6-1-2010

Graph algorithms for NMR resonance assignment and cross-link experiment planning

Fei Xiong
Dartmouth College

Graph algorithms for NMR resonance assignment and cross-link experiment planning

Dartmouth Computer Science Technical Report TR2010-675

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

in

Computer Science

by

Fei Xiong

DARTMOUTH COLLEGE

Hanover, New Hampshire

June 2010

Examining Committee:

(chair) Chris Bailey-Kellogg

Devin Balkcom

Lisa Fleischer

Gopal Pandurangan

Brian W. Pogue, Ph.D.
Dean of Graduate Studies

Abstract

The study of three-dimensional protein structures produces insights into protein function at the molecular level. Graphs provide a natural representation of protein structures and associated experimental data, and enable the development of graph algorithms to analyze the structures and data. This thesis develops such graph representations and algorithms for two novel applications: structure-based NMR resonance assignment and disulfide cross-link experiment planning for protein fold determination. The first application seeks to identify correspondences between spectral peaks in NMR data and backbone atoms in a structure (from x-ray crystallography or homology modeling), by computing correspondences between a contact graph representing the structure and an analogous but very noisy and ambiguous graph representing the data. The assignment then supports further NMR studies of protein dynamics and protein-ligand interactions. A hierarchical grow-and-match algorithm was developed for smaller assignment problems, ensuring completeness of assignment, while a random graph approach was developed for larger problems, provably determining unique matches in polynomial time with high probability. Test results show that our algorithms are robust to typical levels of structural variation, noise, and missings, and achieve very good overall assignment accuracy. The second application aims to rapidly determine the overall organization of secondary structure elements of a target protein by probing it with a set of planned disulfide cross-links. A set of informative pairs of secondary structure elements is selected from graphs representing topologies of predicted structure models. For each pair in this "fingerprint", a set of informative disulfide probes is selected from graphs representing residue proximity in the models. Information-theoretic planning algorithms were developed to maximize information gain while minimizing experimental complexity, and Bayes error plan assessment frameworks were developed to characterize the probability of making correct decisions given experimental data. Evaluation of the approach on a number of structure prediction case studies shows that the optimized plans have low risk of

error while testing only a very small portion of the quadratic number of possible cross-link candidates.

Acknowledgements

This work would never have been done without the support from many people. It is my pleasure to thank these people here.

I can never overstate my gratitude to my Ph.D. advisor, Prof. Chris Bailey-Kellogg. Without his encouragement, inspiration, sound advice, and excellent teaching, I would never be able to make my accomplishment. I was always impressed by his insight in science, his enthusiasm in computational biology research, his great efforts to explain things precisely and concisely, and his patience throughout my whole Ph.D. study and the thesis-writing period.

I would like to acknowledge the help of Prof. Gopal Pandurangan (Math, Nanyang Technological University). He and Prof. Bailey-Kellogg inspired me with the idea of using randomized graph to solve the NMR resonance assignment problem. Without his help, this work could not have been done in such an elegant manner, and finally been published in the ISMB proceedings. He also gave me many valuable advice during my writing of the PhD dissertation and helped to expedite my final completion. My sincere thanks will go to Prof. Lisa Fleischer (CS, Dartmouth) and Prof. Devin Balkcom (CS, Dartmouth) as well. They kindly served as my thesis committee members and provided a lot of insightful comments to improve the quality of my work. I also would like to extend my gratitude to Prof. Bruce Donald (CS, Duke) and Prof. Ryan Lilien (CS, Toronto), who gave me great suggestions for the future work in NMR assignment when we met in the ISMB conference.

I wish to thank many faculty and student colleagues. I thank Prof. Alan Friedman (Bi-

ology, Purdue), Prof. Bruce Craig (Statistics, Purdue), and Michal Gajda (IIMCB, Poland), for stimulating discussion about NMR resonance assignment, cross-link based protein fold determination, and lots of other interesting topics.

I am grateful to all my friends in the CBK lab. They have generously helped me in many aspects. These lovely people are: Dr. Shobha Potluri, Dr. Xiaoduan Ye, Dr. John Thomas, Dr. Jason Vertrees, Dr. Wei Zheng, Bornika Ghosh, Himanshu Chandola, Andrew Parker, Lu He, and Tuobin Wang.

At last, I would like to thank my parents and my girl friend, Yan Gao. They have supplied tremendous support to me over the past years. They shared my joy and tears, and their patience and encouragement helped me sustain through the long time study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Proteins play crucial roles in almost every biological process. Each protein starts from a sequence of amino acids, but it is not functional until folded into some particular three-dimensional structure. Thus structural studies of proteins are central to post-genomic tasks in modeling, predicting behaviors of, and controlling the molecular machinery of the cell. There are four distinct aspects of a protein's structure (Figure 1.1). The *primary structure* is the sequence of amino acids in a polypeptide chain. The *secondary structure* consists of the geometry of segments in forms such as alpha-helices or beta-sheets. The *tertiary structure*, or *fold*, describes the spatial arrangement of the secondary structures and the relative positions of all the atoms in a chain. The *quaternary structure* describes how two or more protein chains interact to form a complex. To understand protein functions at the molecular level, we need to know structures at these various levels. This is the topic of structural biology, which employs experimental techniques such as x-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy.

In addition to providing insights into biomolecular function, analysis of protein structures has a significant impact on important applications. It can guide the development of inhibitors that can disable undesired protein functions (i.e., drug design), or help to re-design amino acid sequences to introduce/improve particular enzymatic activities (i.e., protein de-

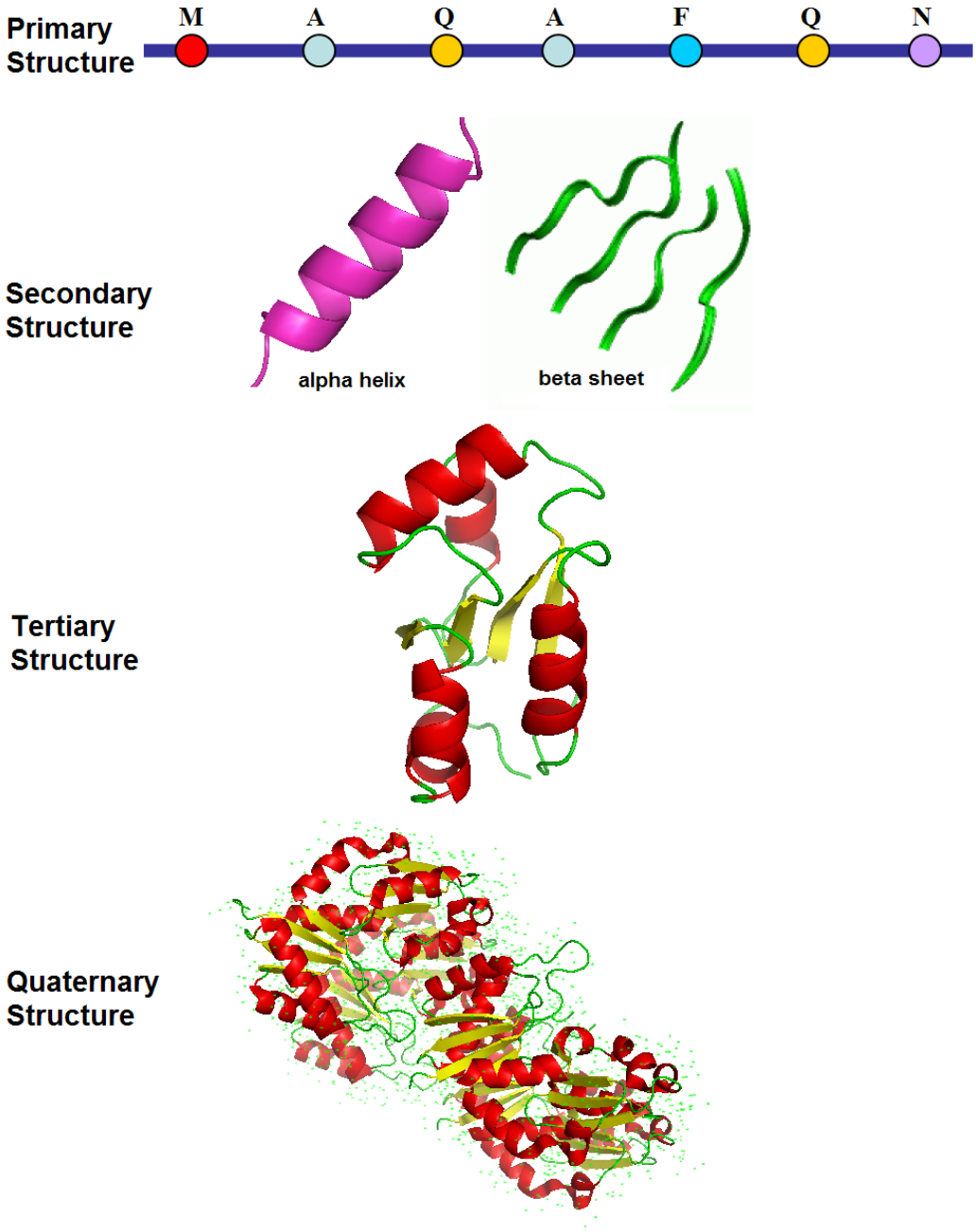


Figure 1.1: Main levels of protein structure.

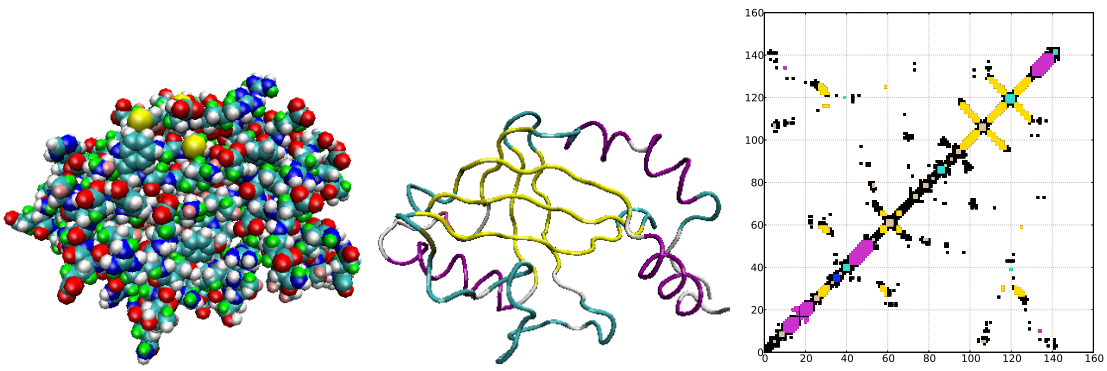


Figure 1.2: A protein is a linear chain (sequence) of amino acids, which folds up into a three-dimensional structure in solution. We can visualize a structure by plotting all atoms in a space-filling representation (left) or a trace of just the backbone atoms (middle), including both α-helices (purple, outer) and β-strands (yellow, center). Alternatively, the structure can be represented as a contact graph (right), where vertices are amino acids and edges are the physical contacts or proximity of vertices. Here the contact graph is visualized as a matrix with rows/columns for vertices and non-zero cells for edges. When a contact pair is within an α-helix or β-strand, its color matches that in the middle panel.

sign). However, there are still significant challenges in determining, analyzing and utilizing protein structure, due to the difficulty of experiments and the consequent noise/error in the collected data. They in return require the development of models and algorithms to properly process the information content of the data, adequately explore the solution space, and rigorously characterize the results.

To analyze a protein's structure, we can naturally consider it as a graph. For example, a "contact graph" (Figure 1.2) can be generated where the vertices are backbone residues and the edges capture the chemical bonds/spatial interactions between atoms in different residues. A graph representation of protein structure is natural due to the underlying physics — atoms are located in space and only a limited set of "neighbor" atoms have the strongest interactions. Also, atomic interactions are often the key features in studies of protein folding, dynamics and function. A graph representation of structure is also sufficient in that it uniquely encodes coordinate information into spatial proximity, when the edges have interatomic distance measurements [26].

Based on graph representations, graph algorithms have been widely used to investigate protein structure. Example applications include the clique-finding approach for comparative modeling [59], the frequent subgraph mining approach for structure motif finding [5, 27, 77], protein contact graph mining for folding pathways [81], secondary structure based protein topological model [42], and tree-decomposition approaches for protein structure prediction [72, 73]. These approaches have proved that advanced graph algorithms, such as graph pattern mining or graph-based feature selection, can be successfully applied to discover substantial properties of proteins.

Research conducted in this thesis was motivated by the success of using graph algorithms to analyze protein structures, and introduces two significant new applications — contact replacement for NMR backbone resonance assignment and cross-link experiment planning for protein fold determination. We have successfully converted these into graph problems (graph matching and graph-based feature selection). The next two sections will introduce the problem scopes and overview our solutions.

1.1 Contact Replacement for NMR Resonance Assignment

Nuclear magnetic resonance (NMR) spectroscopy is playing an increasingly important role in studies of proteins beyond the determination of their three-dimensional structures. For example, since NMR is performed in solution, it can gather information regarding dynamics [33, 48] and structure-function relationships under varying conditions [43]. Similarly, solution NMR is a vital tool in assessing ligand binding for drug development [24, 61] and can also help characterize protein-protein interactions [11]. These applications of NMR are significant even if the structure has already been determined by x-ray crystallography or a high quality homology model is available.

Therefore we would like to develop a novel algorithm that exploits an available structure to interpret NMR data, supporting studies of dynamics and interactions; see Figure 1.3. In

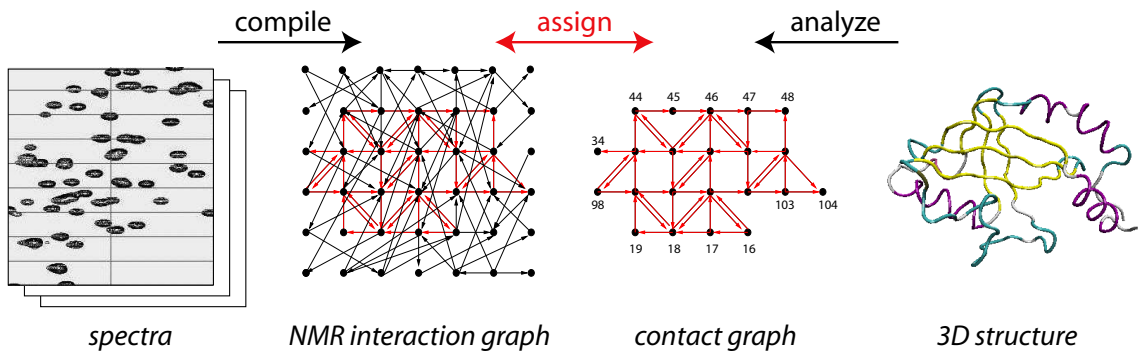


Figure 1.3: Contact-based NMR resonance assignment. Both an existing three-dimensional structure and NMR data (based on the through-space NOESY experiment) are represented as graphs. The interaction graph representing the NMR data is essentially a corrupted, ambiguous version of the contact graph representing the structure. The goal is to uncover the correspondence.

the targeted scenarios, the key bottleneck is to determine the *resonance assignment* mapping NMR data to specific atoms in the protein (e.g., those affected by ligand binding). While backbone resonance assignment (i.e., resonance assignment of the backbone atoms) in support of structure determination has been well-studied (e.g., [3, 30, 32, 40, 44, 64, 65, 76, 84]), we aim essentially to invert the process, and use the structure in support of assignment.

Here we formulate the problem of assignment given a structure and minimalist NMR data as the *contact replacement problem* (Figure 1.3). A contact graph representing a protein structure has vertices for the individual amino acid residues in the protein and edges between nearby pairs. A particular form of "interaction graph" representing NMR data has vertices for NMR-probed "pseudoresidues" (which correspond via an unknown mapping to the real residues), and edges between pairs that, if they were nearby, would explain the data. The NMR edges are essentially the contact edges, significantly corrupted by experimental noise and ambiguity (around 5 noisy edges per correct one). The contact replacement problem is then to uncover the correspondences between these graphs for a given protein.

The name "contact replacement" for our problem is inspired by the names for the analo-

gous problems "molecular replacement" in x-ray crystallography [57] and "nuclear vector replacement" in NMR [37, 39]. In molecular replacement, initial data interpretation is aided by matching against available structural information from a related protein. Likewise, in nuclear vector replacement, residual dipolar coupling data are matched against predictions from an available structure (or high-quality model). Contact replacement and nuclear vector replacement are complementary, relying on different types of NMR data with different information content (distances vs. orientations). The contact replacement problem is related to threading (sequence-structure alignment), but for threading, residues are in sequential order for both the sequence and the structure, whereas here we have no information about the sequential order of the pseudoresidues.

Given this problem definition, we have developed two complementary approaches which focus on different aspects of this problem and have different objectives. We first developed a Hierarchical Grow-and-Match (HGM) algorithm to effectively search the whole solution space to ensure the completeness of assignment results. Our algorithm decomposes the contact graph into sequential fragments with relatively dense interactions, and then combines possible assignments for the fragments, searching over the combinations with effective but conservative pruning criteria. Our algorithm is guaranteed to identify all solutions consistent with the data within a likelihood threshold of the optimal solution. It also deals correctly and uniformly with missing edges, which are quite common when graphs are compiled from the experimental data.

However, HGM only applies to well formed secondary structure elements. The guarantee of completeness was based on the conservative branch-and-bound search, which may take a few hours to a few days to assign a single protein. To better serve the needs of large-scale NMR assignment, we developed an alternative random graph approach that gives up the completeness guarantee but assigns the entire backbone and returns high quality results significantly faster in both theory and practice. In this approach, we first show that by com-

bining connectivity and amino acid type information, and exploiting the random structure of the noise, one can provably determine unique correspondences in polynomial time with high probability, even in the presence of significant noise, i.e., a constant number of noisy edges per vertex. We then detail an efficient randomized algorithm and show that, over a variety of experimental and synthetic datasets, it is robust to typical levels of structural variation, noise and missing.

Our main contributions in contact replacement for NMR resonance assignment include:

- We have introduced the contact replacement approach to do NMR backbone resonance assignment by referring to a known structure and using a minimalist set of experiments, primarily the through-space NOESY.

- We have developed two complementary algorithms for contact replacement: a hierarchical grow-and-match algorithm focused on SSEs, which is guaranteed to find all solutions consistent with the data, and a random graph approach, for the whole protein, which can provably determine unique matches in polynomial time with high probability.

- We have demonstrated the effectiveness of our algorithms on both experimental and synthetic data, with significant noise and missings. We have also shown that the random graph approach has good tolerance to moderate structural uncertainty.

1.2 Protein fold determination

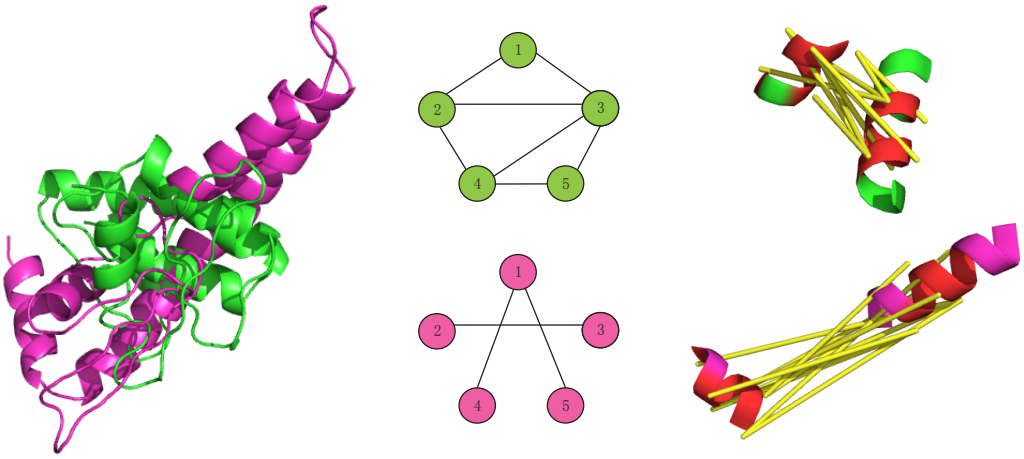
Although a vast number of protein sequences have been identified, just a tiny fraction ($\approx 1\%$) of them have experimentally determined three-dimensional structures [1, 19, 68], due to the difficulty of related wet-lab experiments. At the same time, since structure is more conserved than sequence, only a small number of folds (overall structural organizations) exist in nature [20, 47, 83] and many have been recorded in public protein databases [6].

