

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

Chris Bailey-Kellogg[†] Alik Widge[†] John J. Kelley, III^{†‡}
Marcelo J. Berardi[‡] John H. Bushweller[§] Bruce Randall Donald^{†¶}

October 4, 1999

Dartmouth Computer Science Technical Report No. PCS TR-99-358

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 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. JIGSAW's deferment of assignment until after secondary structure identification differs greatly from traditional approaches, which begin by correlating peaks among dozens of experiments. By deferring assignment, JIGSAW not only eliminates this bottleneck, it also allows the number of experiments to be reduced from dozens to four, none of which requires ¹³C-labeled protein. This in turn dramatically reduces the amount and expense of wet lab molecular biology for protein expression and purification, as well as the total spectrometer time to collect data.

Our results for three test proteins demonstrate that we are able to identify and align approximately 80 percent of α -helical and 60 percent of β -sheet structure. JIGSAW is extremely 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, utilizing a suite of graph analysis algorithms to compensate for the data sparseness. JIGSAW could be used for quick structural assays to speed data to the biologist early in the process of investigation, and could in principle be applied in an automation-like fashion to a large fraction of the proteome.

*This research is supported by the following grants to B.R.D. from the National Science Foundation: NSF IIS-9906790, NSF EIA-9901407, NSF 9802068, NSF CDA-9726389, NSF EIA-9818299, NSF CISE/CDA-9805548, NSF IRI-9896020, NSF IRI-9530785, and by an equipment grant from Microsoft Research.

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

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

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

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

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. [4, 14, 17]). 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 indicating pairs of protons interacting through space. Assignment is also directly useful in techniques such as structure-activity relation (SAR) by NMR [27, 12], which compares NMR spectra for an isolated protein and protein-ligand complex.

JIGSAW is a novel algorithm for automated secondary structure and main-chain assignment. It has been successfully applied to experimental spectra for three different proteins: Human Glutaredoxin [29], Core Binding Factor-Beta [15], and Vaccinia Glutaredoxin-1 [16]. 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), 80ms 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.

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. [28, 8, 23]). MCD was developed for homonuclear spectra, and was applied to experimental data for only one small protein, human Ubiquitin [28]. JIGSAW, in comparison, 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,

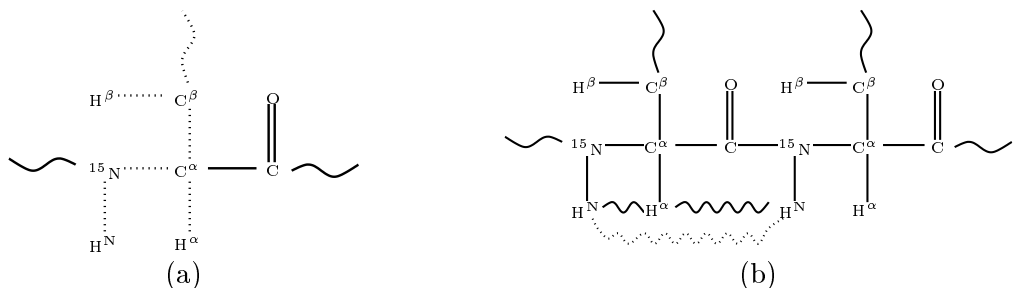


Figure 1: Atom nomenclature and interactions in a protein. (a) Through-bond interactions shown with dotted lines (HSQC: $^{15}\text{N}-\text{H}^{\text{N}}$; HNHA: $^{15}\text{N}-\text{H}^{\text{N}}-\text{H}^{\alpha}$; TOCSY: $^{15}\text{N}-\text{H}^{\text{N}}-\text{H}^{\alpha}-\text{H}^{\beta}-\dots$). (b) Through-space interactions in NOESY shown with wavy lines ($d_{\alpha\text{N}}$ solid and d_{NN} dashed).

...), either to establish sequential connectivities by through-bond experiments (e.g. AUTOASIGN [34] and PASTA [19]), or to match chemical shift patterns (e.g. [20] and [5]). ^{13}C -labeling of a protein is an expensive and time-consuming task, making these approaches unsuitable for high-throughput structural studies. As discussed above, JIGSAW uses only four experiments and requires only ^{15}N -labeling of a protein, a much cheaper process.

Many modern automated assignment packages boot-strap the assignment process. For example, NOAH [21, 22] uses assignments from through-bond spectra to assign the NOESY. GARANT [1] 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. [13], [22], [24]).

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 ^1H 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.¹ Figure 1 illustrates the experiments utilized by the JIGSAW algorithm:

- **HSQC:** An HSQC spectrum [3, pp. 411-447] identifies unique pairs of through-bond correlated ^{15}N and H^{N} atoms. Every residue has such a unique $^{15}\text{N}-\text{H}^{\text{N}}$ 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.²

- **HNHA:** An HNHA spectrum [3, pp. 524-528] captures interacting intraresidue $^{15}\text{N}-\text{H}^{\text{N}}-\text{H}^{\alpha}$; peak intensities estimate the *J coupling constant* $^3J_{\text{H}^{\text{N}}\text{H}^{\alpha}}$ 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

¹Automated peak picking is an interesting and well-studied signal processing problem (e.g. AUTOPSY [18]).

