The NOESY Jigsaw: Automated Protein Secondary Structure and Main-Chain Assignment from Sparse, Unassigned NMR Data

Chris Bailey-Kellogg[1]   Alik Widge[1]   John J. Kelley, III[1,2]

Marcelo J. Berardi[2]   John H. Bushweller[3]   Bruce Randall Donald[1,4]

[1]Dartmouth Computer Science Department, Hanover, NH 03755, USA

[2]Dartmouth Chemistry Department, Hanover, NH 03755, USA

[3]Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22906, USA

[4]Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. *email*: brd@cs.dartmouth.edu

Abstract

High-throughput, data-directed computational protocols for *Structural Genomics* (or *Proteomics*) are required in order to evaluate the protein products of genes for structure and function at rates comparable to current gene-sequencing technology. This paper presents the Jigsaw algorithm, a novel high-throughput, automated approach to protein structure characterization with nuclear magnetic resonance (NMR). Jigsaw applies graph algorithms and probabilistic reasoning techniques, enforcing first-principles consistency rules in order to overcome a 5-10% signal-to-noise ratio. It consists of two main components: (1) graph-based secondary structure pattern identification in *unassigned* heteronuclear NMR data, and (2) assignment of spectral peaks by probabilistic alignment of identified secondary structure elements against the primary sequence. Deferring assignment eliminates the bottleneck faced by traditional approaches, which begin by correlating peaks among dozens of experiments. Jigsaw utilizes only four experiments, none of which requires ^{13}C-labeled protein, thus dramatically reducing both the amount and expense of wet lab molecular biology and the total spectrometer time. Results for three test proteins demonstrate that Jigsaw correctly identifies 79-100% of α-helical and 46-65% of β-sheet NOE connectivities, and correctly aligns 33-100% of secondary structure elements. Jigsaw is very fast, running in minutes on a Pentium-class Linux workstation. This approach yields quick and reasonably accurate (as opposed to the traditional slow and extremely accurate) structure calculations. It could be useful for quick structural assays to speed data to the biologist early in an investigation, and could in principle be applied in an automation-like fashion to a large fraction of the proteome.

1   Introduction

Modern automated techniques are revolutionizing many aspects of biology, for example, supporting extremely fast gene sequencing and massively parallel gene expression testing (e.g. [5, 18, 20]). Protein structure determination, however, remains a long, hard, and expensive task. High-throughput structural genomics is required in order to apply modern techniques such as computer-aided drug design on a much larger scale. In particular, a key bottleneck in structure determination by nuclear magnetic resonance (NMR) is the *resonance assignment* problem — the mapping of spectral peaks to tuples of interacting atoms in a protein. For example, spectral peaks in a 3D nuclear Overhauser enhancement spectroscopy (NOESY) experiment establish distance restraints on a protein's structure by identifying pairs of protons interacting through space. Assignment is also directly useful in techniques such as structure-activity relation (SAR) by NMR [35, 16] and chemical shift mapping [6], which compare NMR spectra for an isolated protein and a protein-ligand or protein-protein complex.

JIGSAW is a novel algorithm for automated main-chain assignment and secondary structure determination. It has been successfully applied to experimental spectra for three different proteins: Human Glutaredoxin [38], Core Binding Factor-Beta [19], and Vaccinia Glutaredoxin-1 [22]. In order to enable high-throughput data collection, JIGSAW utilizes only four NMR experiments: heteronuclear single quantum coherence spectroscopy (HSQC), H^N-H^α-correlation spectroscopy (HNHA), 80 ms total correlation spectroscopy (TOCSY), and NOESY. This set of experiments requires only days of spectrometer time, rather than the months required for the traditional set of dozens of experiments. Furthermore, JIGSAW only requires a protein to be ^{15}N-labeled, a much cheaper and easier process than ^{13}C labeling. From a computational standpoint, JIGSAW adopts a minimalist approach, demonstrating the large amount of information available in a few key spectra.

Given the set of four spectra listed above, JIGSAW identifies spectral peaks belonging to secondary structure elements, and assigns them to the corresponding residues in the protein's primary sequence. In contrast to theoretical and statistical approaches for secondary structure (e.g. [9, 8]) and global fold (e.g. [39]), JIGSAW works in a data-driven manner. The continued necessity of experimental approaches is illustrated by the fact that one of our test proteins, CBF-β, has a unique fold, so that homology-based structure determination would not be applicable. In contrast to secondary structure predictors, JIGSAW provides not only an indication of secondary structure, but also tertiary β-sheet connectivity. Finally, as noted above, the spectral assignment produced by JIGSAW is itself an important product. One use of assigned NMR data in addition to structure determination is the analysis of protein structural dynamics from nuclear spin relaxation (e.g. [31, 30, 21]). Assignment is necessary to determine the residues implicated in the dynamics data. Another important use of NMR assignments, previously mentioned, is SAR by NMR, one of the most important recent breakthroughs in experimental methods for high-throughput drug activity screening. Even if a crystal structure is already known, these studies perform NMR experiments in order to analyze chemical shift changes and determine ligand binding modes. JIGSAW offers a high-throughput mechanism for the required assignment process.

In order to identify and assign spectral peaks belonging to secondary structure, JIGSAW relies on two key insights: *graph-based secondary structure pattern discovery*, and *assignment by alignment*. Atoms in regular secondary structure interact in prototypical patterns experimentally observable in a NOESY spectrum. Traditional NMR techniques determine residue sequentiality from a set of through-bond experiments, and then use NOE connectivities to test the secondary structure type of the residues. JIGSAW, on the other hand, starts by looking for these patterns, and uses their existence as evidence of residue sequentiality. JIGSAW applies a set of first-principles constraints on valid groups of NOE interactions to manage the large search space of possible secondary structure patterns. Subsequently, JIGSAW assigns spectral peaks by aligning identified residue sequences to the protein's primary sequence. To do this, JIGSAW uses side-chain peaks identified in a TOCSY spectrum to estimate probable amino acid types for the residue sequence. It finds such a sequence in the protein's primary sequence, and assigns the spectral data accordingly.

In its philosophy of starting with NOESY connectivities, JIGSAW is in the same spirit as the partially automated Main-Chain Directed (MCD) approach of Wand and co-workers (e.g. [37, 10, 29]). MCD was developed for homonuclear spectra, and was applied to experimental data for only one small protein, human Ubiquitin [37]. JIGSAW, on the other hand, is fully automated and has been successfully applied to experimental heteronuclear spectra for three different larger proteins (for example, CBF-β is nearly twice the size

of Ubiquitin). Jigsaw takes the steps necessary to deal with the significant amount of degeneracy in spectra for large proteins; it also provides a formal graph-theoretic framework for understanding and analyzing the algorithm. Finally, Jigsaw utilizes a novel TOCSY-based method for aligning residue sequences to the primary sequence.

The Jigsaw and MCD approaches differ greatly from other (automated and partially automated) assignment protocols used today in the NMR community. Most modern approaches rely on a large suite of ^{13}C-labeled triple resonance NMR spectra (e.g. HNCA, HNCACB, HN(CO)CACB, . . .), either to establish sequential connectivities by through-bond experiments (e.g. Autoassign [46] and Pasta [25]), or to match chemical shift patterns (e.g. [26] and [7]). As previously discussed, Jigsaw requires only four spectra, making it much more suitable for high-throughput studies. Many automated assignment packages boot-strap the assignment process. For example, Noah [27, 28] uses assignments from through-bond spectra to assign the NOESY. Garant [2] correlates observed peaks across multiple spectra with peaks predicted by a sophisticated model. Partially-computed structures can be used to refine peak predictions (e.g. [17], [28], [32]).

The ^{13}C labeling of a protein required by most automated assignment approaches is quite expensive, making these approaches unsuitable for large-scale structural studies. In return, these protocols yield a great deal of information (e.g. extensive side chain interactions). In contrast, Jigsaw is much cheaper and faster, but does not obtain as much information. Thus Jigsaw is especially suitable for quick structural assays to speed data to the biologist early in an investigation, and could in principle be applied in an automation-like fashion to a large fraction of the proteome. Furthermore, the Jigsaw approach could also both help and benefit from current work on large proteins and sparse NOE sets. For example, protocols developed for the analysis of large proteins use complete predeuteration to alleviate spectral crowding and to sharpen resolution in NOESY spectra [14, 40, 12]. These protocols yield only H^N-H^N interactions and perhaps sparse H^N-^1H interactions, yet have proven useful in structural studies even though they do not yield the extensive amount of information used by most ^{13}C-based approaches. Synergies between such protocols and Jigsaw work in both directions. On one hand, Jigsaw also uses H^N-H^N and sparse H^N-^1H interactions to perform its assignment, and thus the protocols could be combined for studies of larger proteins. On the other hand, Jigsaw could potentially compute complete three-dimensional structures even with its limited set of spectra by leveraging the techniques developed to determine global folds from sparse NOEs (e.g. [1, 43, 36]).

Solving the NMR jigsaw puzzle raises a number of interesting algorithmic pattern-matching and combinatorial issues. This paper presents an analysis of the problem, algorithms to solve it, and experimental results. Section 2 reviews the information content available in the NMR spectra used by Jigsaw. Section 3 presents the graph-based formalism and algorithm for finding secondary structure elements in NOESY spectra. Section 4 discusses the alignment process. Sections 3.3 and 4.1 provide results on experimental data from three different proteins.

2 NMR Data

NMR spectra capture interactions between atoms as peaks in $\mathbb{R}^2$ or $\mathbb{R}^3$, where the axes indicate resonance frequencies (*chemical shifts*) of atoms. In the ^{15}N spectra used by JIGSAW, peaks correspond to an ^{15}N atom, an H^N atom, and possibly another ^{1}H atom, of particular resonance frequencies. JIGSAW takes as input, in addition to a protein primary sequence, lists of peak maxima and intensities, correlated across spectra.[5]

2.1 NMR Spectra

Figure 1 illustrates the experiments utilized by the JIGSAW algorithm, and Figure 2 shows how the information content is encoded in data structures used by JIGSAW.

HSQC An HSQC spectrum [4, pp. 411-447] identifies unique pairs of through-bond correlated H^N and ^{15}N atoms. Every residue has a unique such H^N-^{15}N pair on the protein backbone; the coordinates for the pair are shared by all interactions within that residue and serve to reference interactions across all spectra.[6] Thus the HSQC serves to identify *nodes* (putative residues) for JIGSAW.