Fold recognition (including "threading") [19, 74, 75] takes advantage of this by performing a protein database search and using the available 3D structures to predict the structure of a target sequence. The outputs are crude structures and may vary significantly due to different choices of score function [45, 82]. A near-native structure is often among a set of predictions with high scores, but not necessarily the highest one. Therefore, to correctly and confidently determine the fold, it is necessary to use additional wet-lab approaches to reduce the ambiguity and evaluate the set of proposed solutions.

To bridge computational structure prediction and experimental structure determination, different research groups [10, 13, 21, 35, 78–80] have worked on the problem of "structure elucidation" where predicted structures can be selected by relatively quick biophysical/biochemical experiments instead of the more expensive and time-consuming X-ray or NMR technologies. One experimental type that particularly suits this need is *cross-linking*. Compared with NMR-based techniques, cross-linking is generally a cheaper experimental approach but provides only the relatively coarse information of a pair of residues being within a relatively large distance (e.g., up to 19 Å [79]) at some moment. Although it may not be sufficient to determine the protein structure, previous research [35, 79, 80] showed that cross-linking can be used to select the correct model from a set of high quality predictions. Disulfide cross-linking [9, 28, 36] is a form of cross-linking with advantages in easy and reliable experimentation, quality of resulting information content [79], and plannability of residue locations. In disulfide cross-linking, cysteines are specifically introduced at selected residue positions; the resulting presence / absence of disulfide bonds can then be evaluated by alteration in electrophoretic mobility [9, 36, 79].

In general, it is not practical to conduct disulfide cross-linking over all the possible residue pairs of a even small-size (100 residues) protein. Therefore, the major problem of using cross-links in protein structure analysis is to identify the best cross-link sites (pairs of residue positions to test experimentally). Overall, the geometric feasibility (or simply

the direct distance) of sites is what determines the likelihood of cross-link formation. Thus residue positions with distinct geometric feasibility across different models will be more informative. But potential model errors and structure flexibility will make it difficult to locate the exact cross-linking sites, and we need to carefully model and compensate such kind of noise.



Predicted protein structures SSE contact graph Cross-link experiment plan

Figure 1.4: Protein fold determination by disulfide cross-linking. Protein structures are predicted in 3D representations (left), and then compiled into 2D SSE contact graphs (middle). Structure differences are evaluated through contact graphs, with the most distinctive patterns (*topological fingerprints*) selected through a graph-based feature selection approach. Then an experiment planning algorithm is applied to select a set of residue pairs for disulfide cross-linking (right), to be experimentally evaluated in order to identify the true protein fold.

In contrast to previous researches that focus more on probing the 3D structure geometry and trying to select specific models, in this thesis we introduce a new approach called *protein fold determination*. We target a higher level of structure characterization—"fold"—the spatial organization of protein secondary structure elements (SSEs). An integrated computational-experimental method has been developed to determine the fold of a target protein by probing it with a set of planned disulfide cross-links. We start with predicted structural models obtained by standard fold recognition techniques. In a first stage (Figure 1.4, left – middle), we characterize the fold-level differences between the models in

terms of topological fingerprints, and select a small set of SSE pairs that differentiate the folds. In a second stage (Figure 1.4, middle – right), we determine a set of residue-level cross-links to probe the selected SSE pairs. Each stage employs an information-theoretic planning algorithm to maximize information gain while minimizing experimental complexity, along with a Bayes error plan assessment framework to characterize the probability of making a correct decision once data for the plan are collected. By focusing on overall topological differences and planning cross-linking experiments to probe them, our fold determination approach is robust to noise and uncertainty in the models (e.g., threading misalignment) and in the actual structure (e.g., flexibility). We demonstrate the effectiveness of our approach in case studies for a number of targets in CASP database [45], showing that the optimized plans have low risk of error while testing only a very small portion of the quadratic number of possible cross-link candidates. Fold determination can overcome scoring limitations in purely computational fold recognition methods, while requiring less experimental effort than traditional protein structure determination approaches.

Our main contributions in protein fold determination include:

- We have introduced the protein fold determination approach to characterize protein's structure topology with an integrated computational-experimental solution.
- We have developed information-theoretic selection algorithms for both topological fingerprint identification and robust cross-link experiment planning.
- We have introduced Bayes error frameworks for quality evaluation of the experiment plans.
- We have demonstrated the effectiveness of our approach with analysis of different error/risk factors and conducted end-to-end simulation studies. We have also shown the approach is robust to errors in models, and can account for the case when none of the models is correct.

Chapter 2

Contact Replacement for NMR Resonance Assignment: A Hierarchical Grow-and-Match (HGM) Algorithm

2.1 Overview

Nuclear magnetic resonance (NMR) spectroscopy enables analysis of protein structure, dynamics, and interactions in near-physiological conditions. While much work has focused on the use of NMR in structure determination (including at a genomic scale [55]), applications in studies of dynamics and interactions can be equally important, *even if the structure has already been determined by crystallography or modeled computationally*. NMR-based methods enable rapid and cost-effective screening for binding, and have become a vital tool in drug development [24,61] as well as characterization of protein-protein interactions [11]. Since NMR does not require crystallization of the sample, conditions can be varied in order to study structure-function relationships [43]. Nuclear spin relaxation provides insights

[1]This work has been published [69] in the proceedings of IEEE International Conference on Bioinformatics and Bioengineering (BIBE) 2007.

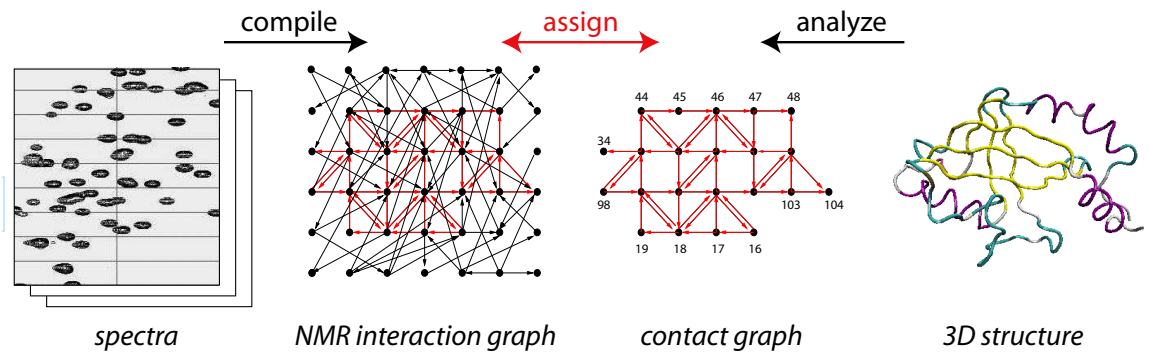


Figure 2.1: Assignment given 3D structure. Both an existing three-dimensional structure and NMR data (based on the through-space NOESY experiment) are represented as graphs. The interaction graph representing the NMR data is essentially a corrupted, ambiguous version of the contact graph representing the structure. The goal is to uncover the correspondence.

into protein structural dynamics (again, in solution) [33, 48].

Figure 2.1 shows the flow of information in NMR-based studies. We then formulate the problem of assignments given a structure and minimalist NMR data as the *contact replacement* problem. As discussed above, the already available structure provides the necessary information about the intrinsic properties of that protein, and those properties should be encoded in the NMR experimental data as well. Thus we consider it as an "inversed" process of the structure determination problem: Given a 3D structure, use the constraints it reveals to process data generated from NMR experiments. Our approach represents both the NMR data and the structure as graphs (the "NMR interaction graph" and the "contact graph", respectively); the NMR graph is essentially a corrupted version of the contact graph, with an unknown correspondence between vertices (and thereby edges). Our goal is to find the correspondence (middle of Figure 2.1). We note that this problem is different from those addressed by previous graph-based approaches [3, 4, 32], which focused on uncovering patterns in an NMR graph rather than matching it to a specified contact graph. Significant challenges still exist to develop and analyze efficient algorithms that guarantee robust performance and enable to explore the information content in both

protein structures and NMR data.

In this work we developed a novel algorithm that utilizes the known structure of a protein to guide its NMR backbone resonance assignment. The key features of our method are as follows:

- It is *complete*, guaranteed to determine all solutions consistent with the data (to within a likelihood threshold of the optimal one). This is particularly important for sparse datasets and with somewhat subjective scoring functions, where we must be careful to characterize the similarities and differences among competing high-quality solutions [64].
- It takes advantage of a *structure* to effectively decompose the solution space, and then efficiently search through it by hierarchically merging partial solutions and eliminating those that provably cannot lead to complete solutions of sufficient quality.
- It is *minimalist*, requiring only 4 spectra from ^{15}N-labeled protein, saving substantial spectrometer time and substantial expense compared to standard triple-resonance-based assignment methods.

We demonstrate the effectiveness of our algorithm in studies of a number of proteins with experimental data and with simulation studies under varying amounts of noise and sparsity. In comparison with some available techniques, Section 2.5 further elaborates on the general context of our work.

2.2 Problem Definition

We first summarize the representations of the input contact graph and NMR interaction graph; for details see [3, 4, 32].

Contact graph. $G^* = (V^*, E^*)$, where V^* is a set of residue positions and E^* is a set of pairs of nearby residue positions. In particular, we place an edge when a pair of protons

is within a specified distance threshold (say, 3, 4, or 5 Å). Each vertex v is labeled with its amino acid type, $a(v)$.

NMR interaction graph. $G = (V, E)$, where V is a set of *pseudoresidues* of unknown correspondence to the residues and E is a set of pairs of pseudoresidues that may have interacting protons (i.e., an interaction would explain a peak in the NOESY spectrum). Such a graph can be compiled from a set of four ^{15}N spectra (HSQC, HNHA, TOCSY, and NOESY), and has a number of properties [32, 56, 69]:

- Each vertex is labeled with a *secondary structure type*, either α or β, as determined from HNHA.
- Each vertex is labeled with a list ℓ of *possible amino acid types*. We use here the classes output by RESCUE [52], which employs a two-level neural network to estimate amino acid type from proton chemical shifts. The first level associates a pseudoresidue with one of the ten type classes (IL, A, G, P, T, V, KR, FYWHDNC, EQM, and S) with very high accuracy (avg: 91.9%, min:88.1%); amino acids within a class are treated as indistinguishable.
- Each edge is labeled with an *interaction type* based on the chemical shift ranges. We use only H^N and H^α, since a structure model's side-chain atomic coordinates are usually less reliable, and we have not found their inclusion to aid the results.
- Each edge has a *match score* s, evaluating the quality of the edge as an explanation for the peak. Typical scoring rules (e.g., [22, 64, 66, 84]) compare absolute or squared difference in chemical shift; except for noise (reasonably modeled as Gaussian), the correct edge should match exactly and have the best score. Here we score edges by error probability, i.e., how likely it is that an edge could be generated by noise. In this way, missing edges are naturally penalized since they contribute a score of zero. The score is thus $-\log\left(1 - \frac{1}{\sqrt{2\pi}\sigma}\exp^{-\frac{\Delta(e)^2}{2\sigma^2}}\right)$ where $\Delta(e)$ is the chemical

shift difference for edge e, and σ is the standard deviation of chemical shift difference distribution.

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

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

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

2.3 Methods

We develop an algorithm based on two insights: multiple consistent edges are necessary to obtain effective constraint on an assignment, and sequential edges (i.e., residue i to residue $i + 1$) provide an appropriate basis for uncovering a correspondence. Noise edges in the NMR graph result from chemical shift degeneracy, and are not correlated with spatial proximity. Thus we are more confident in a set of NMR edges that consistently match a set of contact edges; it is not likely that many false positives can "conspire" to match properly. Confidence is gained by both the number of edges and their *density* (number of edges di-

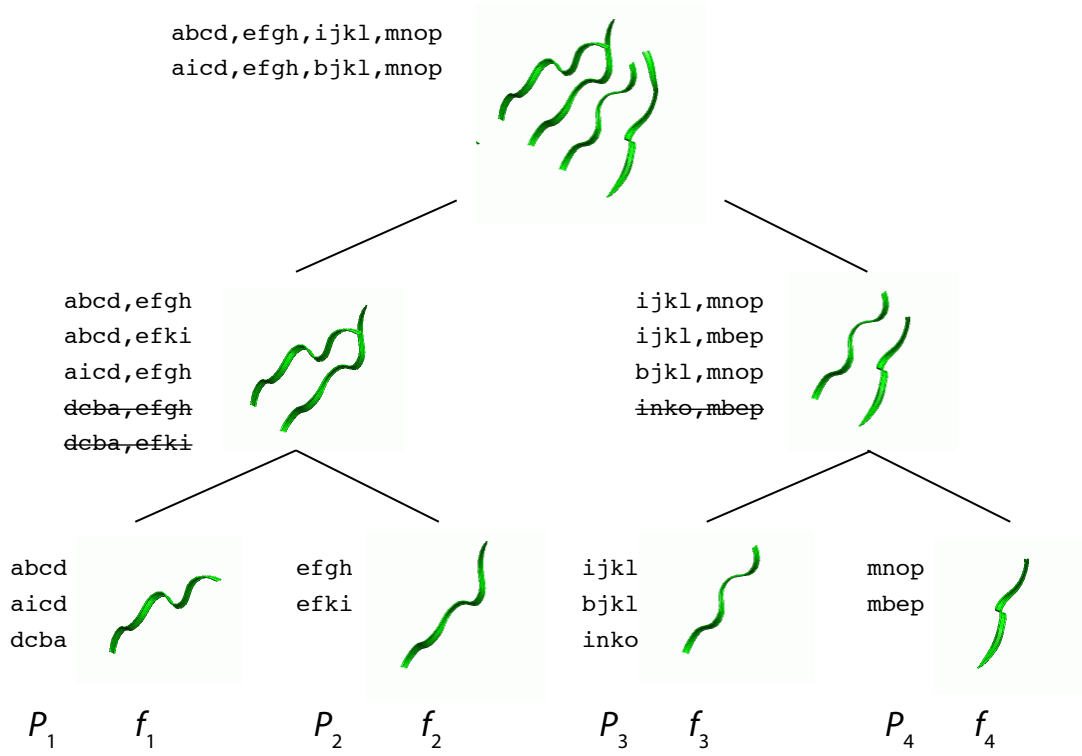


Figure 2.2: Intuition for our Hierarchical Grow-and-Match algorithm. The 3D structure is decomposed into fragments; here a β-sheet is decomposed into $f_1, \ldots, f_4$ (in practice, each strand may be composed of multiple fragments). The decomposition accounts for both contact density and the combinatorics of the sets of possible assignments; here $P_1, \ldots, P_4$ are the sets of possible "pseudofragments", each listing pseudoresidues (indicated by different letters) that could correspond to the fragment's residues. Pseudofragments are merged according to a hierarchical "merge tree"; those that are inconsistent (use the same pseudoresidue) are immediately pruned and are not shown. A conservative bound (not illustrated) eliminates some combinations that provably cannot lead to a near-optimal solution; these are illustrated by strike-throughs.

vided by number of vertices). While all contacts are important, to construct an assignment, we want to focus on edges likely to match well, and sequential residues reliably are in contact and reliably generate NOESY peaks. Furthermore, the one-dimensional structure of sequential residues makes strings of them relatively easy to manipulate. To incorporate both density and sequentiality, our algorithm first forms fragments based on sequential interactions, and then hierarchically merges them according to the contact density of edges across two fragments (Figure 2.2). We now detail this process.

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

Previous work [64, 65] on assignment using sequential fragments raised two important considerations: the fragments should be long enough to provide sufficient constraint, but short enough to keep under control the combinatorial number of corresponding pseudofragments. The same holds here, with an extra twist for our context: the fragments should account for contact density. We thus take advantage of the natural organization of secondary structure elements—the decomposition should "respect" helices and sheets, using them as core fragments (perhaps subdivided, to control pseudofragment combinatorics). We follow the basic previous approach [64, 65] of growing one fragment until there would be too many corresponding pseudofragments (according to a threshold θ), and then starting a new fragment. We also start a new fragment when the secondary structure type changes. If a single-residue fragment results, we merge it into the previous fragment. The result is a set $S = \{f_1, f_2, \ldots\}$ of sequential fragments and a set of sets $\mathcal{P} = \{P_1, P_2, \ldots\}$, where $P_i = \{p_{i1}, p_{i2}, \ldots\}$ is a set of alternative pseudofragments.

Merge Tree The decomposition of the graphs into fragments and pseudofragments yields high-quality sequential "building blocks"; we combine these based on contact density. β-sheets provide intuition (see Figure 2.2)—there are relatively dense connections between adjacent strands, so it makes sense to merge sequential fragments for the strands into (no longer sequential) fragments for the sheet. More generally, let us define a *merge tree*, such that parents represent the unions of their children, and the leaves are the sequential fragments. We call the unions *fragments* (not necessarily sequential), and thereby extend

the set S of sequential fragments to a set F of fragments, each of which is the union of the sequential fragment leaves below a node in the tree. The root thus represents a fragment including all residues, to which we want to assign a pseudofragment including all pseudoresidues.

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

Scoring and Bounding An assignment (fragment and corresponding pseudofragment) must satisfy the "hard" constraints of consistency of amino acid type and secondary structure type, and uniqueness of residues and pseudoresidues. It can then be evaluated for how well it explains the data:

$$s(f,p) = \sum_{(e^*,e)\in m(f,p)} w(e) + \phi^*(f,p) + \phi(f,p) \tag{2.1}$$

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

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

$$s_2(f_i, f_j; \mathcal{P}') = \min_{p_i \in P_i, p_j \in P_j} \left(\frac{s(f_i, p_i)}{n_i} + \frac{s(f_j, p_j)}{n_j} + s(f_i, f_j, p_i, p_j) \right) \tag{2.2}$$

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

$$b(F', \mathcal{P}') = \sum_{f_i \neq f_j \in F'} s_2(f_i, f_j; \mathcal{P}') \tag{2.3}$$

We can prove that (2.3) is a lower bound on the match score of any complete pseudofragment, and our results below (Figure 2.4) also indicate that it is fairly tight.

```
s_* ← ∞
define HGM(k, p, P)
    if k > |F|
        // complete fragment
        output p; s_* ← min {s_*, s(f_{k−1}, p)}
    else
        let f_i, f_j be the fragments merged to form f_k
        foreach (p_i, p_j) ∈ P_i × P_j,
                    sorted by s(f_i ∪ f_j, p_i ∪ p_j)
            if p_i ∩ p_j = ∅ and p ∩ (p_i ∪ p_j) = ∅
                // no shared pseudoresidues
                p' ← p ∪ p_i ∪ p_j
                let P' be a copy of P with P_k fixed to {p'}
                if b(F, P') < s_* + Δ
                    // satisfied bound
                    HGM(k + 1, p', P')
```

Figure 2.3: Hierarchical Grow-and-Match Search. The arguments are k: fragment index; p: pseudofragment; $\mathcal{P}$: set of sets of alternative pseudofragments. An ensemble is output, and the best score is maintained in s_*.

Search Algorithm Based on the merge tree, we can assemble larger and larger pseudofragments; based on the bound, we can eliminate pseudofragments that are guaranteed to be sufficiently suboptimal. Thus we perform Hierarchical-Grow-and-Match as a depth-first search with conservative pruning (i.e., only eliminate partial solutions guaranteed to be suboptimal). The search (Figure 2.3) follows the structure established by the merge tree: to merge a pair of fragments, branch on the possible pairs of pseudofragments, ordered by score. (We assume binary trees, but the generalization is straightforward.) The search proceeds bottom-up, left-to-right, through the tree; the initial invocation is for the first non-leaf node, with an empty pseudofragment. We assume that the fragments are numbered accordingly—S is the set of sequential fragments and F has S followed by merged fragments in order. Thus to assemble fragment f_k, we merge fragments f_i and f_j, choosing one pseudofragment each from sets P_i and P_j. Pseudoresidues already used in earlier pseudofragments cannot be reused later in the same search branch. Upon merging the selected pseudofragments (setting P_k to the merged result), we verify that the bound is satisfied and

then recurse to the next fragment in the tree (f_{k+1}), with the pseudofragment assembled so far (p') and the choices of pseudofragments for the remaining fragments ($\mathcal{P}'$). The algorithm will find the optimal assignment, and by setting threshold Δ to be greater than zero, it will also find a complete ensemble of nearly-optimal solutions. Variations of this depth-first approach are straightforward; e.g., we also used a beam-search-like approach that propagates several choices simultaneously, in order to more rapidly identify a solution.