²Some side chains, such as Gln, have their own $^{15}\text{N}-\text{H}^{\text{N}}$ pairs as well. These can be removed in preprocessing, or detected and handled specially.

it as an estimator of the secondary structure type.

- **TOCSY:** A TOCSY spectrum [10] includes through-bond interactions with ^1H atoms on a residue’s side chain; the 80ms TOCSY in particular reaches many atoms on a residue’s side chain. Since the chemical shifts of ^1H 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 NOESY experiment [10] 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 [31], a d_{NN} represents an $\text{H}^{\text{N}}\text{-H}^{\text{N}}$ pair, while a $d_{\alpha\text{N}}$ represents an $\text{H}^{\alpha}\text{-H}^{\text{N}}$ pair (see Figure 1(b)); these can be distinguished by the characteristically different chemical shifts of H^{α} and H^{N} atoms.

The main JIGSAW 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 ^{15}N and H^{N} coordinates of one residue and the ^1H coordinate of the H^{α} or H^{N} proton of another residue. The HSQC indicates which is the first residue by its unique ^{15}N and H^{N} coordinates. The TOCSY and HNHA indicate residues whose H^{α} or H^{N} has the given ^1H coordinate. Unfortunately, projection onto the ^1H 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). 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) *The NOESY interaction graph $G = (V, E)$ is a labeled, directed multigraph defined as follows:*

- *Vertices V are residues.*
- *Edges $E \subset V \times V \times \{d_{\text{NN}}, d_{\alpha\text{N}}\} \times \mathbb{R}^+ \times \mathbb{R}^+$ with $e = (v_1, v_2, t, m, d) \in E$ iff there is a NOESY interaction between a proton of v_1 and a proton of v_2 :*
 - *Interaction type t indicates a $d_{\alpha\text{N}}$ or d_{NN} interaction.*
 - *Match score m is the ^1H frequency difference between the observed peak and the shift of the correlated H^{α} or H^{N} .*
 - *Atom distance d , 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 [31, 28], and is more amenable to search and analysis.

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.

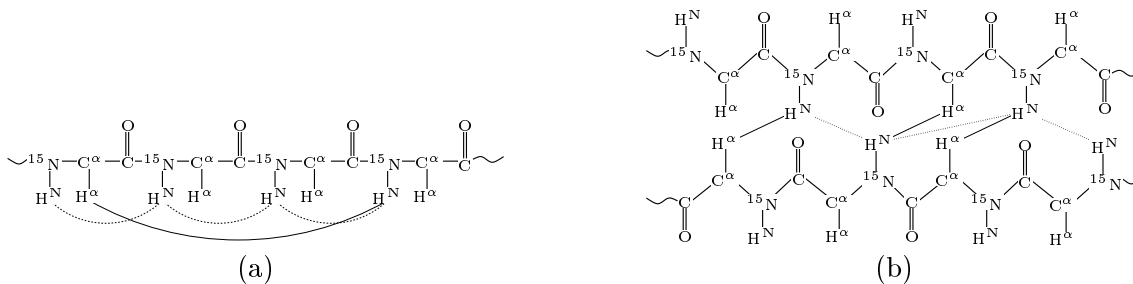


Figure 2: NOESY $d_{\alpha N}$ (solid) and d_{NN} (dotted) interactions in (a) α -helices and (b) β -sheets.

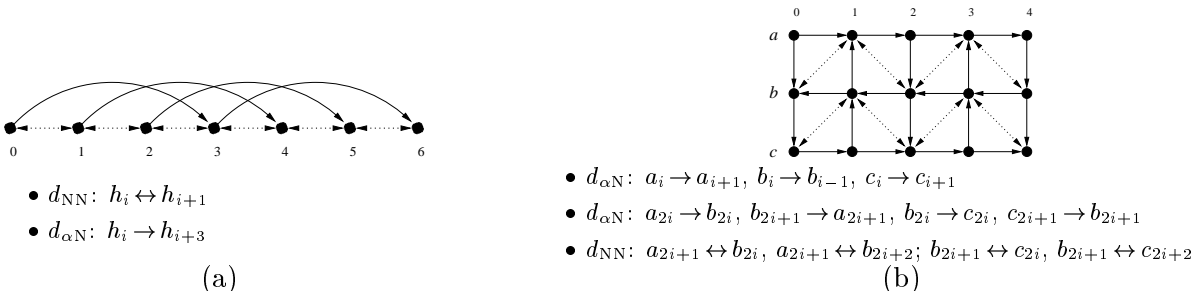


Figure 3: Interaction graphs ($d_{\alpha N}$ edges solid and d_{NN} dotted) and constraints for (a) α -helices and (b) β -sheets.