HNHA An HNHA spectrum [4, pp. 524-528] captures interacting intraresidue through-bond H^N-^{15}N-H^α; peak intensities estimate the *J coupling constant* ^{3}J$_{H^N H^\alpha}$ which is correlated with the ϕ bond angle of a residue. Since this angle is characteristically different for α-helices and β-sheets, JIGSAW uses it as an estimator of the secondary structure type.

TOCSY A TOCSY spectrum [13] includes through-bond interactions with ^{1}H atoms on a residue's side chain; the 80 ms TOCSY in particular reaches many atoms on a residue's side chain. Since the chemical shifts of ^{1}H atoms for different amino acid types are characteristically different, JIGSAW uses the shifts of a TOCSY as a *fingerprint* of the amino acid type.

NOESY The 3D ^{15}N-edited NOESY experiment [13] correlates an amide proton H^N and its ^{15}N with a second proton that interacts through space at a distance less than 6 Å, via the Nuclear Overhauser Effect (NOE). In the terminology of [42], a d_{NN} interaction represents an H^N-H^N pair, while a $d_{\alpha N}$ interaction represents an H^α-H^N pair (see Figure 1(b)); these can be distinguished by the characteristically different chemical shifts of H^α and H^N atoms. JIGSAW uses the NOE peaks to form *edges* between nodes for potentially interacting residues.

2.2 NMR Data Structures

Using the information content of NMR spectra discussed in the preceding subsection, JIGSAW builds two data structures: an interaction graph connecting residue nodes, and a set of fingerprints for each node.

The first data structure, the *NOESY interaction graph*, is an abstraction of a NOESY spectrum that indicates potential residue interactions that could explain the peaks in a spectrum. Each 3D interresidue NOE peak has the H^N and ^{15}N coordinates of one residue and the ^{1}H coordinate of the H^α or H^N proton of another residue. The HSQC indicates which is the first residue by its unique H^N and ^{15}N coordinates. The TOCSY and HNHA indicate residues whose H^α or H^N has the given ^{1}H coordinate. Unfortunately, projection onto the ^{1}H dimension yields a large amount of *spectral overlap* — many protons have the same chemical shift, within a tolerance. For example, there are 10-20 possible explanations for each peak in the NOESY spectrum of CBF-β (see Section 3.3), yielding a 5-10% signal-to-noise ratio. This spectral overlap is the major source of complexity in the JIGSAW approach. The NOESY interaction graph captures the complete set of possible explanations for the peaks; the JIGSAW search algorithm then determines the correct ones.

Definition 1 (NOESY Interaction Graph) *A NOESY interaction graph $G = (V, E)$ is a labeled, directed multigraph with vertices V corresponding to residues and edges $E \subset V \times V$ such that $e = (v_1, v_2) \in E$ iff there is a NOESY interaction between a proton of v_1 and a proton of v_2. Vertices and edges are labeled as follows:*

---

[5] Automated peak picking is an interesting and well-studied signal processing problem (e.g. AUTOPSY [23]).

[6] Some side chains, such as `Gln`, have their own H^N-^{15}N pairs as well. These can be removed in preprocessing, or detected and handled specially.

Secondary structure type label $s : V \rightarrow \{\alpha, \beta, \rho\} \times [0, 1]$ *indicates whether a residue is believed to be in an α-helix, a β-sheet, or other (random-coil) conformation, and the level of confidence in that belief.*

Interaction type $t : E \rightarrow \{d_{\mathrm{NN}}, d_{\alpha\mathrm{N}}\}$ *indicates a $d_{\alpha\mathrm{N}}$ or d_{NN} interaction.*

Match score $m : E \rightarrow \mathbb{R}^+$ *is the ${}^1\mathrm{H}$ frequency difference between the observed peak and the shift of the correlated H^α or H^{N}.*

Atom distance $d : E \rightarrow \mathbb{R}^+$, *computed from the NOE peak intensity, estimates the proximity of the correlated atoms.*

A high match score suggests that a given edge, rather than one of its competitors, is the correct one. In practice, the NOESY interaction graph only includes edges for which the match score is below some threshold (e.g. 0.05 ppm). Different atom distances are expected for atom pairs in different conformations; (e.g. a pair of H^{N} atoms in an α-helix is expected to be quite close).

This data structure provides a more abstract view of the NOESY information than typical atom-based representations [42, 37], and is more amenable to search and analysis.

The second data structure, the *TOCSY fingerprint*, collects all proton chemical shifts associated with a given node.

Definition 2 (Fingerprint) *A* fingerprint *is a set of ${}^1\mathrm{H}$ chemical shifts correlated with a given residue (* H^{N}-${}^{15}\mathrm{N}$ *pair).*[7]

Section 4 uses fingerprints as indications of probable amino acid type, in order to find where in the primary sequence to align a sequence of nodes belonging to a secondary structure element.

---

[7]The main-chain ${}^{15}\mathrm{N}$ chemical shift can also be included in the fingerprint.

3 Graph-Based Secondary Structure Pattern Discovery

In order to find the correct secondary structure of a protein from the highly ambiguous NOESY interaction graph, JIGSAW employs a multi-stage search algorithm that enforces a set of consistency rules in potential groups of edges. The following subsections detail these consistency rules and the JIGSAW graph search algorithm.

3.1 NOESY Interaction Graph Constraints

Figure 3 shows some prototypical NOE interactions in (a) an α-helix and (b) an anti-parallel β-sheet (after [42]).[8] Due to the way a helix is twisted, the H^N of one residue is close to the H^N residue of the next, and the H^α of one residue is close to the H^N of the residue one complete turn up the helix. Since a β-sheet is more stretched out, only the H^α-H^N sequential interactions are experimentally visible in the NOESY, but a rich pattern of cross-strand interactions are possible. Figure 4 represents these patterns in NOESY interaction graphs, and enumerates the *interaction graph constraints* imposed on these graphs by the geometry of helices and sheets.[9]

Definition 3 (Consistency with Interaction Graph Constraints) *A subgraph G' of a NOESY interaction graph G is* consistent *with the interaction graph constraints if there exists an ordering of the vertices $V(G')$ into sequences such that every edge $e \in E(G')$ satisfies one of the forms listed in Figure 4.*

While a NOESY interaction graph from experimental data contains many false edges (and some missing edges as well), the interaction graph constraints strongly limit how the correct edges fit together. As an example, consider the pattern in Figure 3(a). The large amount of noise in a NOESY interaction graph implies that a vertex will have many (around 10 — see Section 2) d_{NN} edges to vertices that could follow it sequentially in an α-helix. However, based on a simple joint probability model (and confirmed by the statistics of Table 5 discussed below), an incorrect d_{NN} edge is less likely also to have its symmetric counterpart. Similarly, the probability of stringing together an incorrect sequence of four vertices and connecting them with an additional $d_{\alpha N}$ edge from the first to the last is even less, and the probability that multiple such sequences adjoin each other is even less. Intuitively, while correct edges consistently reinforce each other, incorrect edges tend to be randomly distributed and thus mutually inconsistent. This insight is repeatedly utilized in the JIGSAW algorithm.

3.2 NOESY Interaction Graph Search

The goal of the JIGSAW NOESY graph search is to find a subgraph of a given interaction graph that encodes the secondary structure of the protein. Such a graph will have interactions indicative of the corresponding secondary structure elements, and thus will satisfy the interaction graph constraints.

Definition 4 (Secondary Structure Graph) *A* secondary structure graph G^* *is a subgraph of a NOESY interaction graph G that is consistent with the interaction graph constraints (Definition 3).*

Since a globally consistent graph consists of multiple locally consistent subgraphs, each of constant size, JIGSAW does not have to solve a large subgraph isomorphism problem to obtain the entire secondary structure.

Figure 5 illustrates the key steps of the JIGSAW graph search algorithm. Given an interaction graph, JIGSAW identifies small fragment subgraphs ("jigsaw pieces") satisfying the interaction graph constraints, merges them into α-helices and pairs of adjacent β-strands, and collects the sequences into entire secondary structure representations. In practice, there are many incorrect fragments among the correct ones, but as discussed at the end of the previous section and supported in the results section, mutual inconsistencies generally keep them from merging into larger graphs. A final step is to rank the best solved jigsaws. The following subsections detail these steps.

---

[8]Parallel β-sheets have similar interactions; we illustrate JIGSAW's approach by concentrating on anti-parallel β-sheets.

[9]Note that since $^{12}C^\alpha$ is not NMR-active, $d_{\alpha N}$ interactions are asymmetric.

3.2.1 Identify Fragments

The first step of Jigsaw is to find small, consistent subgraphs of an interaction graph. Jigsaw searches for *fragment* instances of a set of *fragment patterns* evident in canonical interaction graphs (Figure 4).

Definition 5 (Fragment Pattern) *A* fragment pattern *is a set of constraints on the connectivities, interaction types, match scores, and atom distances for a set of edges, along with the secondary structure type labels for the vertices.*

Definition 6 (Fragment) *A* fragment *is a subgraph of an interaction graph satisfying the constraints of a particular fragment pattern (Definition 5).*

Figure 6 illustrates the connectivities of some such fragment patterns. Each pattern instance groups a small set of edges (representing NOE peaks) that are mutually consistent. Fragment patterns also allow the possibility of missing edges in experimental data. The directions of the missing edges are, however, determined by those of the other edges. For example, in Figure 6(b), patterns 3 and 4 are similar to patterns 1 and 2, respectively; the direction of the missing vertical edge can be inferred from the correspondence.

Fragments are identified by a straightforward graph search: search from each node, forming paths of edges that remain consistent with the pattern. Table 1 provides pseudocode for this search. The algorithm assumes that the connectivities of a fragment pattern are ordered so that the first p_1 edges each connect to exactly one new node, and the remaining edges, up to p total, each connect only already-visited nodes. Since each pattern specifies a connected subgraph, such an ordering is guaranteed to exist. Arguments F and T to the algorithm specify the indices of the from- and to-nodes for an edge, respectively. For example, the search for pattern 1 in Figure 6(a) would start from the leftmost node, find all edges from that node forward to the second node, find all edges from that node forward to the third node (so that the second node found, the to-node of the first edge, serves as the from-node for possible second edges; i.e. $F_2 = 2$), and so forth.