2.4 Results

Table 2.1 summarizes the datasets, both experimental and synthetic, that we used to validate HGM. The proteins are of moderate size for typical NMR studies. Since HGM separates residues and pseudoresidues by secondary structure type, we present here results for separate assignments of α-helices and β-sheets. For all datasets, we generated pseudofragments using a threshold (see Section 2.3 para. 2) of $\theta = 1000$. The likelihood threshold (see Section 2.3 para. 5) Δ was used to collect alternative assignments that are at most 100 times worse *a posteriori* than the optimal one.

2.4.1 Experimental datasets

We used four experimental datasets: three from previous contact-based assignment work [32], including Human Glutaredoxin (PDB ID: 1JHB), Core Binding Factor β (PDB ID: 2JHB), and the catalytic domain of GCN5 histone acetyltranferase (PDB ID: 5GCN); the fourth from the ST2NMR paper [56], *Paracoccus denitrificans* cytochrome C_{552} (PDB ID: 1QL4). For brevity, and since assignment is based on structure, we refer to each protein by its PDB ID. We constructed NMR interaction graphs from the already compiled pseudoresidues, adding vertex labels as described in the Methods. We generated edges from the the ^{15}N-edited NOE peak lists with conservatively large match tolerances: 0.05 for ^{1}H, 0.015 for

H^N, and 0.35 for ^{15}N. We used a proton-proton contact threshold of 4 Å, and selected the most representative contact graph from each deposited ensemble. Overall each correct edge had a mean of 2.5–6.0 noise edges, and the average missing rate is slightly above 30%, except for 1QL4 which has a significant missing rate of 62.5%.

Table 2.2 summarizes assignment results by HGM for the experimental datasets. Note that the search is rather efficient, explicitly testing only a small fraction of states (i.e., number of recursive calls of the HGM algorithm). For example, if we multiply the number of pseudofragments for each fragment, there are about 1.24×10^{10} possible states for 1JHB's β-sheets, of which 2716 are visited, and 1.93×10^{24} for 1QL4, of which 29 million are visited. The difference in number of visited states for different datasets is due to a combination of number of residues and pseudoresidues, ambiguity in amino acid type, number of edges, and noise and missing rates. For instance, 1JHB's 43 α-helix residues were represented in the contact graph by 160 edges, of which 52 are missing in the NMR graph. This leads to a very sparse graph to match and requires HGM to search deeper before being able to safely prune a sub-optimal solution.

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

Table 2.1: Datasets (top 4 experimental; bottom 8 synthetic)

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

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

Table 2.2: HGM results for experimental data

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

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

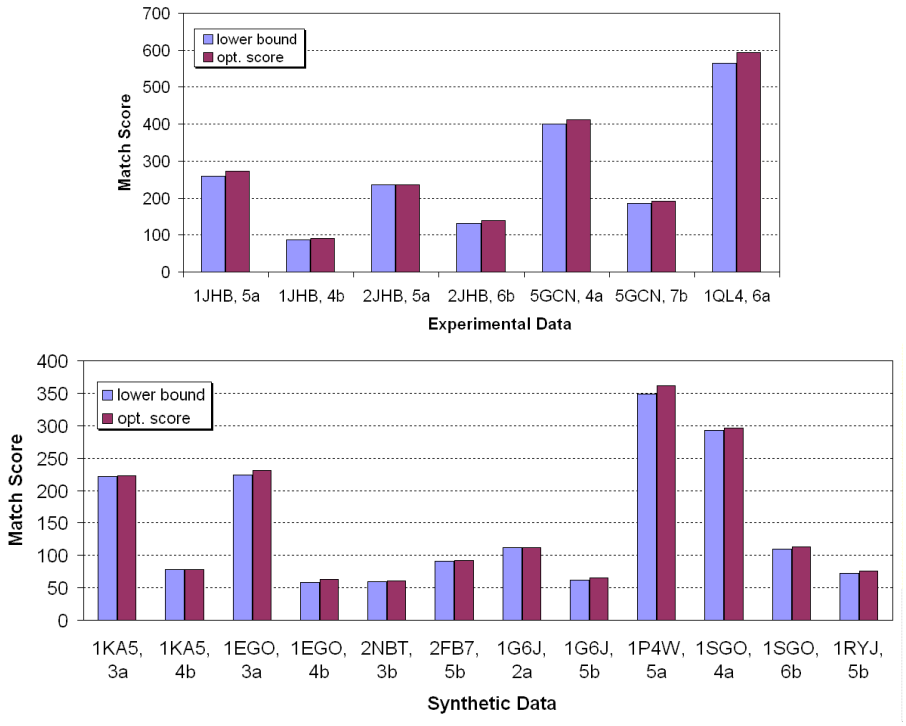
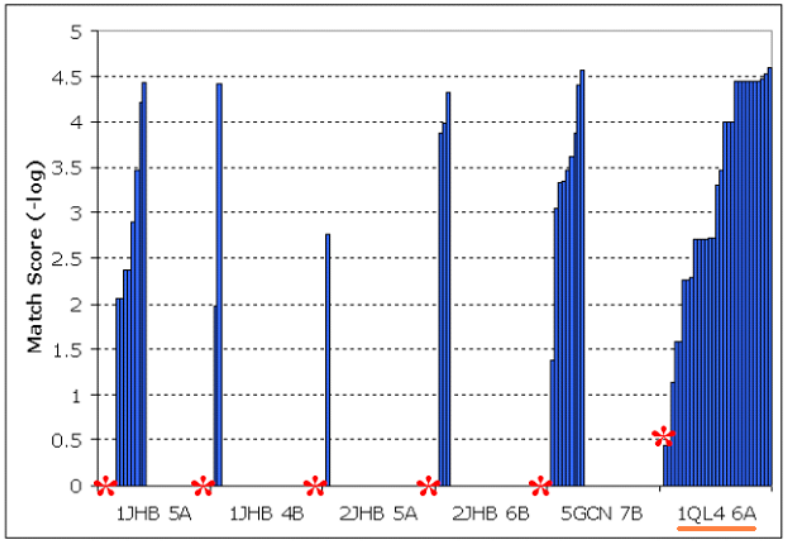
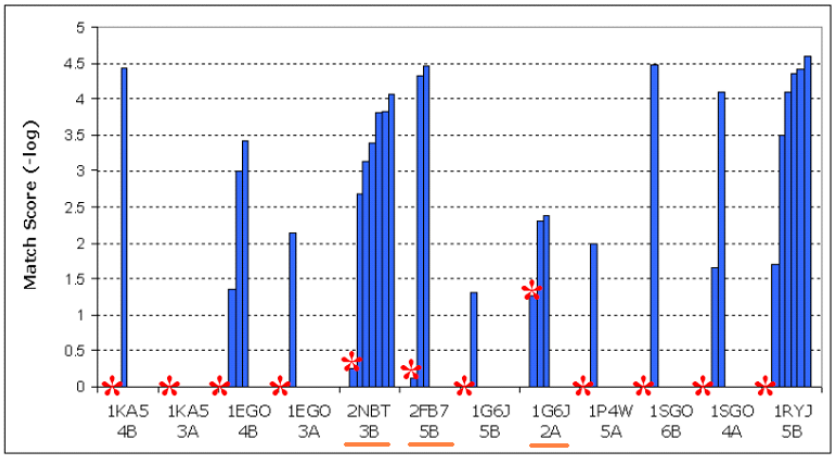


Figure 2.4: Comparison of HGM-identified lower bound on match score with the score of optimal assignment for experimental (top) and synthetic (bottom, 10, 30% missing) datasets.

Figure 2.4 illustrates the lower bound estimated by (2.3) relative to the optimal match score. In general, the bound is tight — the score difference is less than 8%.



(a) Experimental Data



(b) Synthetic Data (10, 30% missing)

Figure 2.5: Score distribution in the result ensembles ($\theta = 100$) of experimental and synthetic data sets. Displayed values have been adjusted regarding to the best score (ground zero) in each ensemble. Red asterisks indicate the reference assignment solutions. Orange underlines indicate where the reference solution is not ranked the best.

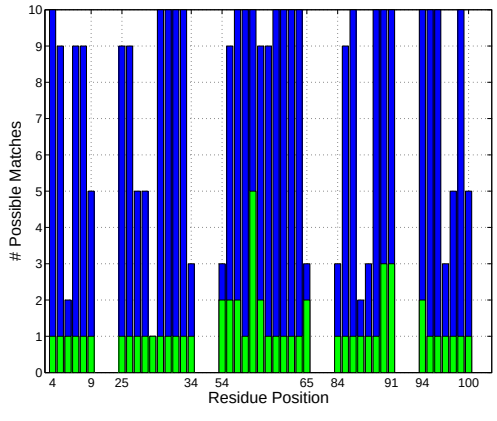
Deposited solutions, determined by expert spectroscopists, serve as 'reference' assignments. As shown in Figure 2.5, for each dataset, HGM found the reference assignment, but also a number of alternative assignments that were about as good (or perhaps even better), under our scoring model. The largest ensemble (30 solutions) was found for 1QL4, due to

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

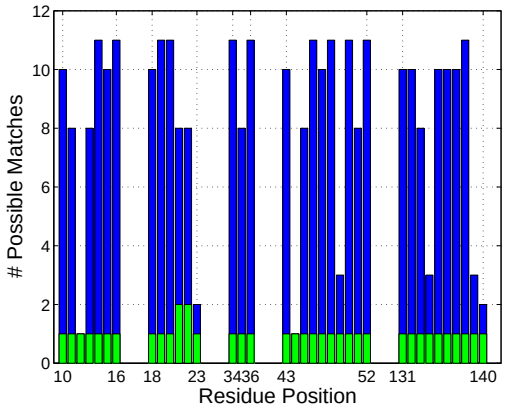
The differences among the high-quality assignments were typically confined to swaps between a few 'equivalent' pseudoresidues. For example, in the α-helices of 1JHB, pseudoresidues assigned to positions 58, 90 and 91 are exchangeable and the top four solutions of the ensemble provided all the possibilities. By computing a complete ensemble of feasible assignments, HGM allows us to carefully evaluate the remaining ambiguity, as Figure 2.6 illustrates. Overall, the average number of possible assignments for each residue position has been reduced from 7.9 to 1.2. The most significant ambiguity was caused by significant numbers of missing edges in particular secondary structure elements (e.g., in the first and the third helices of 1QL4). Figure 2.6 doesn't include 5GCN 4α, since it has only one solution returned by HGM. The mean number of assignments before HGM for 5GCN 4α is 11.86.

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

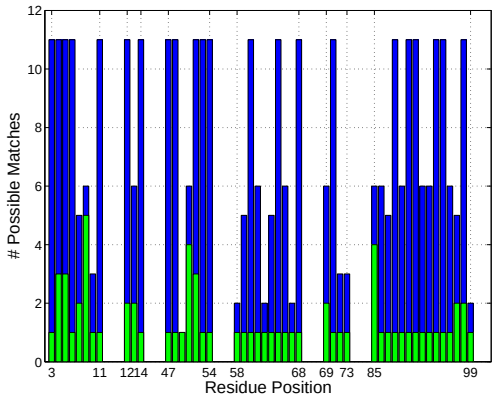
To the best of our knowledge, only one algorithm, ST2NMR [56], has been developed for backbone assignment based on a 3D structure and NOESY data. ST2NMR uses a Monte Carlo approach to optimize an assignment, explaining NOESY peaks in terms of distances in the structure. It was shown to be effective for some example test data, but



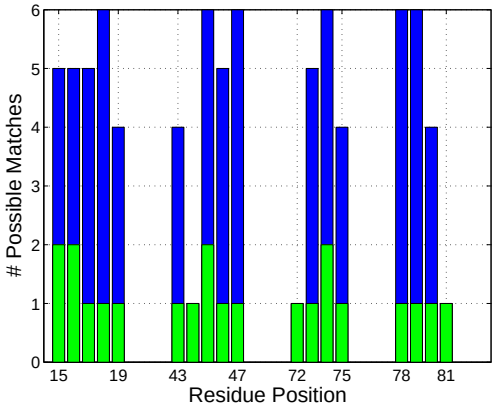
1JHB 5α: 7.65, 1.33



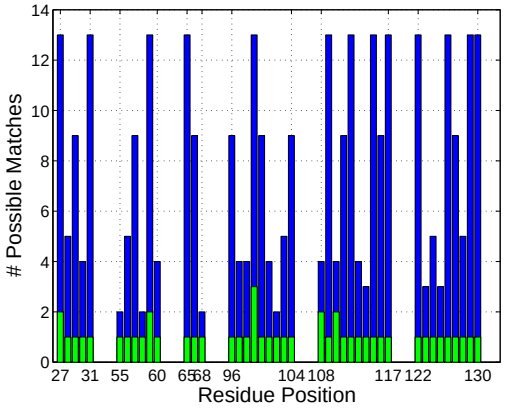
2JHB 5α: 8.33, 1.06



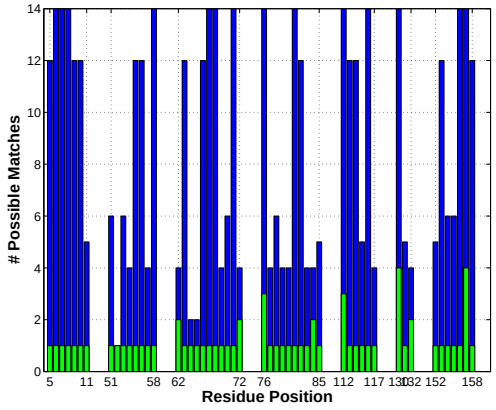
1QL4 6α: 7.60, 1.47



1JHB 4β: 4.44, 1.22



2JHB 6β: 7.71, 1.14



5GCN 7β: 8.73, 1.27

Figure 2.6: Assignment ambiguity before (blue) and after (green) HGM. The bars indicate how many pseudoresidues can be mapped to each residue *a priori* and within the HGM ensemble. The means before and after HGM are listed after their PDB ids.

requires specific experimental set-ups and can provide no guarantees or insights into the information content of the data. For a comparison, we tested ST2NMR on our 3 different experimental datasets (the publication already provided results for 1QL4), but found the resulting assignment to be highly dependent on the order of the input pseudoresidues. For example, over a set of 10 runs with different random pseudoresidue order for 1JHB, the assignment accuracy of ST2NMR varied from 28% to 77%, with a median of 56%. The results were confirmed by communication with the authors, who pointed out that ST2NMR depends critically on the combination of good structure and NOESY spectra, with good matches between them, and that ST2NMR works better with 2D homonuclear NOESY rather than 3D ^{15}N-edited NOESY. This result further emphasizes the need for complete search algorithms, like HGM, to be able to evaluate the reliability of an assignment in a situation with sparse, noisy data.

2.4.2 Synthetic datasets

In order to study our algorithm's performance under varying noise and sparsity levels, we also generated synthetic datasets using chemical shift data deposited in the BMRB [60]. We chose a random set of eight moderate-sized proteins previously tested with RESCUE [52], with varying α-helix and β-sheet content (refer again to Table 2.1). To construct the NMR interaction graph, we first simulated NOE peaks for pairs of interresidue backbone protons within a distance ≤ 4 Å. We likewise restricted the contact graph to 4 Å, thereby essentially ignoring the longer-distance NOEs, since we found in the experimental data that they are missing at a significant frequency and thus the information they provide is not worth the additional computational complexity they require. We randomly deleted peaks according to observed statistics correlating the missing probability with the interatomic distance [12], and tested two different missing rates at different distances: $d \leq 3$ Å, missing either 10% or 21%; $3 < d \leq 4$ Å, missing either 30% or 41%. Note that these rate ranges cover

the rates observed in the experimental datasets, except for the extremely sparse 1QL4. We generated an "observed" interresidue proton chemical shift for each remaining peak by adding Gaussian noise with variance 0.02 (corresponding to the 0.05 ^{1}H match tolerance). We added an edge to the NMR graph for each proton whose chemical shift matches the noisy value within the 0.05 threshold, yielding an average of 1.4–4.3 noise edges per correct edge. As discussed [32], all the noise is on the interresidue side of the interaction, since the intraresidue side has two chemical shifts (H^N, ^{15}N) by which to resolve chemical shift ambiguity. The simulated noise rates are somewhat smaller than the experimental ones, but we did not consider it realistic to increase the chemical shift tolerance in order to artificially inflate them.

Table 2.3 summarizes the results of HGM on the synthetic datasets. Here reference solutions indicate the original BMRB assignments. As with experimental data, for all test cases the reference assignment is included with an optimal or near-optimal score, and only a few assignment swaps differentiate the best solutions. Figure 2.4 illustrates that the lower bound remains quite tight. In general, the HGM algorithm performed very well and only explicitly tested a tiny fraction of the search space before finding the complete solution set. The synthetic tests further demonstrate the effect observed before: an increase in the missing rate yields greater search complexity and ensemble size (the last two sets of columns). These results also indicate that, given a uniform missing rate, α-helices tend to have a better tolerance for missings than do β-sheets since their tertiary structures are usually more compact and thus generate more edge constraints. Also, assignment of β-sheets benefits from the hierarchical merge order which naturally utilizes the spatial proximity of β-strands, even when sequentially separated. In such cases, a poor local assignment, e.g., many missing edges in a strand, can be effectively overcome by way of connections with neighboring strands.

Table 2.3: HGM results for synthetic data at two different missing rates.

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

Columns as described in Table 2.2.

2.5 Related Work

Our work differs from most traditional approaches to backbone resonance assignment in that it is structure based. As we showed in the Results, our guarantee of completeness appears to be very valuable, as ST2NMR (the only other existing approach using structure+NOESY) did not perform well on our sparse, noisy datasets. Our work is complementary to structure-based assignment approaches that reply primarily on experiments other than NOESY. For example, the NVR work by Langmead and Donald [37, 39] uses residual dipolar coupling (RDC) data as global orientational restraints, with only unambiguous NOEs to help prune. It remains interesting future work to fully integrate RDC with NOESY data and thereby perhaps overcome their individual limitations.

There are other techniques for assignment based on the NOESY, but they do not use information from an available 3D structure. The Main-Chain Directed approach represents an early approach to backbone assignment based on the NOESY [46, 62], although that work was developed for homonuclear spectra, only partially automated, and applied to experimental data for only one small protein. The automated Jigsaw approach [4] was successfully applied to uncover α-helix and β-sheet patterns in NOESY data (and thereby assign those regions). More recent work developed an algorithmic basis for the Jigsaw-style approach, with a randomized algorithm that gives optimal performance in expected polynomial time for the special case of uncovering secondary structures in corrupted NMR graphs [3, 32]. We note that we are focusing here on backbone assignment based on the NOESY. Work on NOE assignment (e.g., [67]), including side-chain interactions, is certainly related but typically is addressed only once backbone resonances have been assigned by standard techniques. An algorithm combining these two aspects for simultaneous backbone and side-chain assignment would represent an interesting, and significant, advance.

In our focus on matching graph representations of protein structures, our work is somewhat like sequence-structure alignment (threading); e.g., Xu and co-workers developed a

divide-and-conquer approach that uses hierarchical combinations of sub-alignments [75] and can incorporate *assigned* NOE data to constrain them [76]. A significant difference is that for threading, residues are in sequential order for both the sequence and the structure, while in contact-based assignment, the pseudoresidues come with no explicit order.

2.6 Summary

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

An important step for practical utility of HGM is to characterize its performance on x-ray structures and homology models, the natural inputs for structure-based assignment. We must model and account for any systematic differences between these models (generating contact graph edges) and the native solution state probed by NMR (generating NMR graph edges). Since HGM focuses on high-contact-density regions and the edges most likely to be observed, we believe that it will be fairly robust to modest structural differences. Variable correspondence and lack of correspondence in different subgraphs may also provide evidence to select models, as was done for RDC data by Langmead and Donald [38].