3.1 NOESY Interaction Graph Constraints

Figure 2 shows some prototypical NOE interactions in (a) an α -helix and (b) an anti-parallel β -sheet (after [31]).³ 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 3 represents these patterns in NOESY interaction graphs, and enumerates the *interaction graph constraints* imposed on these graphs by the geometry of helices and sheets.⁴

While a NOESY interaction graph contains many false edges (and in experimental data, some missing edges as well), the interaction graph constraints strongly limit how the correct edges fit together. For example, it is likely that a vertex will have several d_{NN} edges to vertices that could follow it sequentially in an α -helix. However (see Figure 3), it is less likely that an incorrect next vertex also has a symmetric d_{NN} edge, or that an incorrect sequence of vertices is also connected by an additional $h_i \rightarrow h_{i+3}$ $d_{\alpha N}$ edge, or that multiple such sequences adjoin each other. This insight of *mutually inconsistent incorrect hypotheses* is repeatedly utilized in the JIGSAW algorithm.

3.2 NOESY Interaction Graph Search

The JIGSAW NOESY graph search uncovers secondary structure in an interaction graph G as a subgraph G^* of G consistent with the interaction graph constraints. Since a globally consistent graph consists of repeating, locally consistent subgraphs, each of constant size, JIGSAW does not have to solve a large subgraph isomorphism problem for the entire secondary structure.

³Parallel β -sheets have similar interactions; this paper concentrates on anti-parallel β -sheets.

⁴Note that since the $^{12}C^\alpha$ is not NMR-active, $d_{\alpha N}$ interactions are asymmetric.

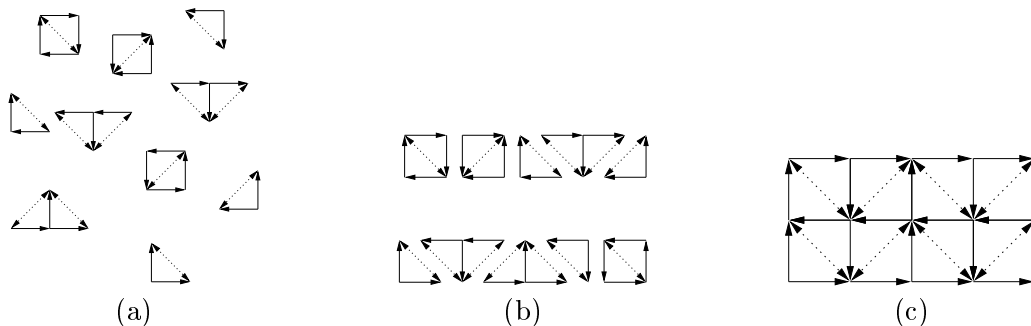


Figure 4: JIGSAW algorithm overview: (a) identify graph fragments, (b) merge them sequentially, and (c) collect them into complete secondary structure graphs.

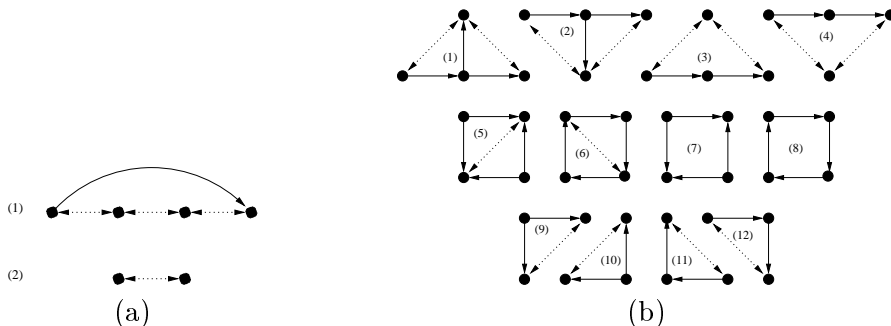


Figure 5: Interaction graph fragment patterns in (a) α -helices and (b) β -sheets.

Figure 4 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 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.

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 3). Figure 5 illustrates some such fragment patterns. These patterns constrain the interaction type, match score, and atom distance for a set of edges, along with the ϕ bond angle (and thus secondary structure type) indicated by the HNHA for the vertices.

Fragment patterns 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 5(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: for a pattern involving p edges, search from each node to depth p along paths that remain consistent with the pattern.

Claim 1 (Computational Complexity of Fragment Pattern Identification) *Given an interaction graph with n edges 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 2 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 consistent fragments that together define either an α -helix or two neighboring strands of a β -sheet. 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 2 (Rooted Fragment Sequence) *Given a set of fragments \mathcal{F} for an interaction graph G and a set of 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 $F = \{f_1, f_2, \dots, f_n\} \subset \mathcal{F}$, where $f_1 \in \mathcal{F}'$.*

Fragment sequences are computed by a straightforward exhaustive search from the root fragments. In the worst case there are an exponential number of sequences — if any fragment can connect to any other, then there are $|\mathcal{F}|^{|\mathcal{F}|}$ possible such sequences. However, as with fragment pattern identification, the interaction graph constraints strongly limit the possible sequences, and in practice (supported by Table 2) each initial fragment generates a fairly small number of sequences.

