noisy edges (a constant number per vertex) and a high degree of vertex label ambiguity, the random structure of the noise and ambiguity allows a polynomial-time algorithm to uncover the correct solutions.

We showed that our approach works quite well in practice, tolerating significant noise (up to 500% noisy edges), missings
(up to 40%) and structural variability (up to 2 Å in α-helices and β-sheets and more in loops), while achieving very good assignment accuracy (60–80% overall). This combination is quite promising, and a significant advance in the state of the art. In particular, our robustness to structural uncertainty suggests that we may even be able to handle a ‘looser’ structural profile, such as the overall relationship among the core elements. This is a compelling challenge for further work.

It is interesting to consider the relationship between contact replacement and nuclear vector replacement (NVR) (Langmead and Donald, 2004; Langmead et al., 2004), both of which use an available structure to perform NMR resonance assignment, but based primarily on different data. (NVR does use some NOESY data, too, but only unambiguously assignable peaks.) At a high level, the residual dipolar coupling data used in NVR is global, giving orientations of bond vectors with respect to a coordinate frame, whereas the NOESY data used here is local, giving distances only between close protons. A natural avenue of work is to study the relative information content of these types of information in order to develop a unified framework incorporating both.

Compared to other graph-based structure matching problems (e.g. threading, structural alignment, structure motif finding, chemical
Contact replacement for resonance assignment

compound query, etc.), contact replacement has no sequential order for one of the graphs (the NMR one). However, the basic insights behind our algorithm (namely reusing partial solutions by making local fix-ups) may still be quite relevant in developing new algorithms for these applications. Alternatively, giving up sequential order in those applications may result in finding more distant relationships.

ACKNOWLEDGEMENTS

Funding: This work is supported in part by US NSF grant IIS-0444544 to CBK.

Conflict of Interest: none declared.

REFERENCES

Bailey-Kellogg C. et al. (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. J. Comp. Biol., 7, 537–556.

Bailey-Kellogg C. et al. (2005) A random graph approach to NMR sequential assignment. J. Comp. Biol., 12, 569–583.

Barthès C. et al. (1997) Garnet—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J. Comp. Chem., 18, 139–149.

Björklund A. et al. (2002) Approximating longest directed paths. Electron. Coll. Comput. Complex., 32, 1–13.

Bollobás B. (2001) Random Graphs. Cambridge University Press, Cambridge, UK.

Chen Y. et al. (1993) Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. J. Biomol. NMR, 14, 123–132.

Erdös P. and Rényi A. (1959) On a problem of graphs and number theory. Acta. Math. Acad. Sci. Hungar., 11, 113–125.

Feder T. and Motwani R. (2005) Finding large cycles in hamiltonian graphs. In Proceedings of the 36th ACM Symposium on the Theory of Computing (STOC), Chicago, IL, USA. pp. 407–416.

Feder T. and Motwani R. (2002) Approximating the longest cycle problem in sparse graphs. SIAM J. Comput., 31, 1996–2017.

Gibson H. N. (2004) Finding paths and cycles of superpolylogarithmic length. In Proceedings of the 36th ACM Symposium on the Theory of Computing (STOC), Chicago, IL, USA, pp. 407–416.

Garey M. and Johnson D. (1979) The NP-completeness of the Hamiltonian cycle problem in planar digraphs with degree bound two. Inform. Process. Lett., 8, 199–201.

Güntert P. (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with program Mapper. J. Biomol. NMR, 17, 129–137.

Hajduk P. et al. (1997) Drug design: discovering high-affinity ligands for proteins. Science, 278, 497–499.

Jung J. and Zweckstetter M. (2004) MARS—robust automatic backbone assignment of proteins. J. Biomol. NMR, 30, 11–32.

Kamisetti H. et al. (2006) An efficient randomized algorithm for contact-based NMR backbone resonance assignment. Bioinformatics, 22, 172–180.

Kay L. (1999) Protein dynamics from NMR. Nat. Struct. Biol., 5 (Suppl), 513–517.

Langmead C. and Donald B. (2004) An expectation maximization maximum likelihood algorithm for automated NMR resonance assignments. J. Biomol. NMR, 29, 111–138.

Langmead C. et al. (2004) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. J. Comp. Biol., 11, 277–298.

Lin G. et al. (2002) An efficient branch-and-bound algorithm for assignment of protein backbone NMR peaks. In Proceedings of the Computer Society Conference on Bioinformatics, Palo Alto, CA, USA, pp. 165–174.

Montelione G. et al. (2000) Protein NMR spectroscopy in structural genomics. Nat. Struct. Biol., 7 (Suppl), 982–985.

Mosley H. and Montelione G. (1999) Automated analysis of NMR assignments and structures for proteins. Curr. Opin. Struct. Biol., 9, 635–642.

Palmer J. A. et al. (1996) Nuclear magnetic resonance studies of biopolymer dynamics. J. Phys. Chem., 100, 13293–13310.

Pandurangan G. (2005) On a simple randomized algorithm for finding a 2-factor in sparse graphs. Inform. Process. Lett., 95, 321–327.

Pless J. (1979) The NP-completeness of the Hamiltonian cycle problem in planar digraphs with degree bound two. Inform. Process. Lett., 8, 199–201.

Ponnuswamy S. and Debarati M. (1999) RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. J. Biomol. NMR, 15, 15–26.

Prouty P. et al. (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—the program ST2NMR. J. Comp. Chem., 23, 335–340.

Rossman M. and Blow D. (1962) The detection of sub-units within the crystallographic asymmetric unit. Acta. Cryst., 15, 24–31.

Shuker S. et al. (1996) Discovering high-affinity ligands for proteins: SAR by NMR. Science, 274, 1513–1514.

Vitek D. et al. (2000) Model-based assignment and inference of protein backbone nuclear magnetic resonances. Stat. Appli. Genet. Mol. Biol., 3, Article 6, 1–33. http://www.bepress.com/sagmb/vol3/iss1/art6

Vitek D. et al. (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. Bioinformatics, 21, i230–i236.

Vitek D. et al. (2006) Inferential backbone assignment for sparse data. J. Biomol. NMR, 35, 187–208.

Xiong F. and Bailey-Kellogg C. (2007) A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. In Proceedings of IEEE Bioinformatics and Bioengineering, Boston, MA, USA, pp. 403–410.

Xu Y. et al. (2000) Protein structure determination using protein threading and sparse NMR data. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan, pp. 299–307.

Zimmermann D. et al. (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J. Mol. Biol., 269, 592–610.