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

Chapter 3

Contact Replacement for NMR Resonance Assignment: Random Graph Model for Contact Replacement

3.1 Introduction

As we discussed before, although the problem of backbone resonance assignment has been well-studied within the context of structure determination, the standard protocols are based merely on the restraints derived from NMR experiments, without any reference to the available structure of a target protein. Therefore, by transforming the backbone resonance assignment problem into the contact replacement problem. we actively utilize the spatial contact information from a protein's known structure, and look for a pattern match from it to the graph compiled from the NMR data.

Various versions of what we are calling here contact replacement have previously been studied. Our Hierarchical Grow-and-Match (HGM) algorithm [69] uses a branch-and-

[2]This work has been published [71] in the proceedings of 16th Annual International Conference Intelligent Systems for Molecular Biology (ISMB) 2008. It also appears in Bioinformatics, 24:i105 $-$ i213, 2008.

bound algorithm to find the complete ensemble of consistent correspondences between contact graphs and NMR graphs, and can handle significant noise and sparsity. However, due to the combinatorics of the problem and the branch-and-bound approach, HGM is effectively restricted to well-defined regions of secondary structure. The ST2NMR program [56] casts assignment given a 3D structure and NMR data as an optimization problem, and uses a Monte Carlo approach to find explanations of the data in terms of distances in the structure. While ST2NMR was shown to be effective for some test data, it requires very specific experimental set-ups and can provide no guarantees or insights into the information content of the data. We tested it on a number of different datasets, and found the accuracy to be fairly low and quite sensitive to the order of the input data [69]. PEP-MORPH [14] uses graph representations of the structure and data, but augments them with residual dipolar coupling data in order to compute matchings. Our earlier work on graph-based approaches to NMR assignment, Jigsaw [4] and random graph algorithms [3, 32], were able to effectively uncover secondary structure patterns; our random graph model enabled us to prove that the randomized methods have optimal performance in expected polynomial time. However, these approaches were all restricted to uncovering generic prototypes of secondary structure elements, rather than matching NMR data to an arbitrary three-dimensional structure.

Contribution. This work presents the first *efficient* algorithm to solve the contact replacement problem for *entire* proteins. We first show that by combining connectivity and type, and by exploiting the random structure of the noisy edges and vertex labels, one can provably determine unique matchings in polynomial time with high probability, even in the presence of significant noise, i.e., a constant number of noisy edges per vertex. Since the NMR interaction graphs we are studying have up to 5 times as many noise edges as correct ones, the ability to handle this degree of noise is important. This result significantly improves over previous results on finding long paths in noisy NMR graphs [3]. We then

detail a simple and efficient randomized algorithm that works very well in practice. To do so, we build upon our earlier work on random graph algorithms in NMR [3,32], which used connectivity information alone to uncover large, regular structures (α-helices and β-sheets) in NMR graphs. We now integrate connectivity information with amino acid type information (ambiguous labels on the vertices) in order to uncover large corresponding fragments in NMR and contact graphs for complete structures. We significantly extend our *reuse* paradigm to efficiently uncover these correspondences. Instead of backtracking upon finding an inconsistency in a growing correspondence, the reuse approach seeks to maintain the (mostly good) structure by applying local fix-up rules to address just the source of the inconsistency. Our empirical results show that this approach is quite effective in practice, relatively insensitive both to noise in the NMR graph and structural variation in the contact graph.

3.2 Our Approach

As defined in Section 2.2, input data are represented as graphs and the NMR resonance assignment problem has been converted into a matching between the contact graph G^* and the NMR interaction graph G. One feature of proteins particularly relevant here is that they are made of chains of amino acids. Thus the contact graph has an embedded Hamiltonian path from N terminus to C terminus (in addition to numerous through-space edges connecting residues at any sequential distance). Ignoring missing edges, the NMR graph has a corresponding Hamiltonian path. Our analysis and randomized algorithm both make use of this property, by focusing on finding the Hamiltonian path while "bringing along" the additional edges for scoring purposes.

We note that the contact replacement problem is NP-hard in general, since it contains as a special case the following NP-hard problem: Given a unweighted Hamiltonian graph (undirected or directed) H find a Hamiltonian path in H (i.e., assuming that there are no

constraints on vertex labels and all edge scores are the same). We note that the above problem remains NP-hard even when restricted to sparse Hamiltonian graphs, e.g., directed Hamiltonian graphs with maximum out-degree two [51] or undirected Hamiltonian graphs with degree at most three [18]. The problem has been shown hard to approximate in directed graphs: it is not possible to find paths even of superpolylogarithmic length in constant out-degree Hamiltonian graphs unless Satisfiability can be solved in subexponential time [7]. For undirected Hamiltonian graphs, the best known algorithms give longer paths (e.g., of length $n^{\Omega(1/\log\log n)}$) in Hamiltonian graphs in polynomial time [15–17]. We note that the above algorithmic results do not apply to our problem because we have additional information (amino acid classes) for the vertices. More importantly, NMR interaction graphs are not arbitrary graphs and indeed have a special structure as captured by the random graph model described next.

3.3 Random Graph Model for NMR Interaction Graph

In order to develop and analyze effective algorithms, we must consider and model the nature of the relationship between the ideal contact graph G^* and the observed NMR interaction graph G. We note that traditional $G(n,p)$ random graph models [8] essentially add noise edges randomly and independently. However, the noisy edges in an NMR interaction graph are not arbitrarily distributed. Instead, *chemical shift degeneracy* is the key source of noise in these graphs, imposing a particular correlation structure among noise edges. We have developed a random graph model that properly captures the noise in NMR interaction graphs [3, 32].

Definition 1 ($M(G^*, w)$ random graph) *The model $M(G^*, w)$ "generates" a random graph from the (correct) graph G^*, where w is a parameter that determines the number of noisy edges generated per correct edge. Let π be a random permutation of V^*. Denote by $\pi(v)$*

the index of $v \in V^*$ in the permutation. We then consider as ambiguous all vertices within a "window" of size w around a particular vertex. For each edge of G^*, additional edges are generated as follows. Consider an edge $(u^*, v^*) \in E^*$. Then for each u in the window of width w around $\pi(u^*)$ (i.e., $|\pi(u^*) - \pi(u)| \leq w$), we add the edge (u, v^*) to the random graph.

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

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

3.4 Theoretical Analysis and Implications

We present a theoretical analysis to show that the contact replacement problem can be solved with high probability in polynomial time. For the analysis, we assume that the NMR interaction graph $G = M(G^*, w)$ is generated from the correct contact graph $G^* = (V^*, E^*)$ which is a Hamiltonian path of length n (= number of amino acids). For now, we assume no edge weights, no breaks, and no other sources of noise (additions and deletions); the result can be generalized. We also assume that for each vertex, the true amino acid type is in the amino acid class labeling the vertex, and that it is a non-trivial class, i.e., it is not A itself (refer to the model above). The contact replacement problem now reduces to finding whether there is a Hamiltonian path in G that is *equivalent* (defined below) to that of the Hamiltonian path G^*.

Definition 2 (equivalence) *We say that a subgraph $H = (V_1, E_1)$ of G is equivalent to a subgraph (i.e., subpath) $H^* = (V_2, E_2)$ of G^*, denoted as $H \equiv H^*$, if and only if there is a bijection $m : V_1 \rightarrow V_2$ such that for every $v_1 \in V_1$, $a(m(v_1)) \in \ell(v_1)$ and there is an edge $e = (u_1, v_1) \in E_1$ iff there is an edge $(m(u_1), m(v_1)) \in E_2$.*

In the following, "with high probability (whp)" means with probability at least $1 - 1/n^{\Omega(1)}$, where n is the number of amino acids in the protein.

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

Proof Without loss of generality, we will assume that $|A| = 2$ (A is the set of amino acid types). The proof can be easily made to work without this restriction. The proof hinges on the following claim.

Claim: Fix a (sub)path P of length $k = c \log n$ in G^*, where $c > 0$ is a constant (fixed in the proof). Then the following hold whp.

(a) There is a unique subgraph H of G, such that $H \equiv P$, i.e., there is no other subgraph H' of G such that $H' \equiv P$.

(b) There is no other subgraph Q of G^* such that $P \equiv Q$.

We will first show (a).

Since $G = M(G^*, w)$ (i.e., generated by our random graph model), we know that there exists a subgraph H of G such that $H \equiv P$. We now show that H is unique. Let the two amino acids be a and b, with probabilities of occurring p_a and $1 - p_a$ respectively. Let $q = \max\{p_a, 1 - p_a\}$. By our assumption on the size of A, q is a constant (< 1). The path G^* induces a natural ordering of vertices of G. We bound the probability of finding another path (subgraph) H' of G that is equivalent to P by the following expression:

$$\Pr\{\exists H' \equiv P\} \leq \sum_{k'=0}^{k-1} n \binom{n}{k'} (\frac{w}{n})^{k'} q^{k-k'}$$

The reasoning is as follows. Let k' be the number of noisy edges in H'; k' can vary between 0 and $k-1$, and hence we sum over all possibilities. The first term is the number of different ways of fixing the starting vertex. There are at most $\binom{n}{k'}$ ways of choosing vertices from which the noisy edges emanate. The third term bounds the probability that the noisy edges form a path between them *with* the amino acid labels matching those of the corresponding vertices in P. The last term is the probability that the amino acid labels for the correct vertices match. We can bound the sum as follows (note that we take $\infty^0 = 1$):

$$\begin{aligned} \Pr\{\exists H' \equiv P\} &\leq \sum_{k'=0}^{k-1} n (\frac{en}{k'})^{k'} (\frac{w}{n})^{k'} q^{k-k'} \\ &\leq \sum_{k'=0}^{k-1} n (\frac{ew}{k'})^{k'} q^{k-k'} \end{aligned}$$

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

$$\begin{aligned}
\Pr\{\exists H' \equiv P\} &\leq n \sum_{k'=0}^{c \log n} (\frac{ew}{k'})^{k'} q^{(c \log n)-k'} \\
&\leq n \sum_{k'=0}^{c \log n} (\frac{ew}{qk'})^{k'} q^{c \log n} \\
&\leq n \sum_{k'=0}^{c \log n} (\frac{ew}{qk'})^{k'} O(n^{c \log q}) \\
&\leq n \sum_{k'=0}^{c \log n} O(1)\, O(1/n^{c \log(1/q)}) \\
&\leq n \sum_{k'=0}^{c \log n} O(1/n^3) \\
&= O(1/n)
\end{aligned}$$

if c is a sufficiently large constant.

We now show Claim (b). We bound the probability that there is some subgraph Q of G^* such that $P \equiv Q$:

$$\Pr\{\exists Q \equiv P\} \leq nq^k.$$

The first term in the bound is the number of different ways of fixing the starting vertex of Q and the second term bounds the probability that a particular Q is identical to P. If $k = c \log n$, for a sufficiently large constant c, the above probability is bounded by $1/n$.

Using the above claims we can design the following polynomial-time algorithm. The algorithm finds a subgraph H in G of length $c \log n$ (where c is fixed in the above claim) such that it is equivalent to some subgraph P of G^*. Once such a subgraph is found, it will be a unique match in G^* whp (by the above claim). The algorithm then repeats this process until the full equivalent mapping is found. The subgraph H can be found by an exhaustive search, starting at some vertex and examining all possible paths of length $c \log n$. There are only at most $w^{c \log n} = O(n^{c \log w})$ (i.e., a polynomial number) of possible paths and hence

the search can be done in polynomial time. One of these paths in G will be a unique match with a corresponding path in G^*. ■

We note that the above theorem can be extended to the case when there are missing (correct) edges in the NMR graph G' as shown in the following corollary.

Corollary 1 *Suppose G contains a path P' of length at least $c \log n$ (where c is as fixed in the above theorem) that is equivalent to a subgraph H^* of G^*. Then P' can be found and matched correctly with H^* whp.*

The above analysis shows that the contact replacement problem can be solved in polynomial time if $w = O(1)$, i.e., there is at most a constant number of noisy edges per vertex. This is significant for two reasons. First, in practice, typically the number of noisy edges per vertex is a constant (around 5). Second, if there is no amino acid information, the randomized algorithm of [3,49] can find long paths (of length at least $\Omega(n/\log n)$) in polynomial time only if the number of noisy edges per vertex is *at most one*. Our analysis here shows that *this threshold barrier can be surmounted* by using amino acid type information. Our experimental results validate this theoretical prediction.

3.5 Methods

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

path result from the previous section, while by including all edges in the score, we take advantage of all available information to better control the search.

A key insight of our algorithm is that *in searching for good matchings, the best ones tend to share a lot of substructure*. (Our results below on assignment ambiguity, Figure 3.4, illustrate.) In branching-based searches, such shared substructure can appear on many different branches, making exhaustive search very inefficient and causing backtracking to perform wasteful undoing and redoing. In contrast, we use more efficient local fixes to resolve inconsistencies and continue searching with most of the structure still intact.

Figure 3.1 gives the pseudocode for our algorithm. The algorithm maintains (and fixes up) a single set F of fragments, with a mapping m to the contact graph that is always consistent (i.e., fragments do not overlap). Some fragments may not be mapped, meaning that under the current matching, they are considered noise. On each iteration, the algorithm sequentially extends one fragment, adding an NMR vertex that will correspond to the next residue position in the sequence. Several things could happen upon growing to that vertex; see Figure 3.2. In the simplest case, the algorithm picks up an unmatched NMR vertex (and its fragment) and simply extends the matching. However, it may run into a conflict and need to fix up the current matching. If a fragment wants to grow to a vertex in the middle of another fragment, then the other fragment is split at the point of conflict to allow its suffix to be taken away. If the growth results in a mismatch of amino acid type or of alignment, then a realignment is attempted. Matching the fragment somewhere else in the contact graph may result in a consistent matching, or may produce another conflict, potentially fixed by replacing part of the conflicting fragment with the new fragment. To keep each step simple enough, we only recursively handle the conflict at this point if it's simple enough to fix. The algorithm repeats until convergence. In practice, we run a fixed number of iterations, and keep track of m through the iterations in order to analyze the distribution of good solutions.

$m \leftarrow \emptyset$ *// matching, from G to G**
$F \leftarrow \{\{v\}$ s.t. v $\in$ V$\}$ *// each v starts in its own fragment*
Repeat until convergence:
  Choose at random a vertex $u \in V$ with no successor in F
  Choose an edge $(u, v) \in E$, for some v, with probability according to score
  // Try to grow from u to v at current alignment
  Let f_u and f_v be the fragments in F containing u and v
  $f_u \leftarrow f_u +$ suffix of f_v starting at v
  $f_v \leftarrow$ prefix of f_v before v
  If f_v is empty, remove it from F
  If $m(v)$ is defined and $m(u)$ is undefined and f_u can be aligned ending at $m(v) - 1$
    Update m to align f_u, i.e., $m(u) \leftarrow m(v) - 1$, etc.
  Else if $m(u)$ is defined and $m(v)$ is undefined and f_v can be aligned starting at $m(u) + 1$
    Update m to align f_v, i.e., $m(v) \leftarrow m(u) + 1$, etc.
  Else
    // Try to realign
    Let $f = \{p_1, ..., p_n\}$ be the fragment with u and v
    Choose an alignment $f' = [i, i + n]$ starting from position i, with probability according to score
    If any portion of f' already has some other fragment aligned there
      Choose to splice that out or to keep it, with probability according to score
    Update m
    Recursively handle spliced-out subfragments, if they are large enough and can be aligned

Figure 3.1: Randomized algorithm for contact replacement: given a contact graph $G^* = (V^*, E^*)$ and NMR graph $G = (V, E)$, determine the matching m.

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

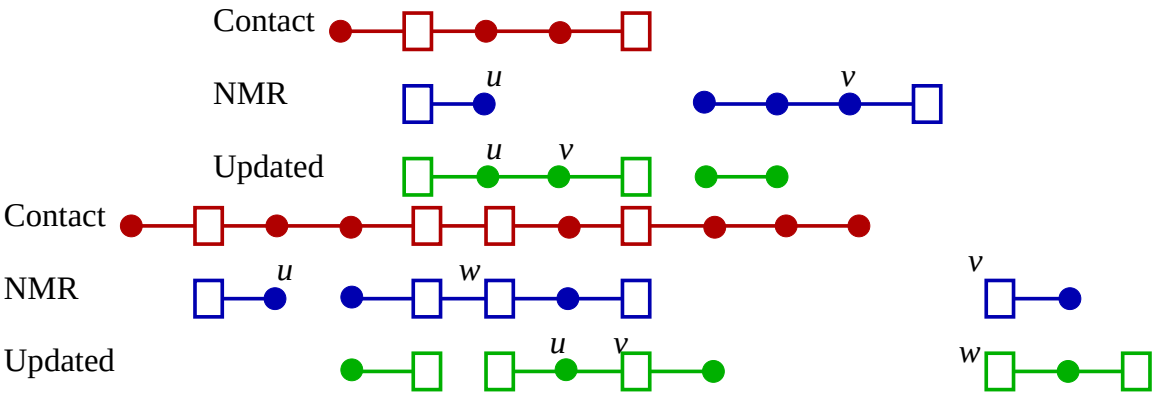


Figure 3.2: Reuse-based growing and aligning. Contact graph and NMR residues in the same column are matched. There are two amino acid types (empty squares and filled circles), which must match. (left) Growing from a matched fragment ending in u to an unmatched fragment with v in the middle leaves behind the prefix of the unmatched fragment in order to append and match the suffix following u. (right) Growing from u to v requires a realignment of the joined fragment. The joined fragment displaces the suffix starting at w of another fragment.

we consider the total of all edges before vs. after the possible change.

3.6 Results

Table 3.1 summarizes the datasets, both experimental and synthetic, that we used to validate our algorithm. The proteins are of moderate size for typical NMR studies, and this collection has representative structural diversity and assignment difficulty. We used three experimental datasets from previous contact-based assignment work [32, 69], including Human Glutaredoxin (PDB ID: 1JHB), Core Binding Factor β (PDB ID: 2JHB), and the catalytic domain of GCN5 histone acetyltranferase (PDB ID: 5GCN). For brevity, and since assignment is based on structure, we refer to each protein by its PDB ID. The noise rate (average number of noisy NMR edges per contact edge) is as high as 5.4 (1JHB α-helices) and the missing rate as high as 51.8% (5GCN loops). Since such complete experimental datasets are a rare commodity, in order to more broadly test our approach, we also used a set of previously-generated synthetic datasets [69] based on chemical shift data deposited in

the BMRB. These synthetic datasets include noise edges according to Gaussian noise with variance 0.02 (corresponding to a standard 0.05 ^{1}H match tolerance) and missing edges according to observed statistics correlating the missing probability with the interatomic distance [12]: $d \leq 3$ Å, missing 21%; $3 < d \leq 4$ Å, missing 41%.

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

Figure 3.3 illustrates some examples of the convergence of the algorithm; other runs and other datasets had similar behavior. In general, the score increases rapidly over the initial iterations (a few hundred steps). During this phase, pseudoresidues are being organized into various "short" paths aligned to the primary sequence, naturally increasing the score. With successive iterations, the short paths will start to grow into each other and conflicts occur, requiring fix-up moves to remove the conflicts. While moves are made so as to prefer increased score, locally bad moves are occasionally made in order to escape local optima. In many cases, the score converges to a value near that of the reference solution. As we will see below, the variation tends to produce only minor ambiguity in the resulting correspondence, and over the ensemble of solutions the correct assignments tend to be found.