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.

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. 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, the interaction graph constraints again bring the search space down to a manageable size (see Table 2).

Since a secondary structure graph is computed as the union of fragment sequences, the completeness result follows immediately from Claim 2.

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 ^1H 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” [18] into this analysis. Normalization over all edges for a peak yields the probability that a particular edge is a good explanation for the peak. This yields a higher probability when a peak closely matches, and when it doesn’t have many good competitors.

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

$$P(\text{good}(e)) = P(\text{interaction}(e)) / \sum_{e' \in G} P(\text{interaction}(e')) \tag{2}$$

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) [29], Core Binding Factor-Beta (CBF- β) [15], and Vaccinia Glutaredoxin-1 (vacGrx) [16].⁵ ^{15}N -edited HSQC, HNHA, 80ms TOCSY, and NOESY spectra were collected on a 500MHz spectrometer and processed with the program PROSA [11]. Peaks were picked manually and in a semi-automated fashion with the program XEASY [2]. JIGSAW was invoked with the appropriate primary sequences and ASCII peak lists, referenced across spectra.⁶ 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 indicative of the known secondary structure, 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 3 with a set of generic constraints on match score and atom distance. Computation took about one to ten minutes, depending on the protein.

As an illustration, Figure 6 depicts the β -sheets JIGSAW uncovered for CBF- β , a 141-residue protein. (An optional appendix for the interested reader depicts the α -helix results for CBF- β and both α -helix and β -sheet results for the other two proteins.) 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.

Table 1 summarizes the results for all three proteins in terms of the number of correct, extra (but still sequential), and incorrect edges discovered by JIGSAW, compared to the actual edges known from the literature. Recall that edges correspond to NOESY peaks, and thus represent interpretations of portions of the spectrum. 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

⁵While huGrx and vacGrx have similar structures, their experimental spectra have significant differences.

⁶For CBF- β , JIGSAW uses manually-computed J-constants, following the NMR protocol of [15].

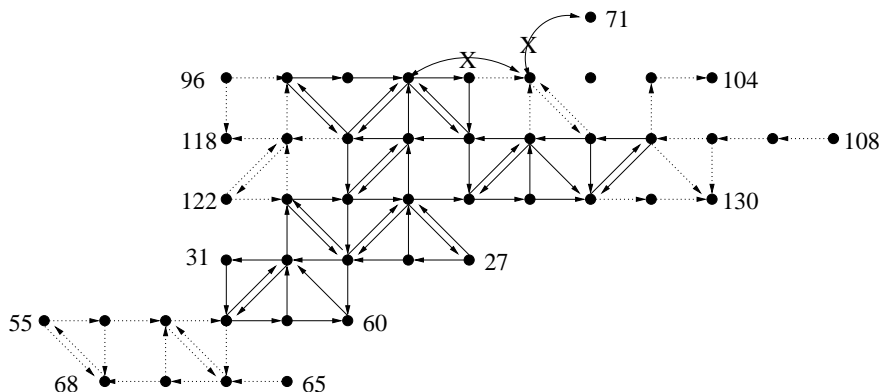


Figure 6: β -sheets of CBF- β computed by JIGSAW. Edges: solid=correct; dotted=false negative; X=false positive.

	huGrx	CBF- β	vacGrx		huGrx	CBF- β
Actual	82	72	80	Desired	28	89
Correct	70; 65	72; 62	63; 63	Correct	13; 13	58; 54
% Correct	85%; 79%	100%; 86%	79%; 79%	% Correct	46%; 46%	65%; 60%
Extra seq.	0; 0	0; 12	0; 8	Extra seq.	0; 0	0; 0
Incorrect	0; 0	0; 4	0; 0	Incorrect	0; 0	0; 2

(a)

(b)

Table 1: 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		huGrx	CBF- β
Edges	1312	2216	807	Edges	1312	2216
Fragments	72	95	64	Fragments	277	1611
Root fragments	36	30	13	Root fragments	2	101
Fragment sequences	147	186	203	Fragment sequences	9	527
2ary structure graphs	647	17279	671	2ary structure graphs	9	6287

(a)

(b)

Table 2: Combinatorics of JIGSAW secondary structure discovery for (a) α -helices and (b) β -sheets.

sequentially-connected residues can be beneficial by providing additional assignments. The β -sheet peaks for both huGrx and vacGrx are so sparse (see the appendix for illustrations) 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 much 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 2 demonstrates that, due to the interaction graph constraints, the actual combinatorics of JIGSAW are much better than the worst-case exponential possibility.