The algorithm starts from each of n nodes and searches to a fixed depth of p for a pattern of p edges, examining only the edges from a specified node at each step. A bound d on the maximum degree of each node permits a bound on the complexity of the search: we perform n searches, each of size $O(d^p)$.

Claim 1 (Computational Complexity of Fragment Pattern Identification) *Given an interaction graph with n nodes and maximum degree d, instances of a fragment pattern involving p edges can be identified in time $O(nd^p)$.*

In practice (as demonstrated in Table 5 below), the interaction graph constraints greatly restrict the search, pruning most paths before they reach a depth of p.

We assume that the fragment patterns generate a *complete* set of fragments. That is, any secondary structure graph G^* for a given interaction graph G can be formed from a union of the fragments identified in G. Due to the large number of incorrect edges, there can also be many incorrect fragments. It remains for the subsequent processing stages (below) to eliminate them.

3.2.2 Merge Sequentially-Consistent Fragments

Given a set of fragment "jigsaw pieces" $\mathcal{F}$, Jigsaw starts solving the puzzle of secondary structure by finding sequences of fragments whose union defines either an α-helix or two neighboring strands of a β-sheet and is consistent with the interaction graph constraints. To reduce the computational cost, it is possible to identify a set of root fragments $\mathcal{F}' \subseteq \mathcal{F}$ that satisfy stronger constraints, and to root the sequences at these fragments.

Definition 7 (Rooted Fragment Sequence) *Given a set of fragments $\mathcal{F}$ for an interaction graph G and a set of chosen* root *fragments $\mathcal{F}' \subseteq \mathcal{F}$, a rooted fragment sequence F is a subgraph of G consistent with the interaction graph constraints for either a single α-helix or a pair of adjacent β-strands, and formed from the union of a set of n fragments $F = \{f_1, f_2, \ldots, f_n\} \subset \mathcal{F}$, where $f_1 \in \mathcal{F}'$.*

Fragment sequences are computed by a straightforward exhaustive search from the root fragments (Table 2 provides pseudocode). In the worst case there are an exponential number of sequences — if any fragment can connect to any other, then there are $|\mathcal{F}|!$ possible such sequences. However, as with fragment pattern identification, the interaction graph constraints limit the possible sequences, and as Table 5 illustrates, the number of sequences generated from an initial fragment is much less than this upper bound.

The completeness of fragment sequences follows immediately from the assumed completeness of fragments, if there is at least one root fragment per helix or strand pair. We state the claim here for completeness of exposition.

Claim 2 (Completeness of Fragment Sequences) *Any secondary structure graph G^* for a given interaction graph G is a union of the fragment sequences for the fragments $\mathcal{F}$ in G.*

3.2.3 Collect Consistent Sequences

To obtain an entire, consistent secondary structure graph for the protein, JIGSAW forms unions of consistent fragment sequences (see Table 3 for the specification). Imposing directionality — first identifying sequences and then joining them — greatly reduces the size and redundancy of the search space. While the merging step is worst-case exponential in the number of fragment sequences, again in practice the interaction graph constraints keep the search sub-exponential (see Table 5) and allow the algorithm to run in only minutes.

As with fragment sequences, the completeness result follows immediately from the definition and is stated here for purposes of formalization.

Claim 3 (Completeness of Secondary Structure Graphs) JIGSAW *finds all consistent secondary structure graphs G^* for a given interaction graph G.*

3.2.4 Identify Best Secondary Structure Graphs

The final step in the JIGSAW graph search is to identify the best secondary structure graphs from the set of collected possibilities. Intuitively, the algorithm should produce a large graph, reaching all the vertices expected to belong to the given secondary structure type. Smaller graphs probably were not expanded due to inconsistencies. Furthermore, as many of the expected edges as possible should belong to the graph (vertices should have high degree), and should have good match scores.

This intuition is formalized with a probabilistic measure of a graph's correctness. For simplicity, we assume a Gaussian *a priori* probability that an edge e indicates the correct interaction represented by a spectral peak, based on comparison of ^{1}H chemical shifts (recall that the match score $m(e)$ encodes the difference — see Definition 1); it remains interesting future work to incorporate actual spectral "line shapes" [23] into this analysis. Normalization over all edges generated for the peak yields the probability that that edge is a good explanation for its peak. This yields a higher probability when a peak closely matches, and when it doesn't have many good competitors:

$$P(\text{interaction}(e)) = G_\sigma(m(e)) \tag{1}$$

$$P(\text{good}(e)) = \frac{P(\text{interaction}(e))}{\displaystyle\sum_{e' \in C(e)} P(\text{interaction}(e'))} \tag{2}$$

where $G_\sigma(\cdot)$ denotes a Gaussian of width σ, and $C(\cdot)$ denotes the set of edges generated for the peak of a given edge.

The *correctness probability* for a secondary structure graph G^* depends the goodness of its edges:

$$P(\text{correct}(G^*)) = 1 - \prod_{e \in G^*} (1 - P(\text{good}(e))) \tag{3}$$

The correctness probability can be applied during fragment sequence enumeration (Section 3.2.2) and secondary structure graph construction (Section 3.2.3), in order to prune graphs with too little *support* (correctness probability too low for the graph size).

3.3 Experimental Results

Jigsaw was tested on experimental data for Human Glutaredoxin (huGrx) [38], Core Binding Factor-Beta (CBF-β) [19], and Vaccinia Glutaredoxin-1 (vacGrx) [22].[10] ^{15}N-edited HSQC, HNHA, 80 ms TOCSY, and NOESY spectra were collected on a 500MHz Varian spectrometer at Dartmouth and processed with the program Prosa [15]. Peaks were picked manually and in a semi-automated fashion with the program Xeasy [3]. Jigsaw was invoked with the appropriate primary sequences and ASCII peak lists, referenced across spectra.[11] In order to distinguish the dependence on HNHA from the dependence on NOESY, Jigsaw was run with two spectral suites: the first with simulated J-coupling constants set at the nominal values for the correct secondary structure type, and the second with J-coupling constants computed from the experimental HNHA data; all other spectra were the same in the two suites. Jigsaw used the patterns of Figure 4 with a set of generic constraints on match score and atom distance. Computation took about one to ten minutes, depending on the protein.

Figures 7, 8, and 9 depict the α-helices discovered by Jigsaw in CBF-β, huGrx, and vacGrx, respectively, with both suites of spectra. The results are similar for both suites, except that α-helices in suite 2 sometimes extend past or fail to reach the end of an α-helix or β-strand, due to misleading J constants. In vacGrx under suite 2, an additional potential rigid piece of secondary structure is uncovered, extending from residue 48 to residue 51.

Figures 10 and 11 show the β-sheets uncovered by Jigsaw in CBF-β and huGrx, respectively, using suite 2. The results for CBF-β with suite 1 are the same as in Figure 10, but with the correct edges to residue 100 rather than the incorrect edges to 101 and 71. The results for huGrx with suite 1 are identical; in both cases, connectivity in the lower two strands of huGrx is too sparse for Jigsaw. Figure 12 shows that the NOESY connectivities for β-sheets in vacGrx are too sparse for general-purpose Jigsaw patterns to detect. These test cases demonstrate that Jigsaw correctly uncovers a significant portion of the β structure, particularly in well-connected portions of the graph. Note that β-sheets are *tertiary structure*, indicating more than just the sequentiality of their strands.

The purpose of the graph search is to identify the small fraction of edges in the NOESY interaction graph that are actually involved in secondary structure. Ultimately, this means that the algorithm is identifying for each NOE peak which putative residues are interacting to cause that peak. Thus, appropriate metrics for Jigsaw's graph search performance are the numbers of correct and incorrect edges/peaks identified, based on the actual assignments known from the literature. Table 4 summarizes the results for all three proteins. It also includes the number of "extra" edges that aren't considered part of the secondary structure elements but are still sequentially correct. With spectral suite 2, Jigsaw is less accurate about the extent of a helix or strand; however, the actual extent is ambiguous, and extending to additional sequentially-connected residues can be beneficial by providing additional assignments. The β-sheet peaks for both huGrx and vacGrx are so sparse that Jigsaw identifies little to no β structure. In general, it is much harder to uncover β-sheets than α-helices, since β-strand sequentiality is specified by the noisier H^α region of the spectrum. We expect proteins with significant β-sheet content, such as CBF-β, to have enough connectivity to support the mutually confirming Jigsaw graph patterns.

Table 5 demonstrates that, due to the interaction graph constraints, the actual combinatorics of Jigsaw are much better than the worst-case exponential possibility. Notice that Jigsaw efficiently explores one to two thousand edges (Table 5, line 1) to find less than one hundred correct ones (Table 4, lines 1-2).

---

[10]While huGrx and vacGrx have similar structures, their experimental spectra have significant differences.

[11]For CBF-β, Jigsaw uses manually-computed J-constants, following the NMR protocol of [19].

4 Fingerprint-Based Sequence Alignment

Fingerprint-based sequence alignment finds sets of sequential residues in the protein sequence corresponding to the vertex sequences identified by the JIGSAW graph search algorithm. This process utilizes the TOCSY fingerprints introduced in Section 2.

The BioMagResBank (BMRB) has collected statistics from a large database of observed chemical shifts [34]. Figure 13 shows the mean chemical shifts for the protons of the 20 different amino acid types. The chemical shifts are affected by local chemical environment, which includes amino acid type and secondary structure. The chemical shift index (CSI) has successfully used this information to predict secondary structure type given chemical shift and amino acid type [41]. JIGSAW takes a different approach: it "inverts" the BMRB to predict amino acid type given chemical shift and secondary structure type.

The first step in alignment is to match each vertex's fingerprint with the canonical BMRB fingerprints. Due to extra and missing peaks, only a partial match might be possible.

Definition 8 (Partial Fingerprint Match) *A partial fingerprint match between vertex fingerprint S_v and BMRB amino acid fingerprint S_a ($a \in A = \{\texttt{Ala}, \texttt{Arg}, \ldots\}$), is a bijection $m : S_v{}' \to S_a{}'$ between subsets $S_v{}' \subseteq S_v$ and $S_a{}' \subseteq S_a$.*

Partial fingerprint matches are scored based on how well corresponding points match, together with penalties for extra and missing points. Assuming Gaussian noise around the expected chemical shift, with standard deviation σ_a for amino acid type a, the match score is defined as follows:

$$\text{partial}(S_v{}', S_a{}') = c_0|S_v - S_v{}'| + c_1|S_a - S_a{}'| + c_2 \prod_{p \in S_v{}'} G_{\sigma_a}(p - m(p)) \tag{4}$$

where c_0, c_1, c_2 are weighting factors.