Table 3.1: Datasets (top 3 experimental; bottom 9 synthetic)

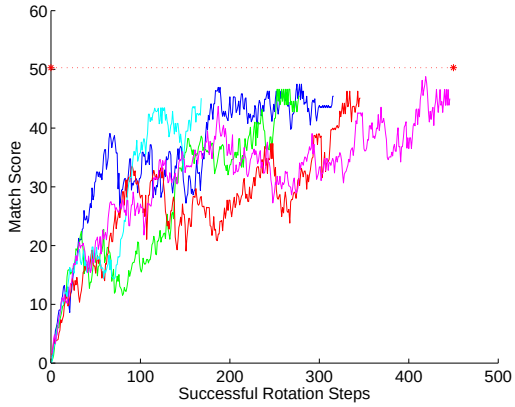
PDB ID	BMRB Entry	$\alpha/\beta/loop$ # elements	# residues	# edges	noise ($\times$)	missing (%)	RMSD (Å)
1JHB	N/A	5/4/10	43/18/44	160/49/81	5.4/2.5/3.0	32.5/38.7/33.3	1.3/0.8/1.6
2JHB	4092	5/6/11	36/42/64	138/99/141	3.5/5.2/3.7	33.3/18.2/41.1	1.5/0.9/2.6
5GCN	4321	4/7/12	56/52/58	245/115/110	4.9/4.6/2.2	32.7/28.7/51.8	1.5/1.6/3.5
1KA5	2030	3/4/8	40/23/25	162/56/58	3.2/2.8/1.9	21, 41	0.8/0.7/0.8
1EGO	2152	3/4/8	39/19/27	165/42/49	2.2/2.7/2.6	21, 41	2.1/1.4/3.6
2FB7	7084	– /5/6	– /32/63	– /74/96	– /3.0/2.4	21, 41	– /1.5/7.7
1G6J	5387	2/5/8	18/22/36	75/47/71	1.4/3.1/3.0	21, 41	1.0/1.1/2.3
1P4W	5615	5/ – /6	66/ – /33	253/ – /29	3.8/ – /2.7	21, 41	1.2/ – /3.5
1SGO	6052	4/6/9	47/26/64	199/68/131	2.9/3.4/6.3	21, 41	2.8/1.3/9.6
1RYJ	5106	1/5/7	9/27/37	31/55/51	1.0/4.3/3.4	21, 41	1.3/1.4/2.6
2NBT	1675	– /3/4	– /16/50	– /36/108	– /1.0/2.9	21, 41	– /1.5/4.5
1YYC	6515	2/9/11	36/72/66	149/165/153	1.2/4.7/2.6	21, 41	2.0/1.7/6.2

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

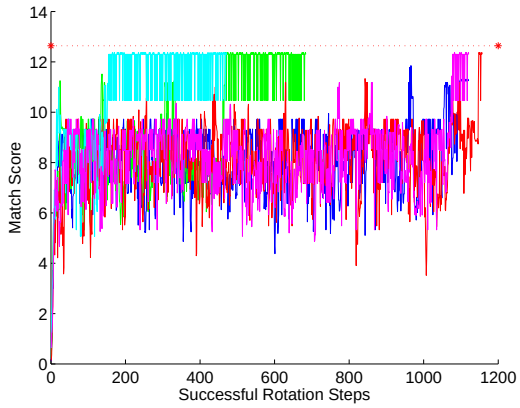
Figure 3.4 illustrates the assignment results for the experimental datasets. Notice that we can assign the *whole* protein, and that for most of the positions, the reference assignments are included in the top-ranked solutions. Exceptions tend to be from areas with many missing edges (e.g. 1JHB 51–57) or residues close to a Proline (e.g. 5GCN 34–35), which necessarily induces a break. The results also show that the high-scoring solutions tend largely to agree. For 1JHB, there are on average 1.7 matches for each residue in α-helices, and 1.2 in β-sheets and loops. For 2JHB the ambiguity level is 1.3 for α-helices, 2.5 for β-sheets, and 2.4 for loops, and for 5GCN we have 1.3, 2.6, and 3. (These numbers can be compared to the expected number of matches *a priori*, which is simply the number of residues in the protein within the same ambiguous amino acid class, anywhere from 2 to 14.) In general, β-sheets and loops are more ambiguous than α-helices since their tertiary structures generate fewer edge constraints. For the nine synthetic datasets, the average ambiguity is as low as 1 for α-helices (1G6J), β-sheets (1KA5) and loops (1EGO); with a maximum of 2.8 (1SGO), 3.6 (1YYC), 9.1 (1SGO), and median of 1.7 (1KA5), 1.5 (1G6J), 2.1 (1G6J) for the three types, respectively. The most ambiguous case is 1SGO loops since it has both the highest noise ratio (6.3) and the largest RMSD (9.6 Å).

We compared these results to the corresponding ones of [3] (limited to α-helices), and found that our algorithm performs much better. Considering each position separately, we can evaluate how frequently the majority of the solution ensemble identifies the correct match. In our results, that is true for 90% of the positions, where it holds for less than 70% of the positions under the earlier method.

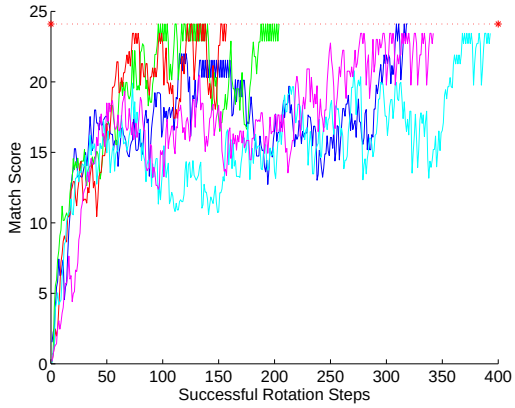
For both the experimental datasets and the synthetic ones, we studied the sensitivity of our algorithm to structural variation. For each dataset, an ensemble of NMR-determined structures had been deposited. We generated a contact graph for each different member of the ensemble, and studied how well the original data could be assigned under the varying structures. The average RMSDs of the ensemble members (all to the reference model) are



1JHB α-helices

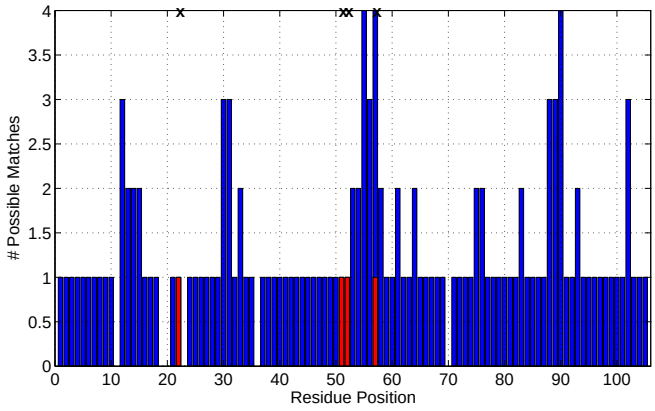


1JHB β-sheets

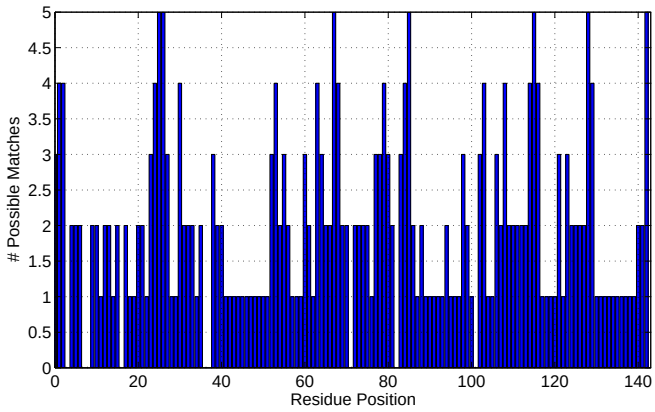


1JHB loops

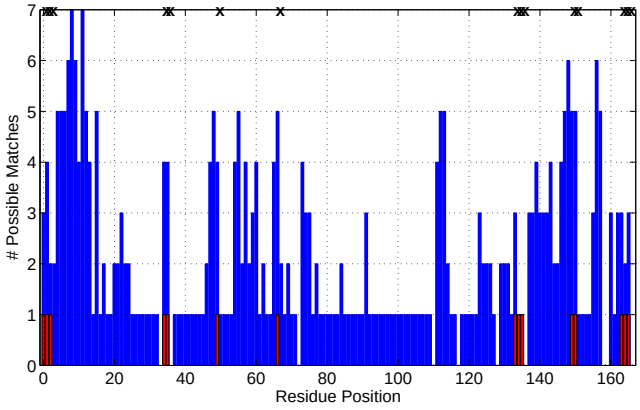
Figure 3.3: Score convergence over $10,000$ iterations for 5 individual test runs for 1JHB. Here a "successful" step indicates that a move has been accepted (a partial move can be rejected during fix-up). The dashed horizontal line at the top indicates the score of the reference assignment.



1JHB



2JHB



5GCN

Figure 3.4: Assignment ambiguity for experimental data sets. The bars indicate how many pseudoresidues can be mapped to each residue in the top 10 solutions. The red bars (also marked by 'X's at the top) indicate positions for which the reference assignment was not present in any solution.

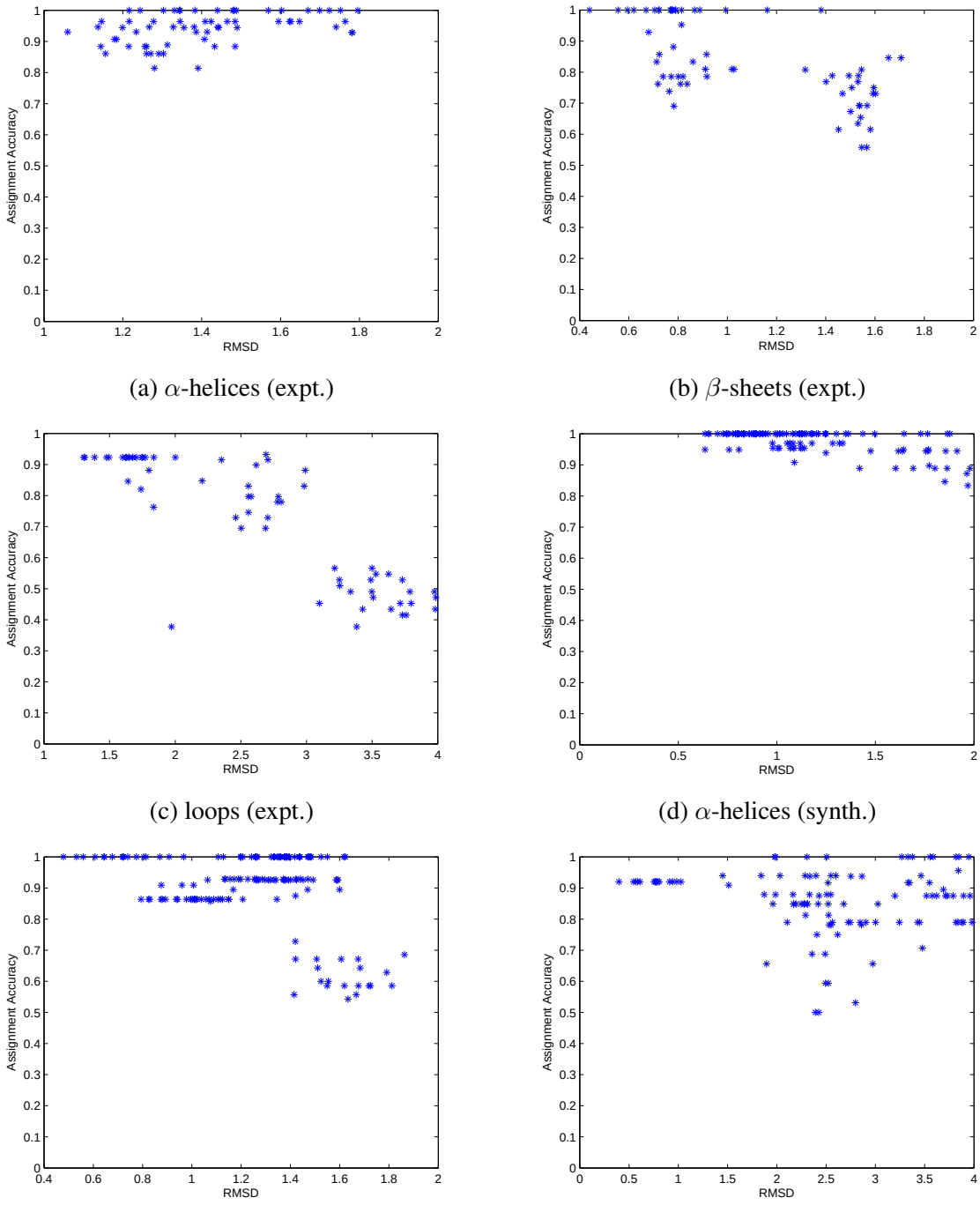
given as the far right column in Table 3.1, and are representative of the extent of structural uncertainty one might expect when assigning NMR data using an x-ray structure or high-quality homology model.

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

Figure 3.6 summarizes the performance of our algorithm for each dataset under the different structure models. These results suggest that, overall, we achieve good accuracy in assignment, above 80% for α-helices, 70% for β-sheets and 60% for loops. Since contacts are discrete, one might initially expect more effects from structural variation. However, recall that our method focuses on matching paths and uses non-sequential edges for scoring. While the score degrades with the loss of non-sequential edges, path connectivity is fairly well maintained regardless of the three-dimensional coordinates.

3.7 Discussion

NMR spectroscopy provides scientists with the ability to collect detailed information regarding protein dynamics and interactions in solution. However, in order to interpret the dynamics and interaction experiments, it is necessary to first obtain a resonance assignment so that the observed spectral peaks may be matched to atoms in the protein (e.g., to localize which atoms are affected by binding). In order to increase the throughput and decrease the expense of performing resonance assignment, this paper develops a new ap-



(a) α-helices (expt.)

(b) β-sheets (expt.)

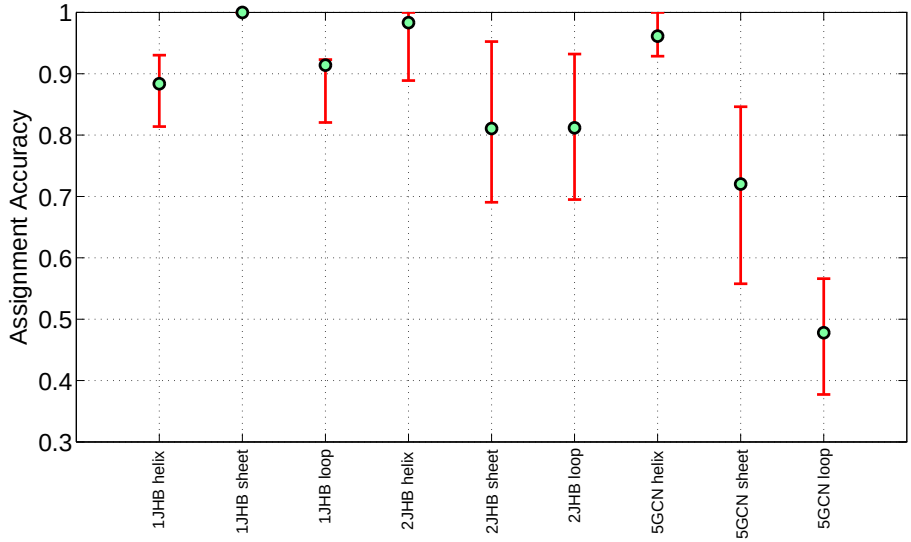
(c) loops (expt.)

(d) α-helices (synth.)

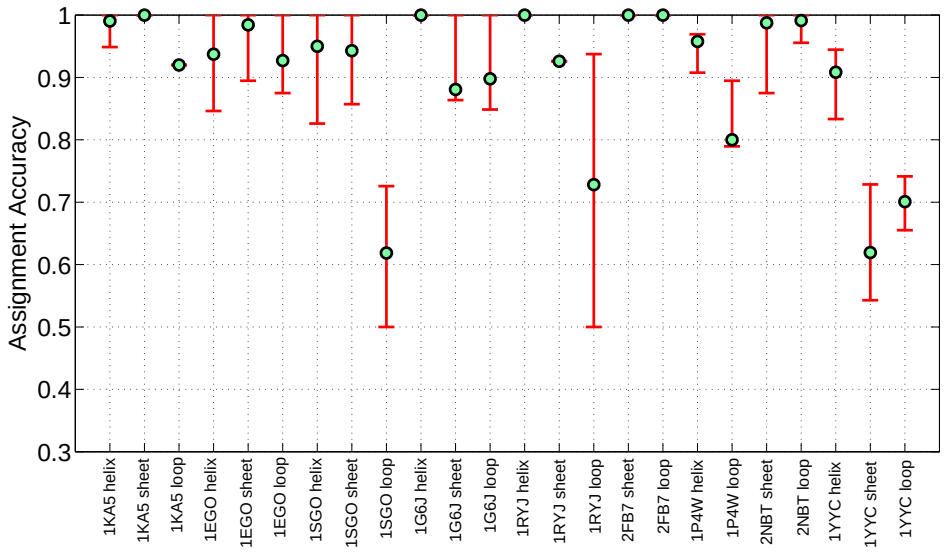
(e) β-sheets (synth.)

(f) loops (synth.)

Figure 3.5: Performance of our algorithm with varying structure, measured in terms of Root Mean Square Distance (RMSD) to the reference model. Each blue asterisk indicates the accuracy of one member of the structural ensemble for one dataset. We only show results for structures with at most 2 Å RMSD for α-helices and β-sheets and 4 Å for loops.



(a) experimental datasets



(b) synthetic datasets

Figure 3.6: Overall performance of our algorithm. Cyan circles indicate average assignment accuracy over all members of an ensemble for a dataset, while bars indicate the best and worst assignments.

proach, *contact replacement*. Contact replacement exploits information from an available three-dimensional structure (from x-ray crystallography or homology modeling) to drive the assignment process, replacing the typical more extensive and expensive set of experiments with a minimalist set. Once contact replacement has been performed, the available assignments can be used to interpret dynamics or perturbation experiments. We note that those are separate experiments not included in the assignment process, and it is an interesting question (regardless of the assignment approach) to propagate uncertainty from assignment to uncertainty in dynamics or interactions.

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

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

It is interesting to consider the relationship between contact replacement and nuclear vector replacement (NVR) [37, 39], both of which use an available structure to perform NMR resonance assignment, but based primarily on different data. (NVR does use some

NOESY data, too, but only unambiguously assignable peaks.) At a high level, the residual dipolar coupling data used in NVR is global, giving orientations of bond vectors wrt a co-ordinate frame, whereas the NOESY data used here is local, giving distances only between close protons. A natural avenue of work is to study the relative information content of these types of information in order to develop a unified framework incorporating both.

Compared to other graph-based structure matching problems (e.g., threading, structural alignment, structure motif finding, chemical compound querying, etc.), contact replacement has no sequential order information for one of the graphs (the NMR one). However, the basic insights behind our algorithm (namely reusing partial solutions by making local fix-ups) may still be quite relevant in developing new algorithms for those applications. Alternatively, giving up sequential order in those applications may result in finding more distant relationships.

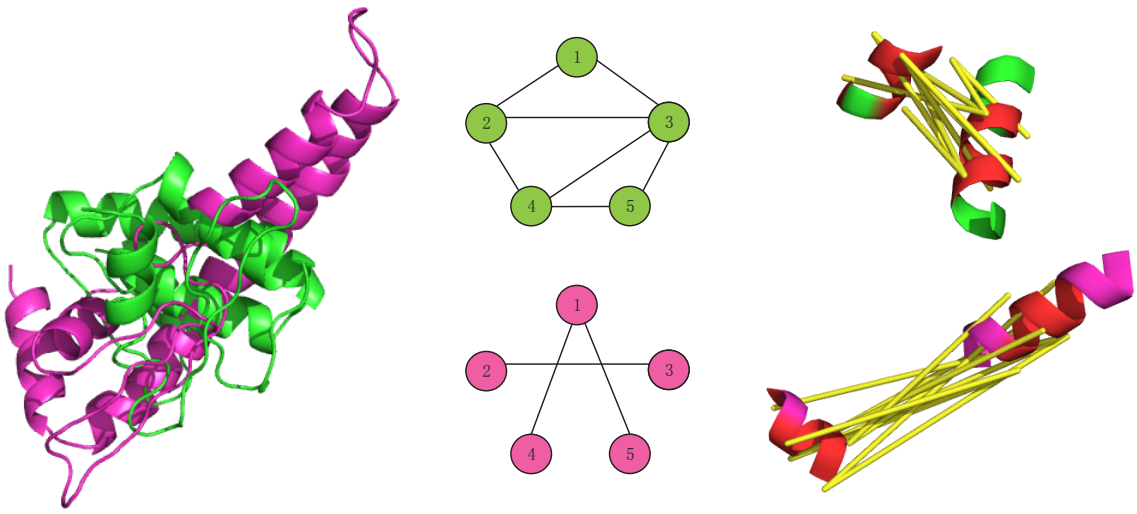
Chapter 4

Protein Fold Determination: Topological Fingerprint-based Cross-link Analysis and Experiment Planning

4.1 Introduction

Despite significant efforts in structural genomics, the vast majority ($> 90\%$ [19]) of available protein sequences do not have experimentally determined three-dimensional structures, due to experimental expense and limitations (e.g., lack of crystallizability). At the same time, since structure is more conserved than sequence, there may be only a small number (a thousand or two [20,47]) of distinct natural "folds" (overall structural organizations), and many of them can already be found in the protein databank (PDB). Fold recognition techniques [19, 74] take advantage of this, and have become increasingly effective at identifying the fold of a given target sequence. However, the series of Critical Assessment of Structure Prediction (CASP) [45] contests demonstrates that, in the absence of sufficient

[3]This work will be published [70] in the proceedings of 9th International Workshop on Data Mining in Bioinformatics (BIOKDD '10).



