B Background and Significance

Modern automated techniques are revolutionizing many aspects of biology, for example, supporting extremely fast gene sequencing and massively parallel gene expression testing (e.g. [28, 64, 73]). 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 [109, 61] and chemical shift mapping [29], which compare NMR spectra for an isolated protein and a protein-ligand or protein-protein complex.

This proposal addresses automated assignment and high-throughput protein structure determination from sparse, unassigned NMR data. JIGSAW is a novel algorithm for automated main-chain assignment and secondary structure determination, developed by the PI Donald and co-workers [6, 7]. It has been successfully applied to experimental spectra for three different proteins: Human Glutaredoxin [112], Core Binding Factor-Beta [66], and Vaccinia virus Glutaredoxin-1 [75]. In order to enable high-throughput data collection, JIGSAW utilizes only four NMR experiments (Sec. D.1.1): heteronuclear single quantum coherence spectroscopy (HSQC), H^N-H^a-correlation spectroscopy (HNHA), 80 ms total correlation spectroscopy (TOCSY), and NOESY. This set of experiments requires only a few 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 cheaper process than ^{13}C labeling (cf. Wuthrich [122]: *"A big asset with regard to future practical applications... [is] ... straightforward, inexpensive experimentation. This applies to the isotope labeling scheme as well as to the NMR spectroscopy..."*) 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. [33, 32]) and global fold (e.g. [113]), 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-b, 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 b-sheet connectivity. Finally, the spectral assignment produced by JIGSAW is itself an important product. Backbone assignments must be obtained before dipolar coupling measurements [114, 103] can be employed for structure determination [90, 34, 50, 68]. One use of assigned NMR data in addition to structure determination is the analysis of protein structural dynamics from nuclear spin relaxation (e.g. [97, 96, 74]). Assignment is necessary to determine the residues implicated in the dynamics data. Another application is high-throughput drug activity screening. Even if a crystal structure is already known, protocols such as SAR by NMR (discussed above) perform NMR experiments 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: (I) *graph-based secondary structure pattern discovery*, and (II) *assignment by alignment*. (I) 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 biophysically-based constraints on valid groups of NOE interactions to manage the large search space of possible secondary structure patterns. (II) 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. For brevity, we call (II) *"alignment,"* and say that JIGSAW *aligns* the secondary structure elements from (I) to the primary sequence.