The *match score* for a vertex and amino acid type is defined as the best partial fingerprint match score; normalization yields the probability that a vertex is of a given amino acid type.

$$\text{match}(S_v, S_a) = \max_{S_v{}' \subset S_v, S_a{}' \subset S_a} \text{partial}(S_v{}', S_a{}') \tag{5}$$

$$P(\text{type}(v, a)) = \frac{\text{match}(S_v, S_a)}{\sum_{b \in A} \text{match}(S_v, S_b)} \tag{6}$$

Then the probability that a sequence of vertices $V = (v_1, v_2, \ldots v_n)$ aligns at position r in the primary sequence L (where $r \leq |L| - |V|$) is the joint type probability over corresponding vertices and amino acid types. The best alignment for a sequence of vertices V relative to a primary sequence s is the position r maximizing the probability.

$$P(\text{align}(V, s, r)) = \prod_{i=1}^{n} P(\text{type}(v_i, s_{r+i-1})) \tag{7}$$

$$\text{alignment}(V, s) = \operatorname*{argmax}_{r \leq |L| - |V|} P(\text{align}(V, s, r)) \tag{8}$$

This alignment process aligns each secondary structure element separately. As it is, this approach provides a basic algorithm for spectral interpretation, explaining peaks in a TOCSY spectrum by identifying which side-chain protons of a particular amino acid type could have caused them. However, in order to achieve the additional goal of finding a complete secondary structure assignment, it is necessary to ensure that no alignments conflict. This problem is similar to that of protein threading [24], where a novel primary sequence must be matched up against secondary structure elements from a known global fold. However, in our case, the ordering of the secondary structure elements is unknown (and of course the scoring function is different). It remains future work to extend threading algorithms to handle this harder task.

4.1 Experimental Results

The purpose of the alignment process is to interpret a TOCSY spectrum by identifying a substring of amino acid types in the primary sequence such that the peaks expected for the side chain protons can explain the observed peaks. The performance of this spectral interpretation process was tested by separately aligning each secondary structure element. Tables 6, 7 and 8 detail the results of fingerprint-based alignment for the TOCSY shifts of known α-helices and β-strands in CBF-β, huGrx, and vacGrx, respectively. Table 9 summarizes the number of correct alignments for all three proteins. The simulated TOCSY is produced from the average chemical shifts of the side-chain protons entered in the BMRB for the given protein. Since the simulated fingerprints are from data correlated among many spectra, they are much more complete and indicative of the amino acid types than are the single experimental TOCSY spectra. While experimental TOCSY yields good alignment results, the simulated results demonstrate that as pulse sequences improve (see e.g. [44, 45]), the experimental results should get even better. In general, long sequences align better than short ones, although unusually noisy data can disrupt the alignment.

As discussed in the preceding section, a generalized threading approach will be necessary in order to determine a consistent alignment for all secondary structure elements of a protein. We tried a simple test in order to evaluate the potential for success of such an algorithm. This test found the top five alignments for each secondary structure element of huGrx; took the cross product to identify sets of alignments, one for each element; eliminated the members that included overlapping alignments; and scored the remaining alignment sets with the product of probabilities for the individual member alignments. The correct alignment set received the best score, by a factor of 100. This motivates the hope that, while individual alignments might not always score best, determining a complete alignment set will correct individual mistakes by eliminating sets with inconsistent members.

5   Conclusions and Future Work

This paper has described the JIGSAW algorithm for automated high-throughput protein structure determination. JIGSAW uses a novel graph formalization and new probabilistic methods to find and align secondary structure fragments in protein data from a few key fast and cheap NMR spectra. A set of first-principles graph consistency rules allow JIGSAW to manage the search space and prevent combinatorial explosion. JIGSAW has proven successful in structure discovery and alignment with experimental data for three different proteins.

JIGSAW offers a novel approach to the automated assignment of NMR data and the determination of protein secondary structure. Since JIGSAW uses only four spectra and ^{15}N-labeled protein, it is applicable in a much higher throughput fashion than traditional techniques, and could be useful for applications such as quick structural assays and SAR by NMR. It demonstrates the large amount of information available in a few key spectra. Finally, JIGSAW formalizes NMR spectral interpretation in terms of graph algorithms and probabilistic reasoning techniques, laying the groundwork for theoretical analysis of spectral information.

We are developing a random graph analysis of the complexity, correctness, and completeness of JIGSAW. This analysis uses a statistical model of the noise (extra and missing edges) in an interaction graph to compute the probability of false positives and false negatives for fragments, fragment sequences, and secondary structure graphs. This is an important direction for future work.

JIGSAW has only been run on the three proteins reported above. We plan to apply JIGSAW to experimental data for additional proteins, and to extend the techniques to analysis of DNA NMR data. We invite structural biologists desiring a fast structural assay to contact us if they wish to run JIGSAW. We anticipate that, since larger proteins have more NOEs, JIGSAW will have to handle more incorrect edges, and thus will require increased computational cost. An accurate noise model will provide a better indication of the dependence of computational complexity on spectrum size. We believe that as long as the noise edges are randomly distributed and do not achieve an overwhelming density, only correct graphs will be able to connect a large number of nodes with a dense set of consistent edges. There are also interesting possible connections between JIGSAW and approaches to computing structures of large proteins via deuteration and sparse NOEs; the introduction discusses the synergy in more detail.

An iterative deepening approach [33, pp. 70-71] could be incorporated into JIGSAW by noticing incompleteness of a secondary structure graph and restarting with looser constraints for fragment generation. For example, circular dichroism data provides an accurate estimate of the total amounts of α-helical and β-sheet structure in a protein [11]. By converting the secondary structure percentage to a count of the number of vertices involved, JIGSAW could recognize that a secondary structure graph was incomplete. Similarly, statistical secondary structure predictors (e.g. [9, 8]) predict the number of residues participating in separate secondary structure elements; JIGSAW could recognize and analyze the difference between its results and such a prediction.

The JIGSAW technique could be extended to assign H^N-^{1}H NOESY peaks on the side chains and to compute the global fold of a protein. Spectral referencing between TOCSY and NOESY gives an indication of which NOESY peaks belong to a given residue; additional interresidue interactions could then be identified in the NOESY and used to constrain the global geometry of α-helices and β-sheets. While such interactions will be sparse in purely ^{15}N-labeled protein, they might be sufficient to aid threading techniques that utilize secondary structure and sparse NOEs (e.g. [1, 43]) or structure determination algorithms from sparse NOE sets (e.g. [36]). Finally, JIGSAW could be re-targeted to include data from ^{13}C-labeled proteins, in order to attack larger proteins, while still requiring smaller sets of data than traditional approaches.

6   Acknowledgments

References

[1] AYERS, D. J., GOOLEY, P. R., WIDMER-COOPER, A., AND TORDA, A. E. Enhanced protein fold recognition using secondary structure information from NMR. *Protein Science 8* (1999), 1127–1133.

[2] BARTELS, C., GÜNTERT, P., BILETER, M., AND WÜTHRICH, K. GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry 18* (1997), 139–149.

[3] BARTELS, C., XIA, T.-H., BILLETER, M., GÜNTERT, P., AND WÜTHRICH, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *Journal of Biomolecular NMR 5* (1995), 1–10.

[4] CAVANAGH, J., FAIRBROTHER, W., PALMER III, A., AND SKELTON, N. *Protein NMR Spectroscopy: Principles and Practice.* Academic Press Inc., 1996.

[5] CHEN, T., FILKOV, V., AND SKIENA, S. Identifying gene regulatory networks from experimental data. In *Proc. RECOMB* (1999), pp. 94–103.

[6] CHEN, Y., REIZER, J., SAIER JR., M. H., FAIRBROTHER, W. J., AND WRIGHT, P. E. Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. *Biochemistry 32*, 1 (1993), 32–37.

[7] CROFT, D., KEMMINK, J., NEIDIG, K.-P., AND OSCHKINAT, H. Tools for the automated assignment high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. *Journal of Biomolecular NMR 10* (1997), 207–219.

[8] CUFF, J., CLAMP, M., SIDDIQUI, A., FINLAY, M., AND BARTON, G. JPRED: A consensus secondary structure prediction server. *Bioinformatics 14* (1998), 892–893.

[9] DEALEAGE, G., TINLAND, B., AND ROUX, B. A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins. *Analytical Biochemistry 163*, 2 (June 1987), 292–297.

[10] ENGLANDER, S., AND WAND, A. Main-chain directed strategy for the assignment of ^{1}H NMR spectra of proteins. *Biochemistry 26* (1987), 5953–5958.

[11] GALAT, A. A note on circular-dichroic-constrained prediction of protein secondary structure. *European Journal of Biochemistry 236* (1996), 428–435.

[12] GARDENER, K., ROSEN, M. K., AND KAY, L. E. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry 36* (1997), 1389–1401.

[13] GRONENBORN, A., BAX, A., WINGFIELD, P., AND CLORE, G. A powerful method of sequential proton resonance assignment in proteins using relayed ^{15}N-^{1}H multiple quantum coherence spectroscopy. *FEBS Letters 243* (1989), 93–98.

[14] GRZESIEK, S., WINGFIELD, P., STAHL, S., KAUFMAN, J. D., AND BAX, A. Four-dimensional ^{15}N-separated NOESY of slowly tumbling predeuterated ^{15}N-enriched proteins, application to HIV-1 Nef. *Journal of the American Chemical Society 117*, 37 (1995), 9594–9595.

[15] GÜNTERT, P., DÖTSCH, V., WIDER, G., AND WÜTHRICH, K. Processing of multi-dimensional NMR data with the new software PROSA. *Journal of Biomolecular NMR 2* (1992), 619–629.

[16] HAJDUK, P., MEADOWS, R., AND FESIK, S. Drug design: Discovering high-affinity ligands for proteins. *Science 278* (1997), 497–499.