(a) (b) (c)

Figure 4.1: Protein fold determination by disulfide cross-linking. The example shows two models, but the method readily handles tens or even hundreds of models. (a) Two models, TS125_3 (green) and TS194_2 (magenta), for CASP target T0351, are of reasonable quality but have rather different topologies. (b) The three-dimensional structures are compiled into graphs on the secondary structure elements (SSEs), representing the topology in terms of contacting SSE pairs. A topological fingerprint is selected based on differences in SSE contacts (e.g., 1-2, 2-4, 3-5, etc.) that together distinguish the models. (c) For each SSE pair in the topological fingerprint, a set of residue pairs is selected for disulfide cross-linking, in order to robustly determine whether or not the SSE pair is actually in contact. The figure shows the selected cross-links (yellow) to test for SSE pair (1, 2). Residues selected for cross-linking are colored red.

sequence identity, it remains difficult for fold recognition methods to always select the correct model. While a native-like model is often among a pool of highly ranked models, it is not necessarily the highest-ranked one, and the model rankings depend sensitively on the scoring function used [45, 82]. Figure 4.1(a) illustrates two possible alternative models for one target from a recent CASP competition.

Alternatively, some structure biologists also use the actual 3D distance cutoffs of the predicted model deviating from the target template to evaluate model quality. Global Distance Test (GDT) [82] is one of the popularly adopted method. It identifies in the prediction the sets of residues deviating from the target by not more than specified CA Distance cutoff using many different superpositions. Nevertheless, it is more about the detailed 3D coordi-

nate comparison of individual residues while doesn't obtain the high-level structure information revealed by topological analysis. For example, it is poor to detect un-protein-like features such as knots or slip-knots [34] which occur more frequently in models generated by protein structure prediction methods. Another widely used metric is the Root-mean-square deviation (RMSD) for the subset of CA atoms from the predicted models. However, it again measures the direct 3D structural difference while doesn't have enough power to present the model's topological abstraction.

Seeking to close the gap between computational structure prediction and experimental structural determination, we [78,79] and others [10,21,80] have developed methods (which we call *structure elucidation*) to select structural models based on relatively rapid biochemical/biophysical experiments. One type of experiment particularly suitable for this purpose is *cross-linking*, which essentially provides distance restraints between specific pairs of residues, based on the formation (or not) of chemical cross-links. While residue-specific (e.g., lysine-specific) cross-linking has been effectively used for this task [25, 35, 80], we previously showed that planned *disulfide* cross-linking has a number of advantages, in terms of the ease and reliability of experiment and the quality of the resulting information content [79]. In disulfide cross-linking (or "trapping") [9, 28, 36], a pair of cysteine substitutions is made and the formation of a disulfide bond after oxidation is evaluated, e.g., by alteration in electrophoretic mobility [9, 36, 79]. An important point for our purposes here is that disulfide cross-links are *plannable*—we control exactly which pair of residues is probed in a particular experiment.

While earlier methods have focused on probing geometry and selecting a model, we target here a more defined characterization of protein structure, ascertaining the overall protein fold. We call this approach *fold determination*, named in contrast to purely computational *fold recognition* and our less defined structure elucidation approach. We first characterize the topological / fold-level differences in a set of models in terms of contact patterns of

secondary structure elements (SSEs); see Figure 4.1(b). The topological representation allows for a robust experimental characterization of the structure, less sensitive to noise and uncertainty in both the models (e.g., threading misalignment) and the actual structure (e.g., flexibility). As a representation with fewer degrees of freedom than the complete threading models, the topological representation also enables us to explicitly consider all possibilities and handle the case when none of the models is correct. Once we have identified a subset of SSE pairs that are most informative for fold determination, we plan disulfide cross-links to evaluate these SSE pairs; see Figure 4.1(c). By specifically planning for each such SSE pair, we can account for the dependence among the cross-links and select a set that will be robust to, and even help characterize, model misalignment and protein flexibility.

The method presented here strikes a balance between very limited cross-linking (e.g., six disulfide pairs in our earlier work [79]) and testing all residue pairs. We assume that robotic genetic manipulation methods (e.g., based on SPLISO [58] and RoboMix [2]) can construct a combinatorial set of dicysteine mutants, but that we still should test a much smaller set than all residue pairs. (Our plans require tens to around a hundred cross-links, depending on error requirements.) Thus we must optimize a plan so as to maximize information gain while minimizing experimental complexity. This is analogous to feature subset selection, where the goal is to choose a subset of features from a dataset such that the reduced set still keeps the most "distinguishing" characteristics of the original [23, 41]. At the topological level (Figure 4.1(b)) the features are SSE pairs, and the objective is to select those that will correctly classify the real structure to a model. At the cross-link level (Figure 4.1(c)) the features are potential disulfide pairs and the objective is to select those that will correctly classify contact/not for the SSE pair. For each level, we optimize a plan by employing an information-theoretic planning algorithm derived from the minimum redundancy maximum relevance approach [50]. We then evaluate a plan with a Bayes error framework that characterizes the probability of making a correct decision from the experi-

mental data.

4.2 Methods

We are given a set M of models. They may be redundant (i.e., some may have the same fold), and they may be incomplete (i.e., a representative of the correct fold may not be included). Our goal is to plan a set of disulfide cross-linking experiments (i.e., identify residue pairs to be individually tested) in order to select among them. As discussed in the introduction, we do this in two stages (Figure 4.1(b) and (c)), first selecting a "topological fingerprint" of SSE pairs to distinguish the folds, and then selecting cross-links to assess these SSE pairs.

4.2.1 Topological fingerprint selection

In order to compare SSE topologies, we need a common set of SSEs across the models. Since secondary structure prediction techniques are fairly stable [29, 31], it is generally the case that models have more-or-less the same set of SSEs, covering more-or-less the same residues ($> 50\%$ overlapping as observed in our test data). Our approach starts with a set S of SSEs that are common to at least a specified fraction (default 50%) of the given models. For example, both models in Figure 4.1 have 5 α-helices, as do 63 other models for the same target. The later cross-link planning stage will account for the fact that the common SSEs may in fact extend over slightly different residues in the different models.

Given the SSE identities, we form for each model $m_i \in M$ an *SSE contact graph* $G_{SSE,i} = (S, C_i)$ in which the nodes S are the SSEs (common to the specified fraction of models, as described in the preceding paragraph) and the edges $C_i \subset S \times S$ are between contacting SSEs (specific to each model). We determine SSE contacts from residue contacts, deeming an SSE pair to be in contact if a sufficient set of residues are. Our current

implementation requires at least 5 contacts (at < 9 Å C^{β}-C^{β} distance), and at least 20% of each SSE's residues to have a contact partner in the other SSE.

Our goal then is to find a minimum subset $F \subset S \times S$ of SSE pairs providing the maximum information content to differentiate the models. As discussed in the introduction, this is much like feature subset selection; in particular, the *max-dependency* feature selection problem seeks to find a set of features with the largest dependency (in term of mutual information) on the target class (here, the predicted structural model) [50]. While max-dependency leads to the minimum classification error, there is unfortunately a combinatorial explosion in the number of possible feature subsets that must be considered. To deal with the combinatorial explosion, we develop here an approach based on the minimum Redundancy Maximum Relevance (mRMR) method [50].

Probabilistic model

First we develop a probabilistic model in order to evaluate the information content in a possible experiment plan. Let us treat each edge as being a binary random variable c representing whether or not the SSE pair is in contact, with $\Pr(c)$ the probability of being in contact ($c = 1$) or not ($c = 0$). We estimate $\Pr(c)$ by counting occurrence frequencies over the contact edge sets C_i for the models:

$$\Pr(c = x) = \frac{\sum_y q(c, x, y) \cdot |\{C_i : y = \mathbf{1}_{C_i}(c)\}|}{\sum_z \sum_y q(c, z, y) \cdot |\{C_i : y = \mathbf{1}_{C_i}(c)\}|} , \tag{4.1}$$

where the summed variables range over $\{0, 1\}$ and the indicator function $\mathbf{1}$ tests for membership of c in set C_i, and thus the set includes those SSE contact graphs for which the contact state of c agrees with y. To allow for noise, when evaluating $x = 1$ we include a contribution from $y = 0$ (false negative) along with that for $y = 1$ (true positive), and similarly when evaluating $x = 0$ we consider both $y = 1$ (false positive) and $y = 0$ (true negative). The q function weights the contributions for the agreeing and disagreeing case.

We currently employ a uniform weighting independent of edge, since we observed in cross-link planning (below) that the expected error rate in evaluating any SSE contact was well below 10% when using a reasonable number of cross-links.

$$q(c, x, y) = \begin{cases} 0.9 & x = y; \\ 0.1 & x \neq y. \end{cases} \tag{4.2}$$

The approach readily extends to be less conservative and to allow different weights for different SSE pairs, e.g., according to cross-link planning (discussed in the next section).

We can likewise compute a joint probability $\Pr(c, c')$ from co-occurrence frequencies:

$$\Pr(c = x, c' = x') = \frac{\sum_{y,y'} q(c, x, y) \cdot q(c', x', y') \cdot |\{C_i : y = \mathbf{1}_{C_i}(c), y' = \mathbf{1}_{C_i}(c')\}|}{\sum_{z,z'} \sum_{y,y'} q(c, z, y) \cdot q(c', z', y') \cdot |\{C_i : y = \mathbf{1}_{C_i}(c), y' = \mathbf{1}_{C_i}(c')\}|} \tag{4.3}$$

where again the sums are over $\{0, 1\}$ and the indicator function is as described above.

Then we can evaluate the *relevance* of each SSE contact edge c in terms of its entropy $H(c)$; a high-entropy edge will help differentiate models while a low-entropy one won't. We can also evaluate the *redundancy* of a pair (c, c') of edges in terms of their mutual information $I(c, c')$; a high mutual-information pair contains redundant information.

$$H(c) = -\sum_{x} \Pr(c = x) \log \Pr(c = x) \tag{4.4}$$

$$I(c, c') = \sum_{x} \sum_{x'} \Pr(c = x, c' = x') \log \frac{\Pr(c = x, c' = x')}{\Pr(c = x) \Pr(c' = x')} \tag{4.5}$$

Experiment planning

The mRMR approach seeks to minimize the total mutual information (redundancy) and maximize the total entropy (relevance). In this chapter, we define the objective function as

the difference of the two terms.

$$s(F) = \frac{1}{|F|} \sum_{c \in F} H(c) - \frac{1}{|F|^2} \sum_{c,c' \in F} I(c, c') \tag{4.6}$$

To optimize this objective function, we employ a first-order incremental search [50], which builds up a set F starting from the empty set and at each step adding to the current F the edge c_* that maximizes

$$c_* = \arg \max_{c \in (S \times S) \setminus F} \Big(H(c) - \frac{1}{|F|} \sum_{c' \in F} I(c, c') \Big) \tag{4.7}$$

The search algorithm stops when the score for c_* drops below a threshold (we use 0.01 for the results shown below).

The original mRMR formulation with first-order incremental search was proved to be equivalent to max-dependency (i.e., to provide the most information about the target classification) [50]. The proof carries over to our version upon substituting our formulations of redundancy and relevance (discrete, with choices of SSE pairs providing information about models) in place of the original ones (continuous, with gene profiles representing different types of cancer or lymphoma). Essentially, it can be proved that the optimal max-dependency value is achieved when each feature variable is maximally dependent on the class of samples, while the pairwise dependency of the variables is minimized. Furthermore, this objective can be obtained by pursuing the mRMR criterion in the "first-order" incremental search (i.e., greedy) where one feature is selected at a time. Therefore we don't need to explicitly compute the complicated multivariate joint probability, but can instead compute just the pair-wise joint probabilities. We thus have an efficient algorithm for finding an optimal set of SSE pairs to differentiate models.

Data interpretation

In the next section, we will describe the planning of disulfide cross-linking experiments to evaluate a given fingerprint. For now, let us assume that the form of experimental data X regarding a fingerprint F is a binary vector indicating for each edge whether or not the SSE pair was found to be in contact. Let us denote by $\mathcal{X} = \{0, 1\}^{|F|}$ the set of possible binary vector values for X. Then the likelihood takes the joint probability over the edges, testing agreement between the observed contact state and that expected under the model:

$$\Pr(X \mid m) = \prod_{i=1}^{|F|} \Pr(F_i = X_i \mid m) \tag{4.8}$$

where we use the subscript to get the i^{th} element of the set. The naive conditional independence assumption here is reasonable, since the elements of F_i (SSE contact states) depend directly on the model, and are thus conditionally independent given the model. We then select the model with the highest likelihood. (If we have informative priors, evaluating model quality, we could instead select based on posterior probabilities.)

Plan evaluation

In the experiment planning phase, we don't yet have the experimental data. However, we can evaluate the potential for making a wrong decision using a given plan by computing the *Bayes error*, ϵ. If we knew which model m were correct and which dataset X we would get, we could evaluate whether or not we would make the wrong decision, choosing a wrong model m' due to its having a higher likelihood for X than the correct model m. The Bayes error considers separately each case where one particular model is correct and one particular dataset results, and sums over all the possibilities. It weights each possibility by its probability—is the model likely to be correct, and if it is, are we likely to get that

dataset. Thus:

$$\epsilon = \sum_{m \in M} \Pr(m) \cdot \sum_{X \in \mathcal{X}} \Pr(X \mid m) \cdot \mathbf{1}(\Pr(X \mid m) < \max_{m' \neq m} \Pr(X \mid m')) \tag{4.9}$$

where $\Pr(m)$ is the prior probability of a model, which we currently take as uniform, but could instead be based on fold recognition scores. Here and in the following formulas we use an indicator function $\mathbf{1}$ that gives 1 if the predicate is true and 0 if it is false. So we assume each different model is correct (at its prior probability), and assess whether or not it would be beaten for each different data set (at probability conditioned on the assumed correct model). This framework thereby gives a probabilistic evaluation of how likely it is that we will make an error, in place of the usual empirical cross-validation that is performed to assess a feature subset selected for classification.

In the case of fold determination, there may not be a single best model—a number of models may in fact have the same fold, and thus be equally consistent with the experimental data. Thus in the data interpretation phase we would not want to declare a single winner, but instead would return a set of the tied-for-optimal models. In the experiment planning phase, we develop a complementary metric to the Bayes error, which we call the *expected tie ratio*, τ:

$$\tau = \sum_{m \in M} \Pr(m) \cdot \sum_{X \in \mathcal{X}} \Pr(X \mid m) \cdot \frac{\sum_{m' \in M, m' \neq m} \mathbf{1}(\Pr(X \mid m) = \Pr(X \mid m'))}{|M|} \tag{4.10}$$

The formula mirrors that for ϵ, but instead of counting the number of incorrect decisions, it counts the fraction of ties. Evaluating τ as we build up a topological fingerprint allows us to track the incremental power to differentiate folds, up to the point where we find that a set of models has the same fold and τ has flat-lined. The metric can readily be extended to account for sets of models whose likelihood is within some threshold of the best.

Finally, the topological fingerprint approach allows us to handle the "none-of-the-above"

scenario, when we decide that no model is sufficiently good; i.e., the correct fold isn't represented by a predicted model. While in other contexts that would be done by comparing the likelihood to some threshold (is the selected model "good enough"?), here we can actually explicitly consider the chance of not considering the correct fold. Note that since a fingerprint typically has a small number of SSE pairs, we can enumerate the space $\mathcal{F} = \{0,1\}^{|F|}$ of its possible values (indicating whether or not each SSE pair in the fingerprint is in contact). Some of those values, $\mathcal{F}_M$, correspond to models in M, while the rest, $\mathcal{F} - \mathcal{F}_M$, are "uncovered". We want to decide if an uncovered fold $f' \in \mathcal{F} - \mathcal{F}_M$ is better than the fold f for the selected model. Moving from models to folds, we can evaluate $\Pr(X \mid f)$ by a formula like (4.8), simply testing whether each X_i has the value specified in f. Then we can decide that it is "none of the above" (models) if $\exists f' \in \mathcal{F} - \mathcal{F}_M$ such that $\Pr(X \mid f') \geq \max_{f \in \mathcal{F}_M} \Pr(X \mid f)$.

Moving from data interpretation to experiment planning, we can again evaluate a plan for the probability of deciding none of the above. If we think of Bayes error as the false positive rate, then we want something more like a false negative rate. We call this metric ν, the *expected none-of-the-above ratio*.

$$\nu = \sum_{f' \in \mathcal{F} \setminus \mathcal{F}_M} \Pr(f') \cdot \sum_{X \in \mathcal{X}} \frac{1}{2^{|F|}} \cdot \mathbf{1}(\Pr(X \mid f') > \max_{f \in \mathcal{F}_M} \Pr(X \mid f)) \tag{4.11}$$

Thus ν is the fraction of experimental datasets for which an uncovered fold will be better than the best covered fold. We currently do not include a prior on X, in order to provide a direct assessment of how many experiments could lead to a none-of-the-above decision. However, we could obtain a weighted value by estimating $\Pr(X)$, e.g., from the priors on the individual SSE pairs (from (4.1)). For the same reason, we treat $\Pr(f')$ as uniform over the uncovered folds f', rather than evaluating it by priors on SSE pairs.

Note that the formula does not include SSE pairs in $(S \times S) \setminus F$; i.e., pairs not in the fingerprint. This is as if they contribute equally to covered and uncovered folds, and thus

do not affect the outcome. In the absence of other information or assumptions about the uncovered folds, this is a reasonable (and conservative) assumption, and yields an interpretable metric.

4.2.2 Cross-link selection

Once a topological fingerprint F has been identified, the next task is to optimize a disulfide cross-linking plan to experimentally evaluate the SSE pairs in the fingerprint. We separately plan for each SSE pair (their conditional independence was discussed in the previous section), optimizing a set of disulfide cross-link experiments (a single cross-link per experiment), such that, taken together, these cross-links will reveal whether or not the SSE pair is in contact. The overall plan is then the union of these SSE-pair plans. Thus we focus here on planning for a single SSE pair. We must account for noise and uncertainty in both the model and the actual protein, as well as for dependency among cross-links. This work represents the first to address these issues.

Different models may place an SSE at somewhat different residues, so when planning cross-links to probe that SSE's contacts, it is advantageous to focus on residues common to many models (and thus able to provide information about cross-linkability in those models). We define for each SSE a set of common residues that may be used in a disulfide plan. Our current implementation includes all residues that appear in at least half of the models that have that SSE. In the following, let R denote the common residues for a target SSE pair.

For each model m_i we construct a *residue cross-link graph* $G_{xlink,i} = (R, D_i)$, in which the nodes are common residues R and there are edges $D_i \subset R \times R$ between possible disulfide pairs (specific to each model). We compute the *cross-linking distance* for a residue pair as the C^β–C^β distance, and take as edges those with distance at most 19 Å, based on an analysis of rates of disulfide formation [9, 79]. Our method could be generalized to include a more detailed geometric evaluation of the likelihood of cross-linking.

Probabilistic model

We must define a probabilistic model in order to evaluate the information content provided by a set of cross-links. We treat possible cross-link (pair of residues) as a binary random variable indicating whether or not there is a cross-link. We start with the model of our earlier work, in which the prior probability of a cross-link wrt a model is 0.95 for distances ≤ 9Å, 0.5 for distances between 9 and 19 Å, and 0.05 for those > 19 Å [79]. However, we also account for two important types of noise in this context: threading misalignment and structural flexibility (Figure 4.2).