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

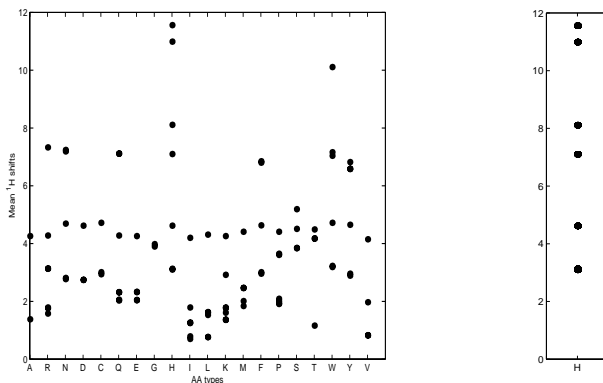


Figure 7: BMRB ^1H mean chemical shifts over different amino acid types. These shifts define “fingerprints” for the expected TOCSY peaks of different amino acid types; the fingerprint for His is isolated as an example.

utilizes the 80ms TOCSY (refer again to Section 2), which identifies a “fingerprint” of ^1H atoms.⁷

The BioMagResBank (BMRB) has collected statistics from a large database of observed chemical shifts [26]. Figure 7 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 [30]. 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 3 (Partial Fingerprint Match) A partial fingerprint match between vertex fingerprint S_v and BMRB amino acid fingerprint S_a ($a \in A = \{\text{Ala}, \text{Arg}, \dots\}$), is a bijection $m : S_v' \rightarrow S_a'$ 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:

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

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.

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

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

Then the probability that a sequence of vertices $V = (v_1, v_2, \dots, v_n)$ aligns at position r in the primary sequence L (where $r \leq |L| - |V|$) is the joint type probability over corresponding vertices

⁷The main-chain ^{15}N chemical shift can be included in the fingerprint.

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	$1 \cdot 10^{19}$

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
$\beta_{1,1}$: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 3: 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.

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

Table 4: Fingerprint-based alignment results summary for both simulated and experimental TOCSY data.

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.

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

$$\text{alignment}(V, s) = \max_{r \leq |L|-|V|} P(\text{align}(V, s, r)) \quad (8)$$

4.1 Experimental Results

Table 3 details the results of fingerprint-based alignment for the TOCSY shifts of known α -helices and β -strands in CBF- β (the optional appendix provides details for huGrx and vacGrx). Table 4 summarizes the number of correct alignments for all three proteins. The simulated TOCSY is produced from the known chemical shifts of the side-chain protons (correlated among many other spectra). While experimental TOCSY yields good alignment results, the simulated results demonstrate that as pulse sequences improve (see e.g. [32, 33]), the experimental results should get even better. In general, long sequences align better than short ones, although unusually noisy data can disrupt the alignment.

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.

One avenue of future work is a random graph analysis of JIGSAW using a statistical model of the noise in an interaction graph to compute the probable correctness and completeness of secondary structure graphs. Another avenue is to apply iterative deepening [25, pp. 70-71] to generate additional fragments, for example, due to suggestions by a statistical secondary structure

predictor (e.g. [7, 6]), circular dichroism data [9], or feedback from fingerprint-based alignment. Finally, the JIGSAW technique could be extended to assign 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.

6 Acknowledgments

We are very grateful to Xuemei Huang and Chaohong Sun for contributing their NMR data on huGrx and CBF- β to this project, for many helpful discussions and suggestions, and to Xuemei for running an invaluable new ^{15}N -TOCSY experiment for us. We would also like to thank Cliff Stein, Tomás Lozano-Pérez, Chris Langmead, Ryan Lilien, and all members of Donald Lab for their comments and suggestions.

References

- [1] C. Bartels, P. Güntert, M. Billeter, and K. Wüthrich. GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry*, 18:139–149, 1997.
- [2] C. Bartels, T.-H. Xia, M. Billeter, P. Güntert, and K. Wüthrich. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *Journal of Biomolecular NMR*, 5:1–10, 1995.
- [3] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, and N.J. Skelton. *Protein NMR Spectroscopy: Principles and Practice*. Academic Press Inc., 1996.
- [4] T. Chen, V. Filkov, and S. Skiena. Identifying gene regulatory networks from experimental data. In *Proc. RECOMB*, pages 94–103, 1999.
- [5] D. Croft, J. Kemmink, K.-P. Neidig, and H. Oschkinat. Tools for the automated assignment high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. *Journal of Biomolecular NMR*, 10:207–219, 1997.
- [6] J.A. Cuff, M.E. Clamp, A.S. Siddiqui, M. Finlay, and G.J. Barton. JPRED: A consensus secondary structure prediction server. *Bioinformatics*, 14:892–893, 1998.
- [7] G. Dealeage, B. Tinland, and B. Roux. A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins. *Analytical Biochemistry*, 163(2):292–297, June 1987.
- [8] S.W. Englander and A.J. Wand. Main-chain directed strategy for the assignment of ^1H NMR spectra of proteins. *Biochemistry*, 26:5953–5958, 1987.
- [9] A. Galat. A note on circular-dichroic-constrained prediction of protein secondary structure. *European Journal of Biochemistry*, 236:428–435, 1996.
- [10] A.M. Gronenborn, A. Bax, P.T. Wingfield, and G.M. Clore. A powerful method of sequential proton resonance assignment in proteins using relayed ^{15}N - ^1H multiple quantum coherence spectroscopy. *FEBS Letters*, 243:93–98, 1989.
- [11] P. Güntert, V. Dötsch, G. Wider, and K. Wüthrich. Processing of multi-dimensional NMR data with the new software PROSA. *Journal of Biomolecular NMR*, 2:619–629, 1992.
- [12] P.J. Hajduk, R.P. Meadows, and S.W. Fesik. Drug design: Discovering high-affinity ligands for proteins. *Science*, 278:497–499, 1997.
- [13] B.J. Hare and G. Wagner. Application of automated NOE assignment to three-dimensional structure refinement of a 28 kD single-chain T cell receptor. Submitted for publication, 1999.
- [14] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, and R. Shamir. An algorithm for clustering cDNAs for gene expression analysis. In *Proc. RECOMB*, pages 188–197, 1999.
- [15] X. Huang, N.A. Speck, and J.H. Bushweller. Complete heteronuclear NMR resonance assignments and secondary structure of core binding factor β (1-141). *Journal of Biomolecular NMR*, 12:459–460, 1998.