[17] HARE, B., AND WAGNER, G. Application of automated NOE assignment to three-dimensional structure refinement of a 28 kD single-chain T cell receptor. *Journal of Biomolecular NMR 15* (1999), 103–113.

[18] HARTUV, E., SCHMITT, A., LANGE, J., MEIER-EWERT, S., LEHRACH, H., AND SHAMIR, R. An algorithm for clustering cDNAs for gene expression analysis. In *Proc. RECOMB* (1999), pp. 188–197.

[19] HUANG, X., SPECK, N., AND BUSHWELLER, J. Complete heteronuclear NMR resonance assignments and secondary structure of core binding factor β (1-141). *Journal of Biomolecular NMR 12* (1998), 459–460.

[20] KARP, R., STOUGHTON, R., AND YEUNG, K. Algorithms for choosing differential gene expression experiments. In *Proc. RECOMB* (1999), pp. 208–217.

[21] KAY, L. E. Protein dynamics from NMR. *Nature Structural Biology 5 Suppl* (1998), 513–517.

[22] KELLEY III, J., AND BUSHWELLER, J. ^{1}H, ^{13}C, and ^{15}N NMR resonance assignments of vaccinia glutaredoxin-1 in the fully reduced form. *Journal of Biomolecular NMR 12* (1998), 353–355.

[23] KORADI, R., BILLETER, M., ENGELI, M., GÜNTERT, P., AND WÜTHRICH, K. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance 135* (1998), 288–297.

[24] LATHROP, R. H., AND SMITH, T. F. Global optimum protein threading with gapped alignment and empirical pair score functions. *Journal of Molecular Biology 255* (1996), 651–665.

[25] LEUTNER, M., GSCHWIND, R., LIERMANN, J., SCHWARZ, C., GEMMECKER, G., AND KESSLER, H. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR 11* (1998), 31–43.

[26] LUKIN, J., GOVE, A., TALUKDAR, S., AND HO, C. Automated probabilistic method for assigning backbone resonances of (^{13}C, ^{15}N)-labeled proteins. *Journal of Biomolecular NMR 9* (1997), 151–166.

[27] MUMENTHALER, C., AND BRAUN, W. Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *Journal of Molecular Biology 254* (1995), 465–480.

[28] MUMENTHALER, C., GÜNTERT, P., BRAUN, W., AND WÜTHRICH, K. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *Journal of Biomolecular NMR 10* (1997), 351–362.

[29] NELSON, S., SCHNEIDER, D., AND WAND, A. Implementation of the main chain directed assignment strategy. *Biophysical Journal 59* (1991), 1113–1122.

[30] PALMER III, A. G. Probing molecular motion by NMR. *Current Opinion in Structural Biology 7* (1997), 732–737.

[31] PALMER III, A. G., WILLIAMS, J., AND MCDERMOTT, A. Nuclear magnetic resonance studies of biopolymer dyanmics. *Journal of Physical Chemistry 100* (1996), 13293–13310.

[32] PEARLMAN, D. Automated detection of problem restraints in NMR data sets using the FINGAR genetic algorithm method. *Journal of Biomolecular NMR 13* (1999), 325–335.

[33] RUSSELL, S., AND NORVIG, P. *Artificial intelligence: a modern approach.* Prentice-Hall, 1995.

[34] SEAVEY, B., FARR, E., WESTLER, W., AND MARKLEY, J. A relational database for sequence-specific protein NMR data. *Journal of Biolmolecular NMR* (1991), 217–236.

[35] SHUKER, S., HAJDUK, P., MEADOWS, R., AND FESIK, S. Discovering high-affinity ligands for proteins: SAR by NMR. *Science 274* (1996), 1531–1534.

[36] STANDLEY, D. M., EYRICH, V. A., FELTS, A. K., FRIESNER, R. A., AND MCDERMOTT, A. E. A branch and bound algorithm for protein structure refinement from sparse NMR data sets. *Journal of Molecular Biology 285* (1999), 1691–1710.

[37] STEFANO, D. D., AND WAND, A. Two-dimensional ^{1}H NMR study of human ubiquitin: a main-chain directed assignment and structure analysis. *Biochemistry 26* (1987), 7272–7281.

[38] SUN, C., HOLMGREN, A., AND BUSHWELLER, J. Complete ^{1}H, ^{13}C, and ^{15}N NMR resonance assignments and secondary structure of human glutaredoxin in the fully reduced form. *Protein Science 6* (1997), 383–390.

[39] Third meeting on the critical assessment of techniques for protein structure prediction. *Proteins: Structure, Function, and Genetics S3* (1999).

[40] VENTERS, R. A., METZLER, W. J., SPICER, L. D., MUELLER, L., AND FARMER, B. T. Use of H_N^1-H_N^1 NOEs to determine protein global folds in predeuterated proteins. *Journal of the American Chemical Society 117*, 37 (1995), 9592–9593.

[41] WISHART, D., SYKES, B., AND RICHARDS, F. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry 31*, 6 (February 1992), 1647–1651.

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

[43] XU, Y., XU, D., CRAWFORD, O. H., EINSTEIN, J. R., AND SERPERSU, E. Protein structure determination using protein threading and sparse NMR data. In *Proc. RECOMB* (2000), pp. 299–307.

[44] ZHU, G., KONG, X., AND SZE, K. Gradient and sensitivity enhancement of 2D TROSY-based experiments. *Journal of Biomolecular NMR 13* (1999), 3–10.

[45] ZHU, G., XIA, Y., SZE, K., AND YAN, X. 2D and 3D TROSY-enhanced NOESY of ^{15}N-labeled proteins. *Journal of Biomolecular NMR 14* (1999), 377–381.

[46] ZIMMERMAN, D., KULIKOWSI, C., HUANG, Y., FENG, W., TASHIRO, M., SHIMOTAKAHARA, S., CHIEN, C., POWERS, R., AND MONTELIONE, G. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology 269* (1997), 592–610.
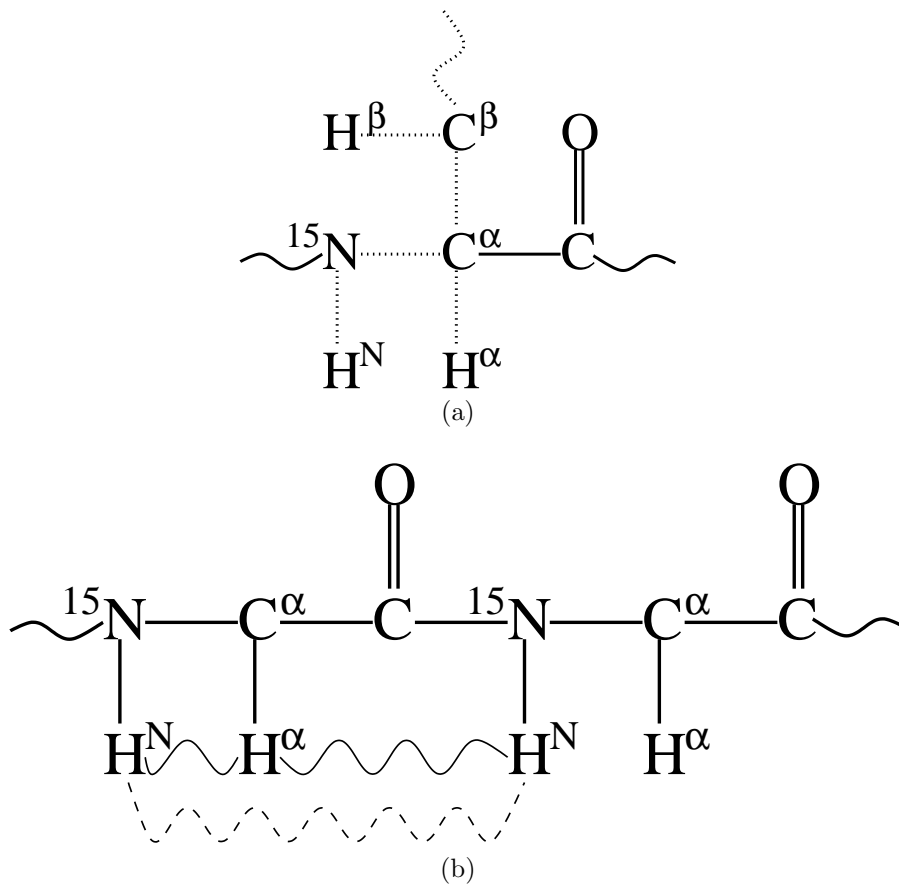
Figure 1: Atom nomenclature and interactions in a protein. (a) Through-bond interactions shown with dotted lines (HSQC: H^N-^{15}N; HNHA: H^N-^{15}N-H^α; TOCSY: H^N-^{15}N-H^α-H^β-...). (b) Through-space interactions in NOESY shown with wavy lines ($d_{\alpha N}$ solid and d_{NN} dashed).
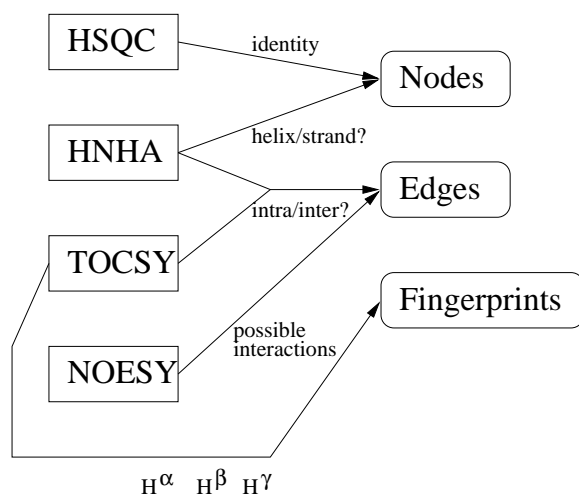
Figure 2: NMR spectra (rectangular boxes) and their uses in JIGSAW data types (oval boxes). The HSQC identifies a putative residue; the HNHA provides the ϕ angle, correlated with membership in α-helix or β-sheet; the TOCSY shows a fingerprint of side-chain proton shifts; the NOESY indicates possible interactions between nodes, with intra- vs. interresidue interactions distinguished by H^α shifts from the HNHA and TOCSY.
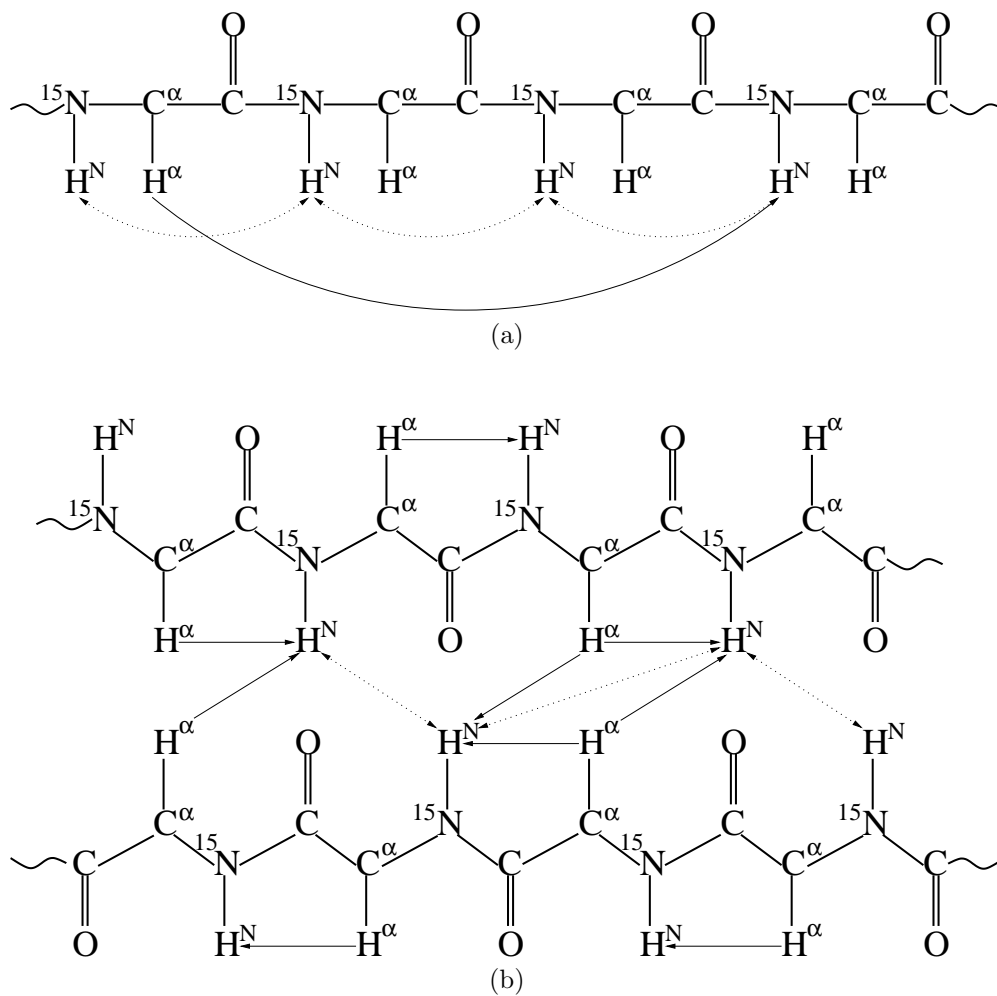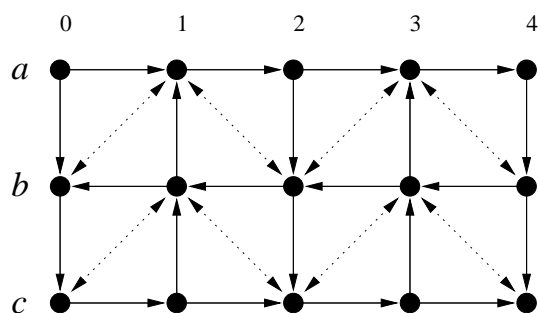
Figure 3: NOESY $d_{\alpha N}$ (solid) and d_{NN} (dotted) interactions in (a) α-helices and (b) β-sheets.