We place a distribution $\Pr(\delta)$ over possible offsets by which an SSE could be misaligned in a model. That is, residue number r in the model is really residue $r + \delta$ in the protein, and thus a cross-link involving residue $r+\delta$ is really testing proximity to residue r. We use a distribution with 0.5 probability at 0 offset, decaying exponentially on both sides up to a maximum offset. Analysis of a model or the secondary structure prediction could provide a more problem-specific distribution. We currently consider each SSE separately; a future extension could model correlated misalignments resulting from threading.

We sample a set of alternative backbones for a model, and place a distribution $\Pr(b)$ over the identities of these alternatives. While there are many ways to sample alternative structures, we currently use Elastic Normal Modes (ENMs) as implemented by $elN\acute{e}mo$ [63], sampling along the lowest non-trivial normal mode. We set $\Pr(b)$ according to the amplitude of the perturbation, using a Hookean potential function derived from ENMs. Future extensions could model different aspects of flexibility, such as local unfolding events during which a cross-link may be captured.

These two factors result in dependence among possible cross-links: if an SSE is misaligned or has moved relative to the original model, all its cross-links will be affected. However, the cross-links are conditionally independent given the particular value of mis-

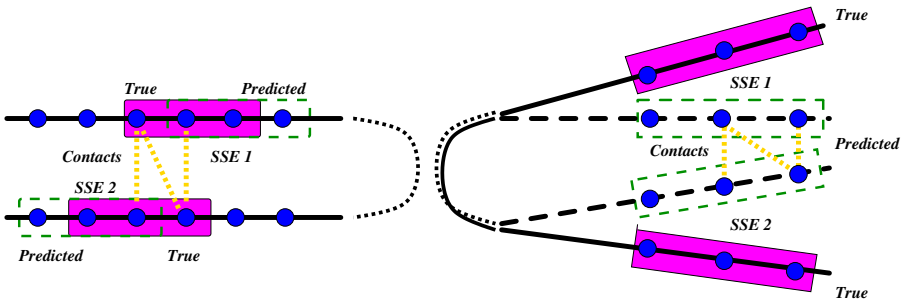


Figure 4.2: Noise factors in cross-link planning: misalignment (left) and flexibility (right). Blue dots represent residues and yellow lines their contacts. Regions in dashed lines are the modeled SSE and those in solid lines those measured by cross-linking experiments.

alignment or backbone choice. Thus we have for any two cross-links ℓ, ℓ':

$$\Pr(\ell, \ell') = \sum_{m} \Pr(m) \cdot \sum_{\delta} \Pr(\ell \mid m, \delta) \cdot \Pr(\ell' \mid m, \delta) \cdot \Pr(\delta) \tag{4.12}$$

and similarly for backbone flexibility. Furthermore, misalignment and flexibility are independent. Detailed steps about how to compute $\Pr(\delta)$ will be discussed in Section 4.5.1.

Experiment planning

Our goal is to select a "good" set of residue pairs $L \subset R \times R$ to experimentally cross-link, in order to assess whether or not the SSE pair is in contact. This is another feature subset selection problem, and we again employ an mRMR-type incremental algorithm. Here a possible cross-link ℓ's relevance is evaluated in terms of the information it provides about whether or not the SSE pair is in contact: $I(\ell, c)$, where c is the binary random variable for contact of a target SSE pair. Redundancy is again evaluated in terms of mutual information. Thus the objective is:

$$s(L) = \frac{1}{|L|} \sum_{\ell \in L} I(\ell, c) - \frac{1}{|L|^2} \sum_{\ell, \ell' \in L} I(\ell, \ell') \tag{4.13}$$

and we incrementally select cross-links to maximize the difference in relevance regarding contact and average redundancy with already-selected cross-links.

Data interpretation

Once we have experimentally assessed cross-link formation for each selected residue pair, we can evaluate the probability of the SSE pair being in contact. Let Y be the set of cross-linking data, indicating for each residue pair in L whether or not a disulfide was detected. To decide whether or not c is in contact, we will compare $\Pr(Y \mid c = 1)$ and $\Pr(Y \mid c = 0)$, and take the one with higher likelihood. Intuitively, the more cross-links that are detected, the more confident we are that the SSE pair is in contact. Thus we currently employ a sigmoidal function to evaluate the likelihood:

$$\Pr(Y \mid c = x) = \frac{1}{1 + e^{(-1)^x \cdot (k - k_0)}} \, . \tag{4.14}$$

Here k is the number of detected cross-links in Y, and k_0 is the minimum number of positive cross-links for us to start believing c is in contact. For example, for $c = 1$, given a default number of 10 experiments, we set $k_0 = 3$ and the likelihoods of $c = 1$ for $k = 0$, 3, 6 are then approximately 0.05, 0.5, and 0.95, respectively. The metric could be extended to reward the broader distribution of cross-links throughout each SSE. However, in our current framework, we find that having a sufficient number of cross-links without regard to location tends to achieve that goal.

Plan evaluation

Finally, in order to assess an experiment plan's robustness, we develop a Bayes error criterion to evaluate the probability of making a wrong decision regarding SSE contact.

$$\epsilon = \sum_{x \in \{0,1\}} \Pr(c = x) \cdot \sum_{Y \in \mathcal{Y}} \Pr(Y \mid c = x) \cdot \mathbf{1}(\Pr(Y \mid c = x) < \Pr(Y \mid c \neq x)) \quad (4.15)$$

As in the previous section, we sum over the possible outcomes (here, in contact or not) and the possible experimental results ($\mathcal{Y} = \{0,1\}^{|L|}$, all binary choices for cross-links in plan L), weighted by their probabilities, and see which yield the wrong decision. In the absence of an informative prior for c (and one that we want to use in interpreting the data), we simply use $\Pr(c = 1) = \Pr(c = 0) = 0.5$.

Note that, if desired, we could use the cross-linking Bayes error as a replacement for q (as $1 - \epsilon$) in evaluating $\Pr(c = x)$. These values could be precomputed for all candidate SSE pairs, or a fingerprint could be reevaluated and perhaps modified upon evaluating its possible cross-link plan.

4.3 Results

We demonstrate the effectiveness of our approach with a representative set of 9 different CASP targets (Table 4.3), including proteins that are all-α, some that all-β, and some that are mixed α and β. For each target, a number of high-quality models have been produced by different groups; we evaluate those of common SSE content, as described in the methods. The models vary in similarity to the crystal structure (the PDB ID indicated), which is unknown at the time of modeling and furthermore not used for experiment planning, as well as to each other (the average root mean squared deviation in atomic coordinates, RMSD,

CASP ID	PDB ID	2^o	# residues	# models	Avg. RMSD
T0283_D1	2hh6	5α	97	162	17.26
T0289_D2	2gu2	5β	74	34	13.45
T0299_D1	2hiy	3α, 3β	91	30	15.23
T0304_D1	2h28	2α, 5β	101	26	15.76
T0306	2hd3	7β	95	45	14.22
T0312_D1	2h6l	2α, 5β	132	55	16.13
T0351	2hq7	5α	117	65	15.42
T0382_D1	2i9c	6α	119	196	12.79
T0383	2hnq	2α, 4β	127	59	11.61

Table 4.1: Test data sets (from CASP7)

between pairs of models is indicated). Our goal is to select for each target an experiment plan to robustly determine the model(s) of the same fold as the crystal structure.

Topological fingerprint selection

Figure 4.3 shows the trends of Bayes error (ϵ), expected tie ratio (τ), and expected none-of-the-above ratio (ν) as more SSE pairs are included in the topological fingerprint. It may seem counterintuitive that ϵ initially increases with the addition of SSE pairs. However, this is because we define the Bayes error of a tie as zero (4.9), and separate out the tie ratio. With few SSE pairs in the fingerprint, τ is generally high—few decisions will be made, as many models look equally good, and the Bayes error is small. Then as SSE pairs are added, τ drops sharply—the fold is more specifically determined, decisions will be made, and the potential for error (as reflected in the Bayes error) increases. Once a sufficient number of SSE pairs has been selected, the specifically-determined fold is distinct, and the decisions are likely to be right, and ϵ will decrease. Thus it is both appropriate and helpful to consider ϵ and τ together, as they provide complementary information in the progress toward obtaining a unique and correct fold.

On the other hand, we observe that the ν value is usually 0 in the first few steps, because at that point there are not distinct folds separated, and it is easy for the SSE graphs from the predicted models to "cover" all the possible folds. ν becomes non-zero when there

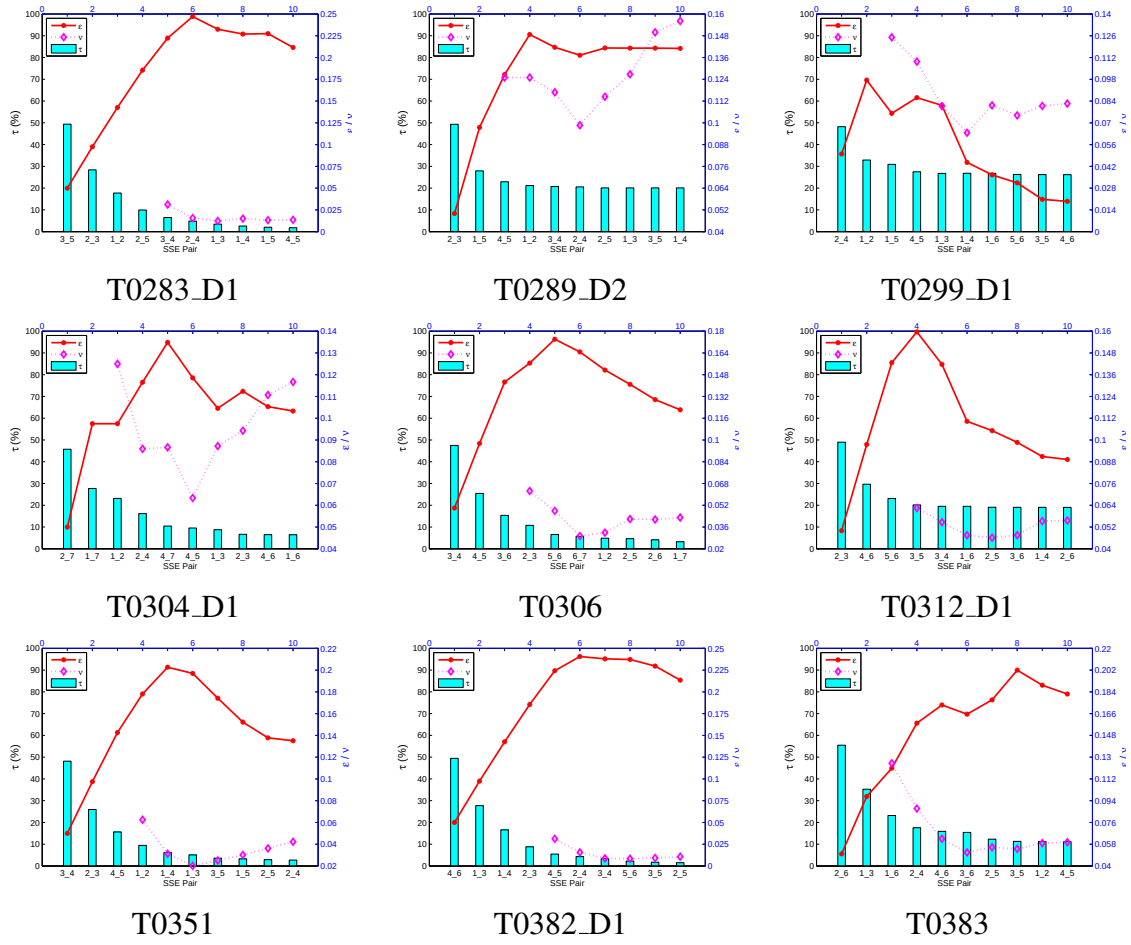


Figure 4.3: Bayes error (ϵ), expected tie ratio (τ), and expected none-of-the-above ratio (ν) with addition of SSE pairs to fingerprints for targets. x-axis: SSE pairs. y-axis (left): τ, (%). y-axis (right): ϵ, ν.

are uncovered folds. Its value first decreases because the number of covered folds and the number of uncovered folds are both increasing as more SSE pairs are included, and ν only gets contributions from an uncovered fold with *greater* (not equal) likelihood as the best covered fold. At some point the number of covered folds stops increasing (due to the limited set of predicted fold types), while the number of uncovered folds is still growing. Then the additional fold possibilities in the uncovered space result in a higher risk of "none-of-the-above", and thus the ν value starts increasing again. This trend is particularly obvious for targets T0289_D2 and T0304_D1; in fact, we return to T0304_D1 below as a real example

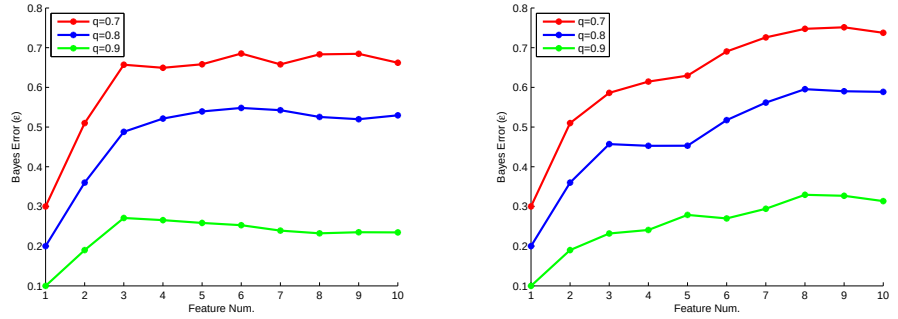
of "none-of-the-above".

The fingerprint evaluation incorporates a parameter in the q function (4.2), essentially indicating the confidence we expect to have in the experimental evaluation of an SSE pair. We performed a sensitivity analysis for three values of q, from 0.7 (fairly ambiguous) to 0.9 (fairly confident). Figure 4.4 shows that for two targets the trends are very similar in all three cases; our algorithm is insensitive to the choice. we omit showing other targets since they display similar insensitivity. A verification with our cross-link plans show their bayes errors fit best with the 0.9 setting.

End-to-end simulation study

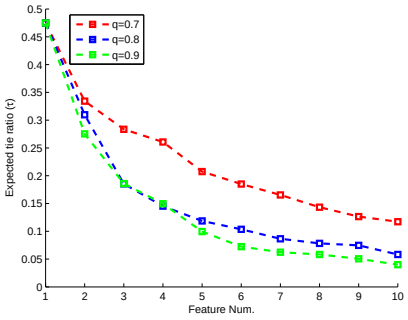
Once we have selected a topological fingerprint, we next design a disulfide cross-linking plan to determine the contact state of the selected SSE pairs. To validate the overall process (fingerprint + disulfides), we perform a simulation study. Given a selected set of residue pairs for cross-linking, we use the crystal structure (PDB entry in Table 4.3) to determine whether or not they should form disulfides (C^{β}-C^{β} distance < 9 Å), and treat those evaluations as the data. We also use the set of all SSE pairs to directly compare the fold of each model with that of the crystal structure, and thereby label each model as being the "correct" fold or not depending on whether or not they have the same SSE contacts for the same SSE pairs. We then evaluate whether or not the simulated data for the selected cross-linking plans result in the same conclusions as the direct comparisons of folds.

To compare the decision based on simulated cross-linking data with that based on fold analysis, we performed a Receiver Operator Characteristic (ROC) analysis. The area under the ROC curve (AUC) measures the probability that our experiment plan will rank a randomly chosen positive instance higher than a randomly chosen negative one. The larger the AUC, the better classification power our algorithm has to detect the right fold. Figure 4.5 illustrates the simulation results on eight example protein targets (ROC analysis

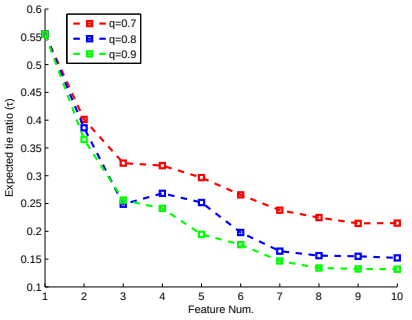


T0306, ϵ

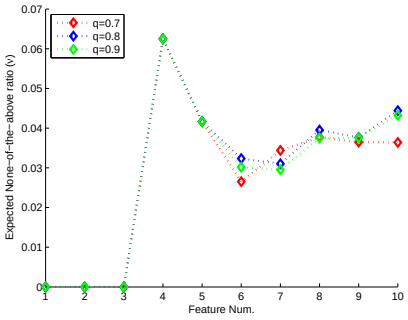
T0383, ϵ



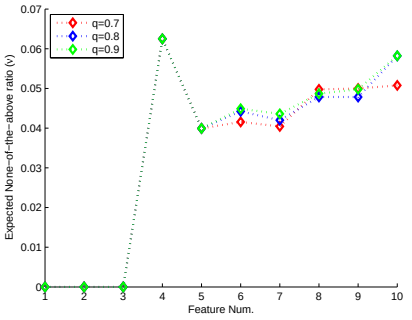
T0306, τ



T0383, τ

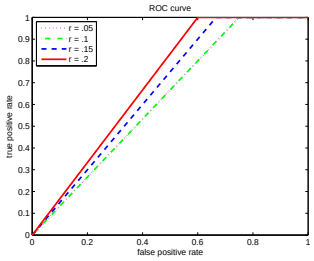


T0306, ν

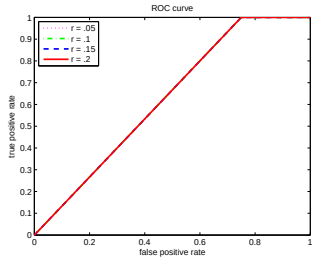


T0383, ν

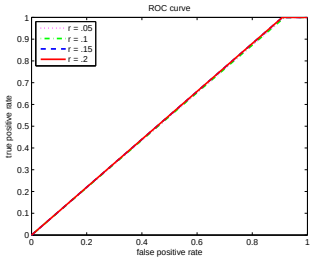
Figure 4.4: Sensitivity analysis for three q function values (0.7, 0.8, and 0.9) for target T0306 and T0383.



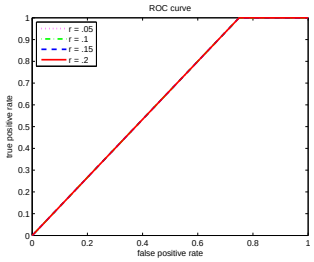
T0283_D1
(AUC: 0.63; 0.63; 0.67; 0.7)



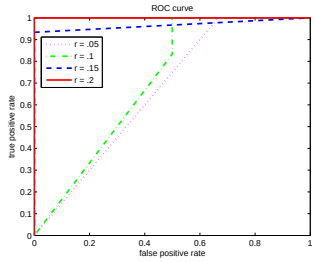
T0289_D2
(AUC: 0.63; 0.63; 0.63; 0.63)



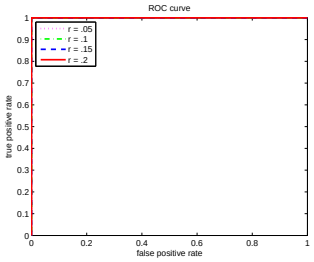
T0299_D1
(AUC: 0.54; 0.54; 0.55; 0.55)



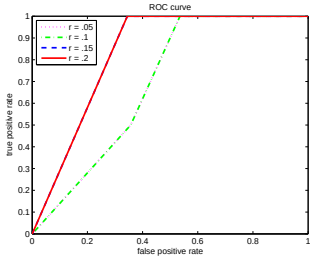
T0306
(AUC: 0.63; 0.63; 0.63; 0.63)



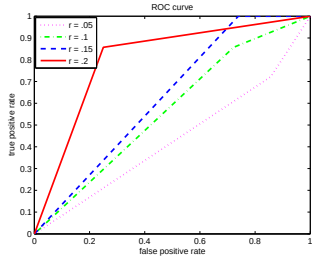
T0312_D1
(AUC: 0.67; 0.71; 0.97; 1.0)



T0351
(AUC: 1.0; 1.0; 1.0; 1.0)



T0382_D1
(AUC: 0.69; 0.69; 0.83; 0.83)



T0383
(AUC: 0.43; 0.57; 0.63; 0.80)

Figure 4.5: ROC curves for eight simulation studies, at different SSE contact fraction thresholds r. T0304_D1 doesn't have a predicted model that matches the crystal structure and thus is analyzed separately (see the discussion for *Robustness*). x-axis: False Positive Rate. y-axis: True Positive Rate. AUC: Area under the ROC curve, for r of 0.05, 0.1, 0.15, and 0.2 respectively.

for T0304_D1 is not applicable and we will discuss it below). ROC curves are shown for different thresholds for the percentage r of residues that must be in contact to declare that the SSE pair is in contact in the structure or model. A high r value results in very few SSE pairs deemed to be in contact (we found that to happen with $r = 0.3$), while a low one yields some fairly weak contacts. As the figure shows, a moderate r value of around 0.2 generally results in quite good fold determination results.