- [16] J.J. Kelley III and J.H. Bushweller. ^1H , ^{13}C , and ^{15}N NMR resonance assignments of vaccinia glutaredoxin-1 in the fully reduced form. *Journal of Biomolecular NMR*, 12:353–355, 1998.
- [17] R.M. Karp, R. Stoughton, and K.Y. Yeung. Algorithms for choosing differential gene expression experiments. In *Proc. RECOMB*, pages 208–217, 1999.
- [18] R. Koradi, M. Billeter, M. Engeli, P. Güntert, and K. Wüthrich. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance*, 135:288–297, 1998.
- [19] M. Leutner, R. Gschwind, Jens Liermann, C. Schwarz, G. Gemmecker, and H. Kessler. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR*, 11:31–43, 1998.
- [20] J.A. Lukin, A.P. Gove, S.N. Talukdar, and C. Ho. Automated probabilistic method for assigning backbone resonances of (^{13}C , ^{15}N)-labeled proteins. *Journal of Biomolecular NMR*, 9:151–166, 1997.
- [21] C. Mumenthaler and W. Braun. Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *Journal of Molecular Biology*, 254:465–480, 1995.
- [22] C. Mumenthaler, P. Güntert, W. Braun, and K. Wüthrich. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *Journal of Biomolecular NMR*, 10:351–362, 1997.
- [23] S. Nelson, D. Schneider, and A.J. Wand. Implementation of the main chain directed assignment strategy. *Biophys. J.*, 59:1113–1122, 1991.
- [24] D. Pearlman. Automated detection of problem restraints in NMR data sets using the FINGAR genetic algorithm method. *Journal of Biomolecular NMR*, 13:325–335, 1999.
- [25] S. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Prentice-Hall, 1995.
- [26] B.R. Seavey, E.A. Farr, W.M. Westler, and J.L. Markley. A relational database for sequence-specific protein NMR data. *Journal of Biomolecular NMR*, pages 217–236, 1991.
- [27] S.B. Shuker, P.J. Hajduk, R.P. Meadows, and S.W. Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274:1531–1534, 1996.
- [28] D.L. Di Stefano and A.J. Wand. Two-dimensional ^1H NMR study of human ubiquitin: a main-chain directed assignment and structure analysis. *Biochemistry*, 26:7272–7281, 1987.
- [29] C. Sun, A. Holmgren, and J. Bushweller. Complete ^1H , ^{13}C , and ^{15}N NMR resonance assignments and secondary structure of human glutaredoxin in the fully reduced form. *Protein Science*, 6:383–390, 1997.
- [30] D.S. Wishart, B.D. Sykes, and F.M. Richards. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651, February 1992.
- [31] K. Wüthrich. *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, 1986.

- [32] G. Zhu, X.M. Kong, and K.H. Sze. Gradient and sensitivity enhancement of 2D TROSY-based experiments. *Journal of Biomolecular NMR*, 13:3–10, 1999.
- [33] G. Zhu, Y. Xia, K.H. Sze, and X. Yan. 2D and 3D TROSY-enhanced NOESY of ^{15}N -labeled proteins. *Journal of Biomolecular NMR*, 14:377–381, 1999.
- [34] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269:592–610, 1997.

Appendix

A Additional Details for Experimental Results

A.1 Secondary Structure Graphs

This section details the α -helices and β -sheets uncovered by JIGSAW as summarized in Table 1. Figures 8, 9, and 10 depict the α -helices uncovered by JIGSAW in CBF- β , huGrx, and vacGrx, respectively, with both suites of spectra. The results are quite 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.