- d_{NN}: $h_i \leftrightarrow h_{i+1}$
- $d_{\alpha\mathrm{N}}$: $h_i \rightarrow h_{i+3}$
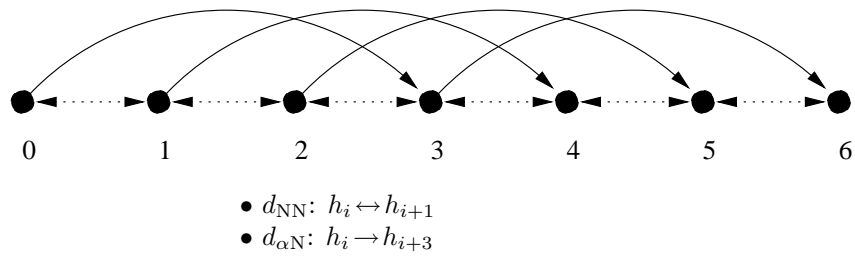
(a)

- $d_{\alpha\mathrm{N}}$: $a_i \rightarrow a_{i+1}$, $b_i \rightarrow b_{i-1}$, $c_i \rightarrow c_{i+1}$
- $d_{\alpha\mathrm{N}}$: $a_{2i} \rightarrow b_{2i}$, $b_{2i+1} \rightarrow a_{2i+1}$, $b_{2i} \rightarrow c_{2i}$, $c_{2i+1} \rightarrow b_{2i+1}$
- d_{NN}: $a_{2i+1} \leftrightarrow b_{2i}$, $a_{2i+1} \leftrightarrow b_{2i+2}$; $b_{2i+1} \leftrightarrow c_{2i}$, $b_{2i+1} \leftrightarrow c_{2i+2}$

(b)

Figure 4: Interaction graphs ($d_{\alpha\mathrm{N}}$ edges solid and d_{NN} dotted) and constraints for (a) α-helices and (b) β-sheets. This figure shows perfect patterns. Interaction graphs in experimental NMR data contain significant noise, manifested as some missing and many extra graph edges.
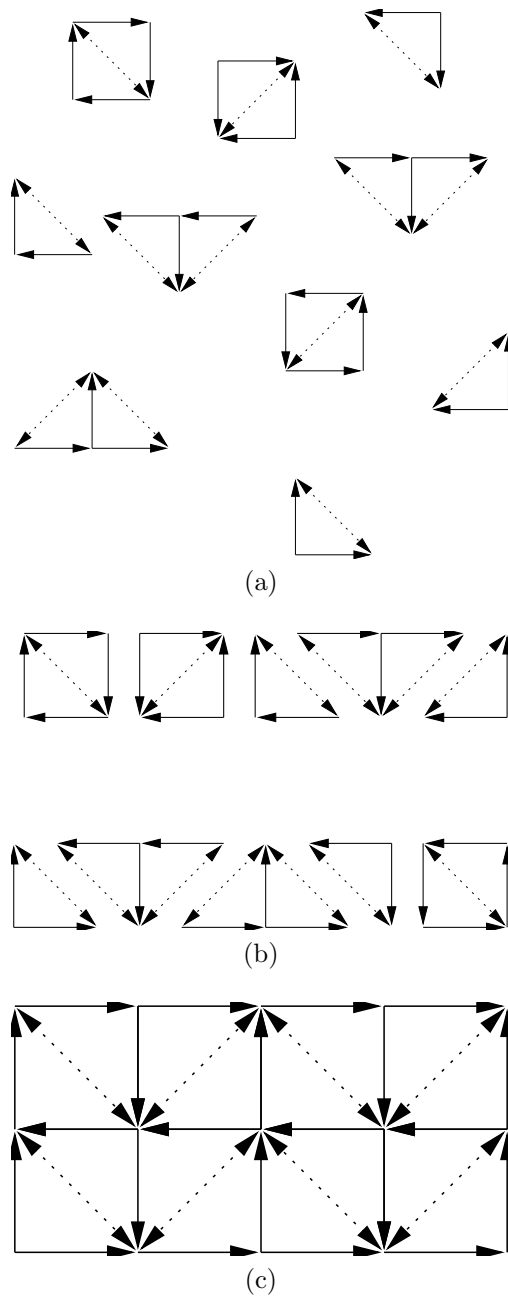
(a)

(b)

(c)