Figure 11 shows the β -sheets uncovered by JIGSAW in huGrx with suite 2. The results with suite 1 are identical; in both cases, connectivity in the lower two strands is too sparse for JIGSAW. The β -sheet results for CBF- β with suite 1 are the same as in Figure 6, but with the correct edges to residue 100 rather than the incorrect edges to 101 and 71. Figure 12 shows that the NOESY connectivities for β -sheets in vacGrx are too sparse for the general-purpose set of JIGSAW patterns to detect.

A.2 Fingerprint-Based Alignment

This section details the fingerprint-based alignment results of huGrx and vacGrx that contributed to Table 4. Tables 5 and 6 list the fingerprint-based alignment results for huGrx and vacGrx, respectively. As with CBF- β , simulated TOCSY data yields almost perfect results, while experimental TOCSY data results are somewhat degraded.

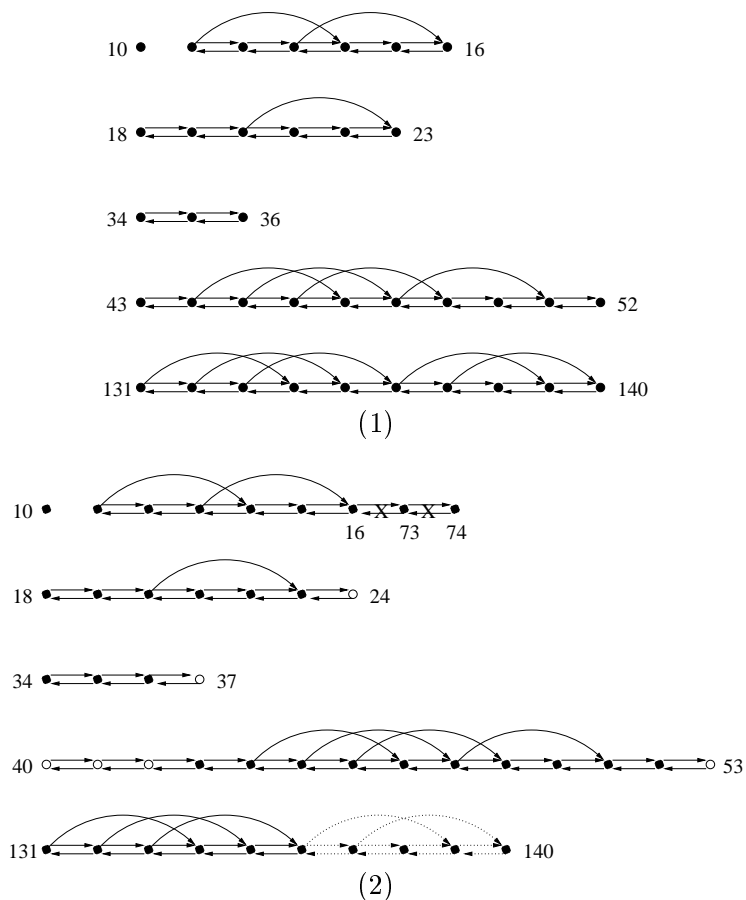


Figure 8: α -helices of CBF- β computed by JIGSAW, using spectral suites 1 and 2. Edges: solid=correct; dotted=false negative; X=false positive. Vertices: solid=correct; empty=sequentially correct but not in α -helix.

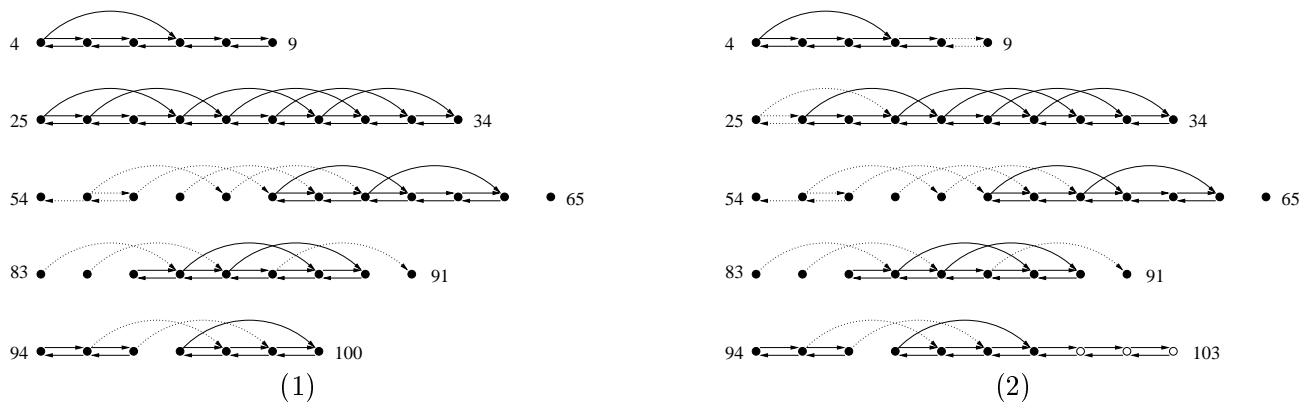


Figure 9: α -helices of huGrx computed by JIGSAW, using spectral suites 1 and 2. Edges: solid=correct; dotted=false negative. Vertices: solid=correct; empty=sequentially correct but not in α -helix.

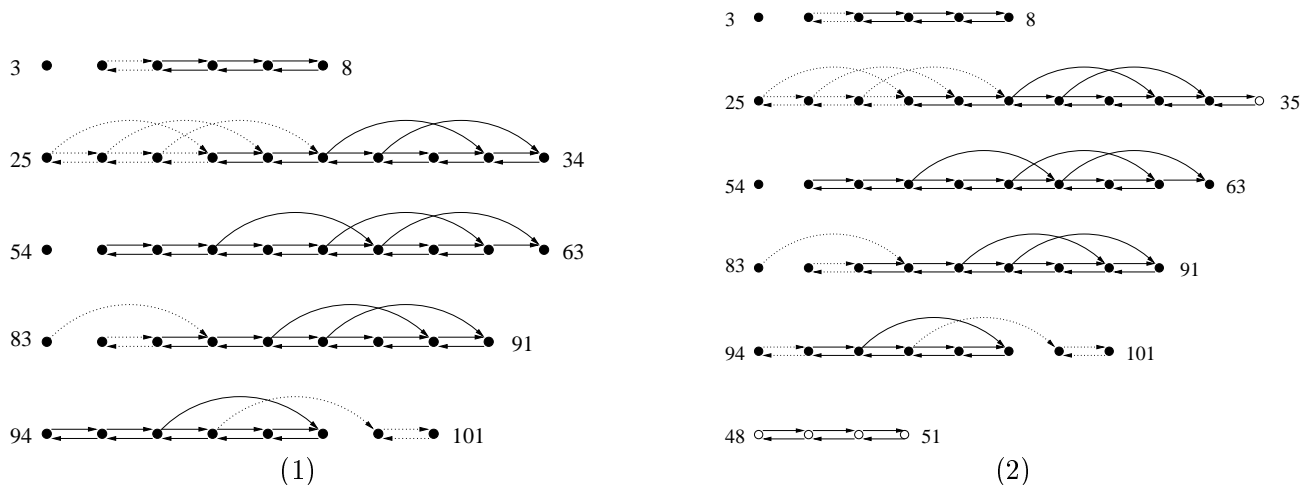


Figure 10: α -helices of vacGrx computed by JIGSAW, using spectral suites 1 and 2. Edges: solid=correct; dotted=false negative. Vertices: solid=correct; empty=sequentially correct but not in α -helix.

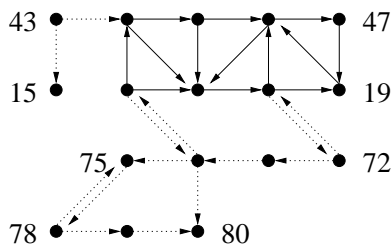


Figure 11: β -sheets of huGrx computed by JIGSAW, using spectral suite 2. Edges: solid=correct; dotted=false negative.

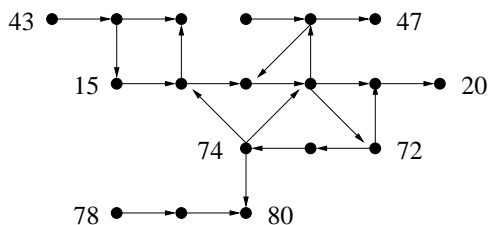


Figure 12: Known β -sheet connectivities in vacGrx. The connectivities are too sparse for the generic JIGSAW algorithm to uncover much structure.

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
$\alpha_1:4-9$	1	$7 \cdot 10^7$	1	$1 \cdot 10^9$
$\alpha_2:25-34$	1	$5 \cdot 10^{17}$	1	$8 \cdot 10^6$
$\alpha_3:54-65$	1	$1 \cdot 10^{16}$	1	$9 \cdot 10^{13}$
$\alpha_4:83-91$	1	$4 \cdot 10^5$	1	$2 \cdot 10^4$
$\alpha_5:94-100$	1	$2 \cdot 10^7$	2	$2 \cdot 10^{-1}$

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
$\beta_{1,1}:43-47$	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 5: 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
$\alpha_1:3-8$	1	$2 \cdot 10^{10}$	5	$3 \cdot 10^{-2}$
$\alpha_2:25-34$	1	$1 \cdot 10^{11}$	2	$3 \cdot 10^{-1}$
$\alpha_3:54-63$	1	$1 \cdot 10^{32}$	1	$2 \cdot 10^3$
$\alpha_4:83-91$	1	$7 \cdot 10^{13}$	4	$5 \cdot 10^{-3}$
$\alpha_5:94-101$	1	$1 \cdot 10^5$	3	$2 \cdot 10^{-2}$

Sequence	Simulated		Experimental	
	Rank	ρ	Rank	ρ
$\beta_{1,1}:42-47$	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 6: 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.