Figure 5: JIGSAW algorithm overview: (a) identify graph fragments, (b) merge them sequentially, and (c) collect them into complete secondary structure graphs. Only correct fragments are shown here. Graphs from experimental data also generate a large number of incorrect fragments, but mutual inconsistencies prevent them from forming either long sequences or large secondary structure graphs.
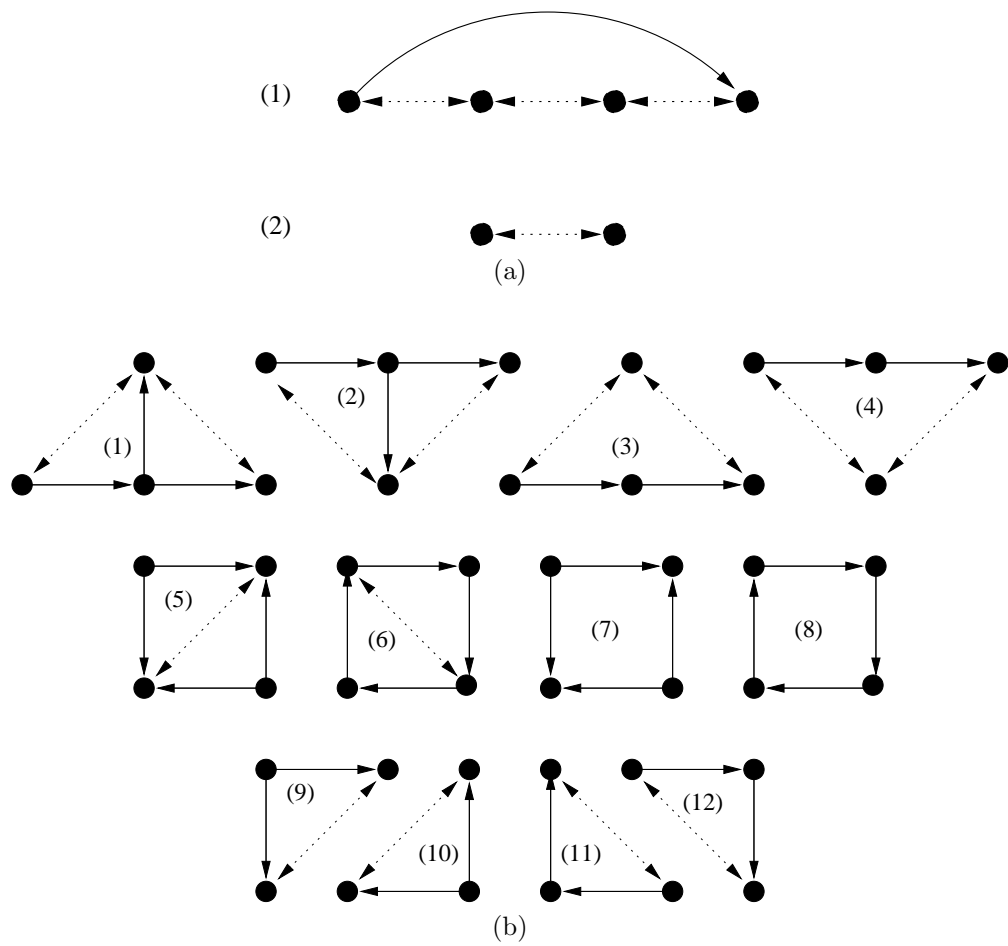
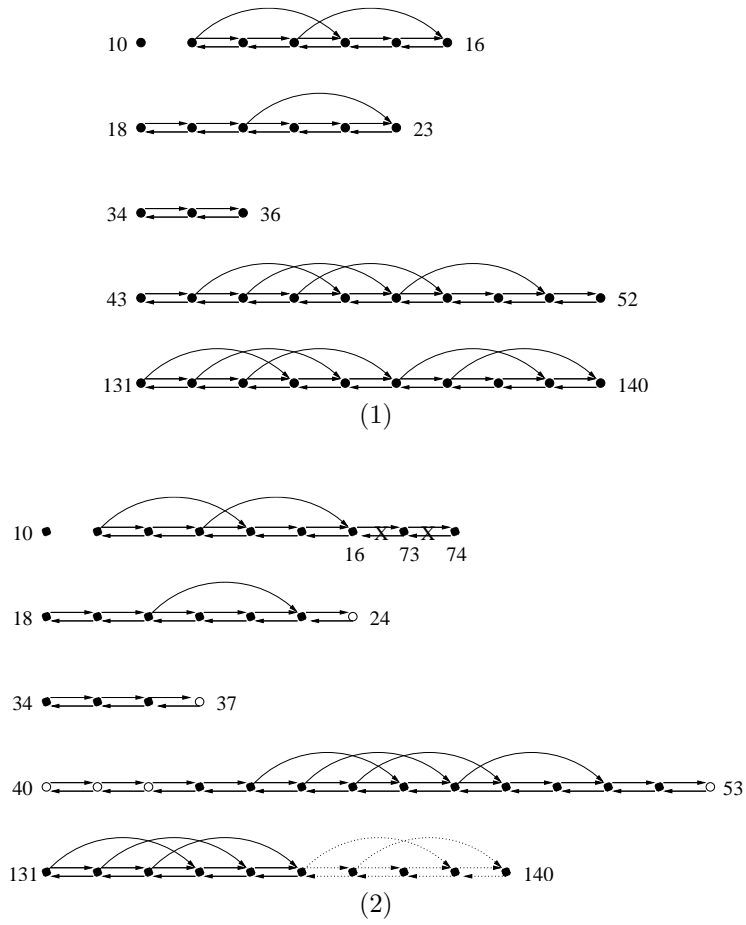Figure 6: Interaction graph fragment patterns in (a) α-helices and (b) β-sheets.

Figure 7: α-helices of CBF-β computed by JIGSAW, using spectral suites 1 and 2. Edges: solid=correct; dotted=false negative; X=false positive. Vertices: solid=correct; empty=sequentially correct but not in α-helix.
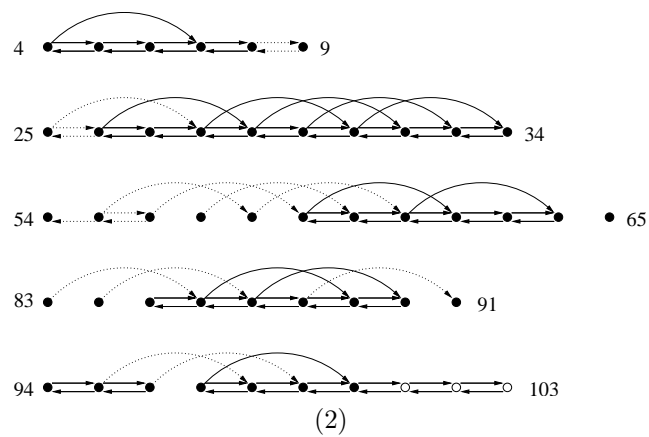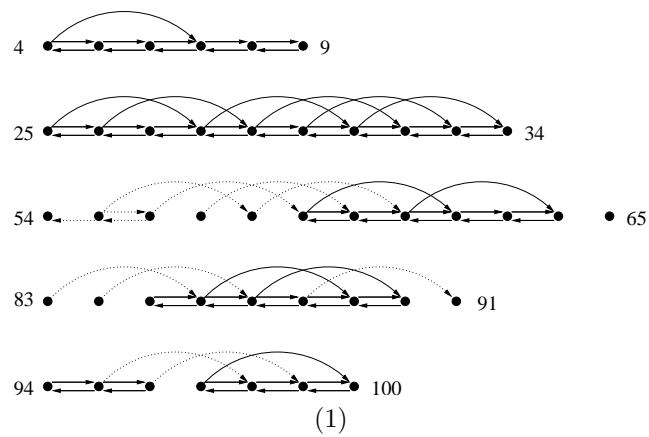
Figure 8: α-helices of huGrx computed by JIGSAW, using spectral suites 1 and 2. Edges: solid=correct; dotted=false negative. Vertices: solid=correct; empty=sequentially correct but not in α-helix.
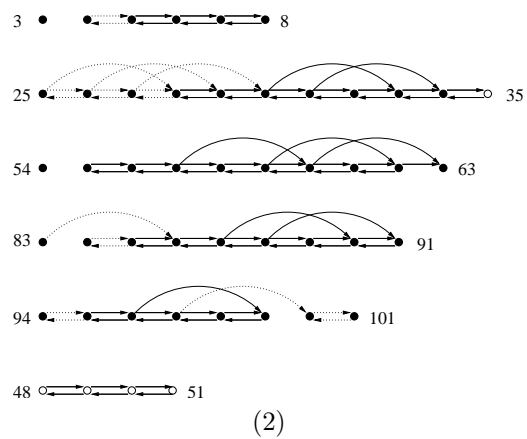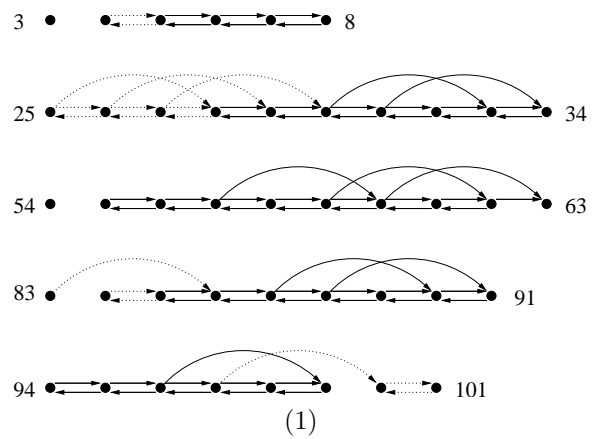
Figure 9: α-helices of vacGrx computed by Jigsaw, using spectral suites 1 and 2. Edges: solid=correct; dotted=false negative. Vertices: solid=correct; empty=sequentially correct but not in α-helix.

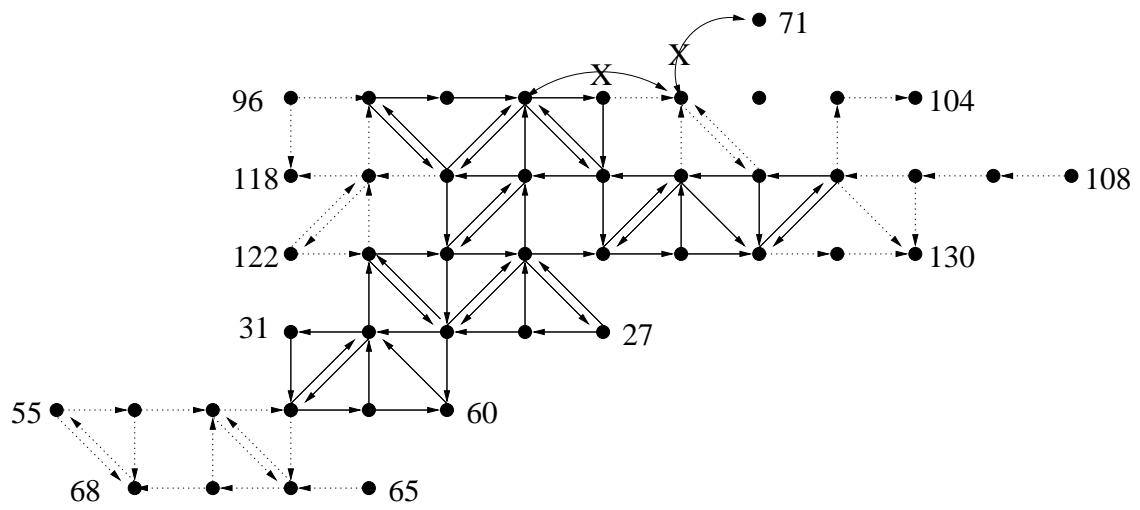Figure 10: β-sheets of CBF-β computed by Jigsaw. Edges: solid=correct; dotted=false negative; X=false positive.
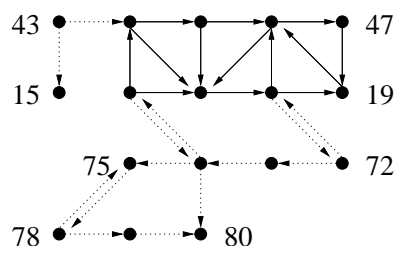
Figure 11: β-sheets of huGrx computed by JIGSAW, using spectral suite 2. Edges: solid=correct; dotted=false negative.
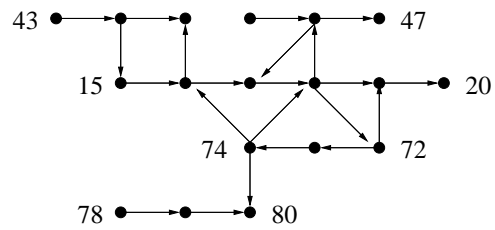
Figure 12: Known β-sheet connectivities in vacGrx. The connectivities are too sparse for the generic JIGSAW algorithm to uncover much structure.
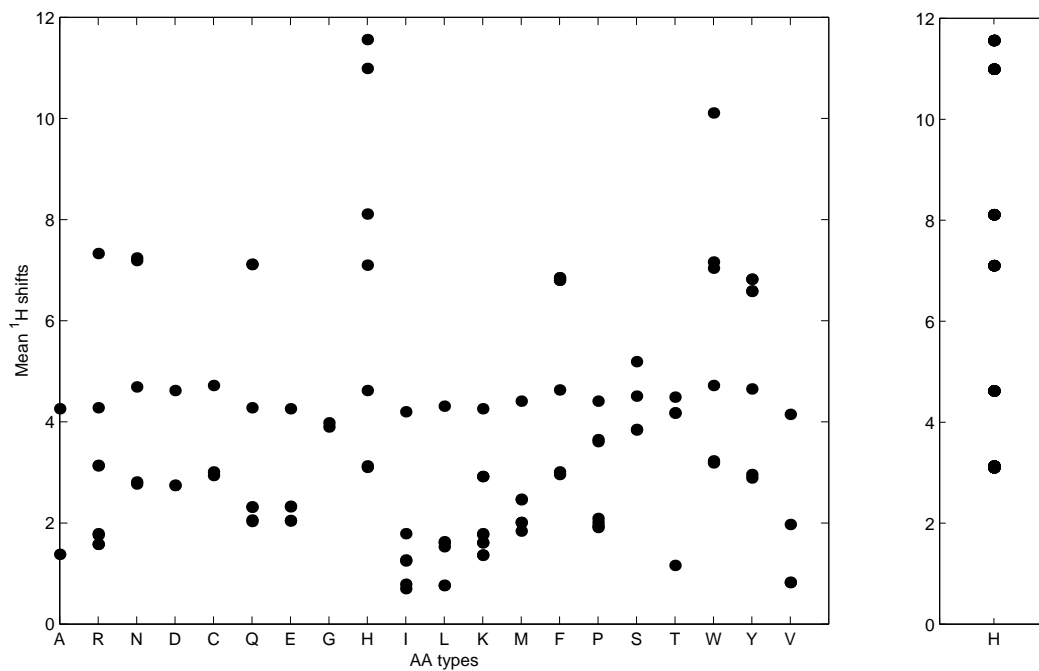
Figure 13: BMRB mean ^{1}H chemical shifts over different amino acid types. These shifts define "fingerprints" for the expected TOCSY peaks of different amino acid types; the fingerprint for His is isolated as an example.

```
Function fragments(G, p_1, p, F, T)
Set 𝒢 ← ∅
For each v ∈ V(G)
        Let 𝒢_1 = {({v, to(e)}, {e}) | from(e) = v}
        For i = 2..p_1
                Let 𝒢_i = {(V(G) ∪ {to(e)}, E(G) ∪ {e}) | (G ∈ 𝒢_{i-1}) ∧ (from(e) = V(G)_{F_i})}
        For i = p_1 + 1..p
                Let 𝒢_i = {(V(G), E(G) ∪ {e}) | (G ∈ 𝒢_{i-1}) ∧ (from(e) = V(G)_{F_i}) ∧ (to(e) = V(G)_{T_i})}
        Set 𝒢 ← 𝒢 ∪ 𝒢_p
return 𝒢
```

Table 1: Pseudocode for JIGSAW graph fragment identification. Add edges to the growing fragments such that each additional edge is from the specified already-found vertex (up to p_1) or between the specified already-found vertices (after p_1). Other constraints (e.g. type and match score) filter the sets but aren't shown here for simplicity.

```
Function sequences(ℱ, ℱ′)
Set 𝒮 ← ℱ′
While 𝒮 continues to grow
    Set 𝒮 ← {(V(S) ∪ V(F), E(S) ∪ E(F)) | (S ∈ 𝒮) ∧ (F ∈ ℱ) ∧ (S ∪ F connected) ∧ (S ∪ F consistent)}
return 𝒮
```

Table 2: Pseudocode for JIGSAW fragment sequence growth. Add fragments to the growing sequences such that each additional fragment is connected to the left or right end of a sequence (i.e. to a node with no forward or no backward adjacency) and the new sequence satisfies the interaction graph constraints.

Function **secondary_structures**($\mathcal{S}$)
return $\{G' = (\bigcup_{G \in S} V(G), \bigcup_{G \in S} E(G)) \,|\, (S \subseteq \mathcal{S}) \wedge (G' \text{ consistent})\}$

Table 3: Pseudocode for Jigsaw sequence collection. Find subsets of the fragment sequences such that the resulting graph satisfies the interaction graph constraints.

	huGrx	CBF-β	vacGrx
Actual	82	72	80
Correct	70; 65	72; 62	63; 63
% Correct	85%; 79%	100%; 86%	79%; 79%
Extra seq.	0; 0	0; 12	0; 8
Incorrect	0; 0	0; 4	0; 0

(a)

	huGrx	CBF-β
Actual	28	89
Correct	13; 13	58; 54
% Correct	46%; 46%	65%; 60%
Extra seq.	0; 0	0; 0
Incorrect	0; 0	0; 2

(b)

Table 4: Summary of results for JIGSAW secondary structure discovery ((a) α-helices and (b) β-sheets), for spectral suites 1 (first) and 2 (second).

	huGrx	CBF-β	vacGrx
Edges	1312	2216	807
Fragments	72	95	64
Root fragments	36	30	13
Fragment sequences	147	186	203
2ary structure graphs	647	17279	671

(a)

	huGrx	CBF-β
Edges	1312	2216
Fragments	277	1611
Root fragments	2	101
Fragment sequences	9	527
2ary structure graphs	9	6287

(b)

Table 5: Combinatorics of Jigsaw secondary structure discovery for (a) α-helices and (b) β-sheets.

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
α_1:10–16	1	$9 \cdot 10^4$	1	$3 \cdot 10^2$
α_2:18–23	1	$2 \cdot 10^4$	17	$4 \cdot 10^{-6}$
α_3:34–36	1	$4 \cdot 10^1$	3	$7 \cdot 10^{-2}$
α_4:43–52	1	$1 \cdot 10^{13}$	1	$2 \cdot 10^4$
α_5:131–140	1	$7 \cdot 10^{14}$	1	$1 \cdot 10^{19}$
$\beta_{1,1}$:27–31	1	$4 \cdot 10^3$	5	$3 \cdot 10^{-2}$
$\beta_{1,2}$:55–60	1	$2 \cdot 10^6$	1	$2 \cdot 10^4$
$\beta_{1,3}$:65–68	1	$2 \cdot 10^1$	1	$1 \cdot 10^3$
$\beta_{2,1}$:96–104	1	$2 \cdot 10^1$	1	$7 \cdot 10^2$
$\beta_{2,2}$:108–117	1	$4 \cdot 10^{10}$	11	$3 \cdot 10^{-5}$
$\beta_{2,3}$:122–130	1	$3 \cdot 10^4$	5	$1 \cdot 10^{-1}$

Table 6: Fingerprint-based alignment results for α-helices and β-strands of CBF-β, with both simulated and experimental TOCSY data. ρ indicates the relative score of the alignment — relative to either the best alignment, if the correct one is not best, or else to the second-best alignment.

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
α_1:4–9	1	$7 \cdot 10^7$	1	$1 \cdot 10^5$
α_2:25–34	1	$5 \cdot 10^{17}$	1	$8 \cdot 10^6$
α_3:54–65	1	$1 \cdot 10^{16}$	1	$9 \cdot 10^{13}$
α_4:83–91	1	$4 \cdot 10^5$	1	$2 \cdot 10^4$
α_5:94–100	1	$2 \cdot 10^7$	2	$2 \cdot 10^{-1}$
$\beta_{1,1}$:43–47	1	$1 \cdot 10^3$	3	$7 \cdot 10^{-3}$
$\beta_{1,2}$:15–19	1	$2 \cdot 10^3$	1	$3 \cdot 10^3$
$\beta_{1,3}$:72–75	1	$1 \cdot 10^3$	4	$2 \cdot 10^{-2}$
$\beta_{1,4}$:78–80	2	$2 \cdot 10^{-1}$	4	$4 \cdot 10^{-2}$

Table 7: Fingerprint-based alignment results for α-helices and β-strands of huGrx, with both simulated and experimental TOCSY data. ρ indicates the relative score of the alignment — relative to either the best alignment, if the correct one is not best, or else to the second-best alignment.

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
α_1:3–8	1	$2 \cdot 10^{10}$	5	$3 \cdot 10^{-2}$
α_2:25–34	1	$1 \cdot 10^{11}$	2	$3 \cdot 10^{-1}$
α_3:54–63	1	$1 \cdot 10^{32}$	1	$2 \cdot 10^3$
α_4:83–91	1	$7 \cdot 10^{13}$	4	$5 \cdot 10^{-3}$
α_5:94–101	1	$1 \cdot 10^5$	3	$2 \cdot 10^{-2}$
$\beta_{1,1}$:42–47	1	$4 \cdot 10^1$	1	$2 \cdot 10^1$
$\beta_{1,2}$:14–20	1	$3 \cdot 10^3$	15	$3 \cdot 10^{-8}$
$\beta_{1,3}$:72–74	1	$4 \cdot 10^2$	10	$5 \cdot 10^{-4}$
$\beta_{1,4}$:78–80	12	$2 \cdot 10^{-3}$	1	$1 \cdot 10^3$

Table 8: Fingerprint-based alignment results for α-helices and β-strands of vacGrx, with both simulated and experimental TOCSY data. ρ indicates the relative score of the alignment — relative to either the best alignment, if the correct one is not best, or else to the second-best alignment.

	huGrx	CBF-β	vacGrx
Correct (simulated TOCSY)	8/9	11/11	8/9
Correct (experimental TOCSY)	6/9	6/11	3/9

Table 9: Fingerprint-based alignment results summary for both simulated and experimental TOCSY data.