Robustness

One of the merits of the fold determination approach is that it is robust to errors in models, and can even account for the case when none of the models is correct. The selected targets provide examples requiring such robustness; we summarize here just a couple. *Misalignment*. In (4.12) we account for being off by up to δ residues in the SSE locations. In the case of T0312_D1, there are 23 models of the correct fold, but with $\delta = 0$, only7 of them agree with the crystal structure regarding all the cross-links in the experimental plan, while with $\delta = 1$ there are 14 that agree, and with $\delta = 2$ there are 16. The remaining unmatched models are looser in structure, and the match is sensitive to the threshold we use to measure SSE contacts. *None-of-the-above.* For target T0304_D1, none of the models has the same SSE contact graph as the crystal structure. The GDT [82] scores of predicted models are in the low 30s, which indicates relatively poor agreement with the crystal structures. As shown in Figure 4.3, the ν value is relatively high, indicating a potential risk of missing the right fold. Indeed once we evaluate the models under the simulated data, we find that the likelihoods are low ($< 2 \times 10^{-3}$), compared to that (≈ 0.66) of the uncovered but correct fold, which is found by enumeration.

4.4 Conclusion

This chapter presents a computational-experimental mechanism to rapidly determine the overall organization of secondary structure elements of a target protein by probing it with a planned set of disulfide cross-links. By casting the experiment planning process as two stages of feature selection—SSE pairs characterizing overall fold and residue pairs characterizing SSE pair contact states—we are able to develop efficient information-theoretic planning algorithms and rigorous Bayes error plan assessment frameworks. Focusing on fold-level analysis results in a novel approach to elucidating three-dimensional protein structure, robust to common forms of noise and uncertainty. At the same time, the approach remains experimentally viable by finding a greatly reduced set of residue pairs (tens to around a hundred, out of hundreds to thousands) that provide sufficient information to determine fold.

4.5 Appendix

4.5.1 Noise model in cross-link experiment planning

In Section 4.2.2, we discussed how to model the noise in the context of threading misalignment and structure flexibility. One of the questions is how to compute $\Pr(\delta)$ and now we will elaborate it in this section.

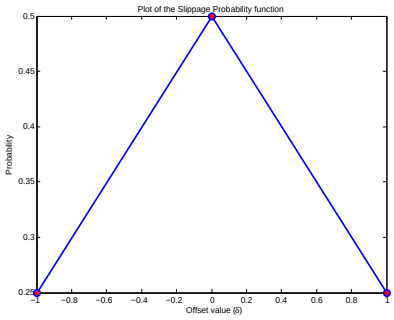
Misalignment model

From Figure 4.2, it is clear that now the likelihood of observing one cross-link for a specific residue pair not only depends on its own distance measure, but also on those of the neighbor edges it could "slip" to. Let Δ be the set of possible offsets of an SSE's alignment, and δ^* be the absolute value of the maximum offset. Then Δ should contain a total $2\delta^* + 1$ misalignment offsets. For example, with $\delta^* = 1$, we have $\Delta = \{-1, 0, +1\}$, where -1, 0, and $+1$ mean backward one step, no change and forward one step, respectively. Suppose the protein target has k SSEs, and each SSE is subject to the misalignment offsets in Δ; then we have the enumeration of structure-wide misalignments as $\sigma = \underbrace{\Delta \otimes \Delta \ldots \otimes \Delta}_{k}$, where $\otimes$ is the outer product and $|\sigma| = |\Delta|^k$.

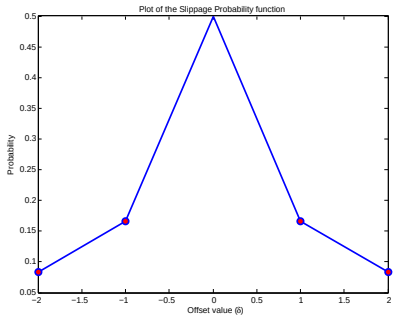
In Δ, each possible offset δ may have different probability of happening. For simplicity, we can either assume that it follows the standard distribution after being normalized to $[0, 1]$ or we can use a piecewise function to simulate it as below (more details as in Figure 4.6).

$$\Pr(\delta) = \begin{cases} 1/2 & \delta = 0\ ; \\ \frac{2^{(\delta^* - |\delta| - 2)}}{2^{\delta^*} - 1} & 1 \leq |\delta| \leq \delta^*\ . \end{cases} \tag{4.16}$$

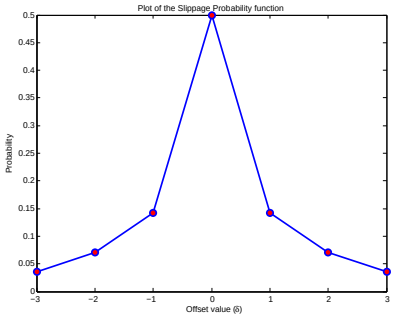
[4]This section covers supplementary methods and results not included in the BIOKDD paper.



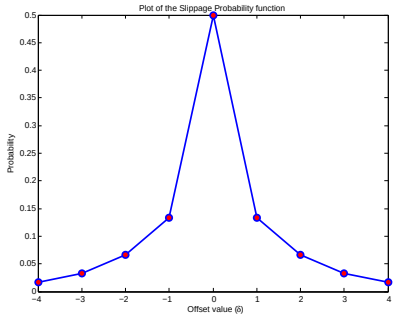
(a) misalignment ($\delta^* = 1$)



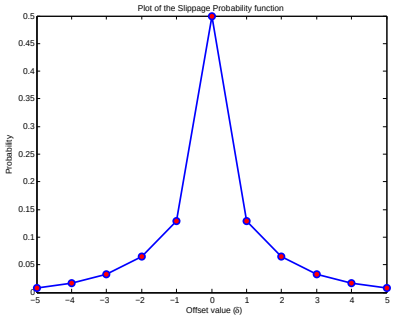
(b) misalignment ($\delta^* = 2$)



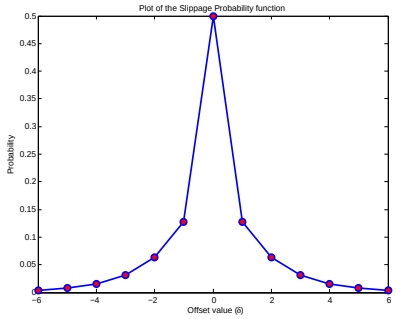
(c) misalignment ($\delta^* = 3$)



(d) misalignment ($\delta^* = 4$)



(e) misalignment ($\delta^* = 5$)



(f) misalignment ($\delta^* = 6$)

Figure 4.6: Plot of the probability function for threading misalignment with different offset limits. x-axis: Offset value (δ). y-axis: Probability.

Flexibility model

Flexibility comes from protein dynamics. While numerous techniques are available for simulation (molecular dynamics, etc.), we study here the case of elastic normal mode (ENM) analysis. ENM provides an analytic description of the dynamics in a macromolecular system near a minimum energy, by using a harmonic approximation of the potential. In this work, we use a software called $elN\acute{e}mo$ [63] to generate the normal modes. It allows the computation of low frequency normal modes for a given protein structure, and provides an analysis of those models at different levels of detail. Since our goal is to study the noise caused by major structure flexibility, normal models with low frequencies for non-rigid-body motion are selected to generate the simulation data. Moreover, each normal mode is projected over 11 amplitudes of perturbation, leading to 11 snapshots of the target structure.

To simplify the analysis, we illustrate here an analysis with just the 7-th normal mode, which has the lowest frequency for non-rigid-body motion and counts for the most important global conformation change. A similar approach as used for threading misalignment can be applied here to compute the likelihood of cross-link flexibility at a specific residue pair. Here δ corresponds to the amplitude of perturbation, and $\Pr(\delta)$ is derived from a Hookean potential function $E_p = c \cdot x^2$ used in the standard ENM, with the spring constant c equal to 1. δ^* is the maximum perturbation allowed along the projection of one normal mode.

$$\Pr(\delta) = \frac{\int_{|\delta|}^{\delta^*+1} x^2 dx}{2 \times \int_0^{\delta^*+1} x^2 dx} \tag{4.17}$$

4.5.2 Proof sketch of the optimality of mRMR on first-order incremental search

In Section 4.2.1, we briefly introduced the Max-Dependency feature selection problem. It can be formulated as

$$\max D(S; c); \text{ where } D = I(S_m; c) \ .$$

Here S is the feature space, c is the set of class labels, and S_m are the selected features. I is the function for multivariate mutual information. We can prove the equivalence of this problem with our (seemingly simpler) mRMR-based problem and algorithm.

Theorem 2 (optimality of mRMR on first-order incremental search) *For the first-order incremental search, mRMR is equivalent to Max-Dependency for the topological fingerprint identification problem.*

Proof sketch: This is very similar with the theorem presented in [50]. Here we just need to change the definition of $J(x_1, x_2, \ldots, x_m)$ ("mutual information"' for multiple scalar variables) from the continuous data format to the discrete data format. Likewise, we can also rewrite the definition of Max-Dependency as $I(S_m; c) = J(S_{m-1}, x_m, c) - J(S_{m-1}, x_m)$. Then Max-Dependency is equivalent to simultaneously maximizing the first term and minimizing the second term. Applying the same inequality as in [50], we know that Max-Dependency is reached when all variables are maximally dependent in $J(x_1, x_2, \ldots, x_m, c)$ but the pair-wise inter-dependency of $x_1, x_2, \ldots, x_m$ is minimized. Since $S_{m-1} = x_1, x_2, \ldots, x_{m-1}$ is fixed at the m-th step, the optimality is equivalent to finding the x_m that maximizes (4.7) as described by our algorithm. Thus we have proved that mRMR is equivalent to Max-Dependency when performed with the first-order incremental search.

4.5.3 Supplementary results

A comparison with the random selection

We also compare our approach with random SSE pair selections, which choose the same number of SSE pairs as in the fingerprint, but in a stochastic manner. Figure 4.7 illustrates the trends for ϵ, τ, and ν for two protein targets, over a set of ten random selections each. Again, we see that random selection may have a better ϵ than our approach in the

beginning due to a relatively high τ ratio. With more SSE pairs included, ϵ starts to drop quickly for our approach, while it doesn't have the same pattern for random selection. This shows that our approach has the ability to choose more informative SSE pairs and will beat random selection by reducing Bayes errors. On the other hand, our approach consistently outperforms random selection in both τ and ν, which further prove its effectiveness.

The likelihood score distribution of matched models

In Section 4.3, we presented an end-to-end study where we use the protein target's crystal structure to simulate cross-link data, and then evaluated the predicted models based upon the simulated data. As an illustration, Figure 4.8 shows the likelihood scores of models to match the appropriate crystal structure, given the simulated data. Here misalignment is allowed to have a higher flexibility of matching a predicted model's fold with the crystal structure's. In general, a larger misalignment offset (δ) will make the match less favorable, as shown by the lower likelihood score. However, we would like to keep the choice open while letting experimentalists set their own preferences between match quality and result completeness.

In addition, Figure 4.9 illustrates how the number of matched true positives (models) changes given various δ values for misalignment.

Plan evaluation for cross-linking experiments

In Section 4.3, we have elaborated how to select topological fingerprints for protein fold determination. We also introduced an end-to-end simulation study which includes cross-link experiment planning to validate those topological patterns in the context of residue contacts. However, we didn't show the explicit results of cross-link planning due to the page limit of paper submission. We include them in this section.

Figure 4.10 illustrates the trend of Bayes error (ϵ) as more cross-link candidates (residue

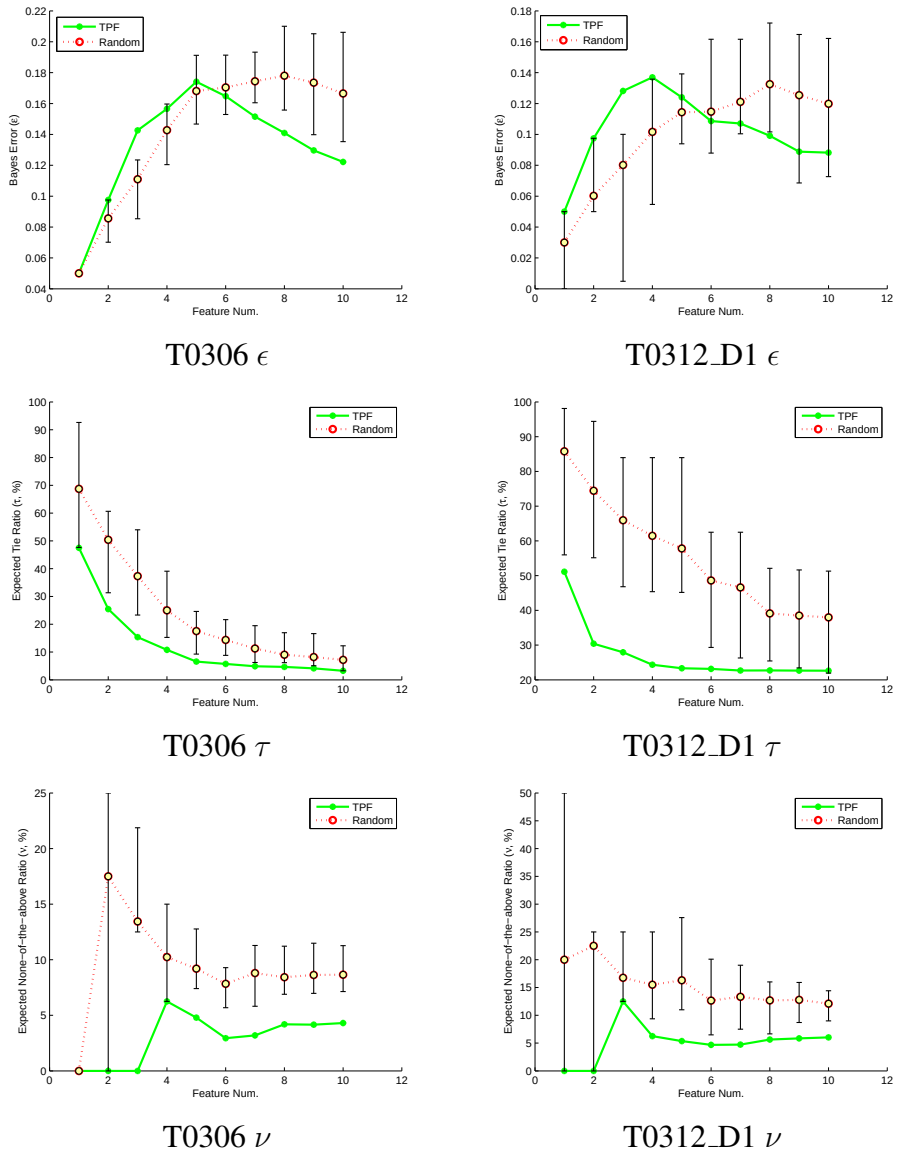
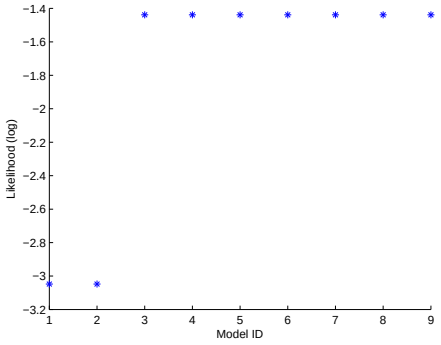
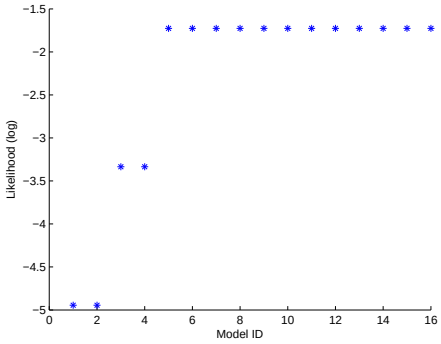


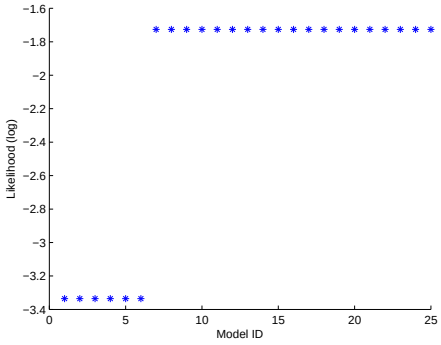
Figure 4.7: Performance comparison for ϵ (top), τ (middle), and ν (bottom) of random selections vs. topological fingerprints for protein targets T0306 (left) and T0312 (right). The green curve is the topological fingerprint selection and the the red curve is the average result of ten random SSE pair selections. Bars indicate the minimum metric value and the maximum metric value over the random selections.



T0283_D1



T0312_D1



T0383

Figure 4.8: Likelihood score distributions for three protein targets, evaluating models against to their reference crystal structures with threading misalignment allowed. Matched models are sorted by their likelihood. Log-transform was applied for a better illustration of the likelihood scores. x-axis: Model ID. y-axis: Likelihood (log).

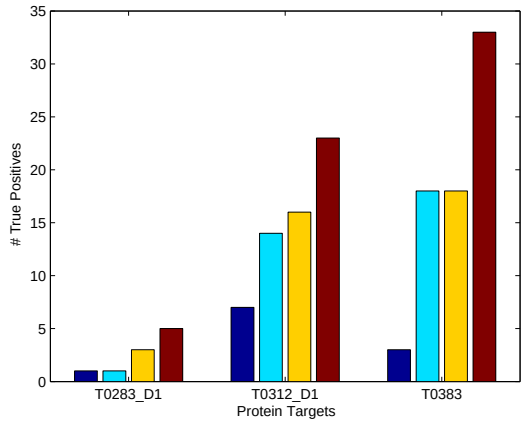
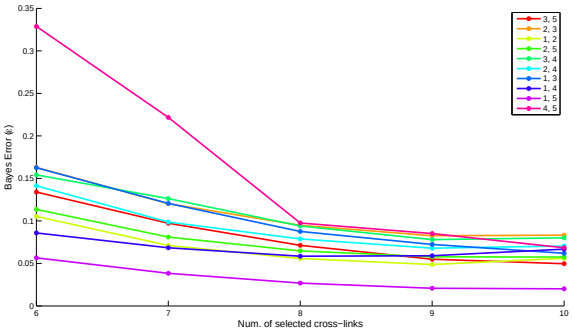
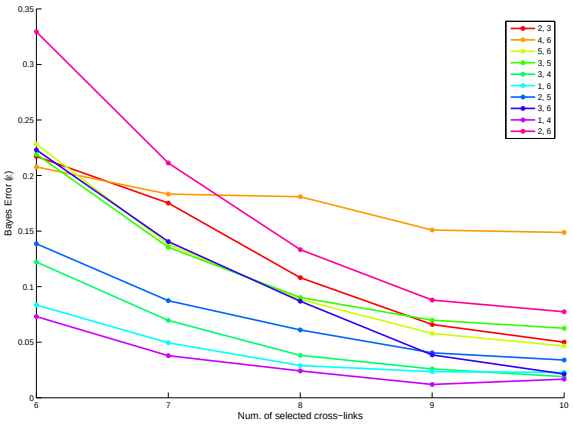


Figure 4.9: Case studies for three protein targets: The true positives (models matched with the crystal structure's fold) when misalignment is allowed, and $\delta = 0$ (blue), 1 (cyan), 2 (yellow). The red bar shows the total number of true positives for each target.

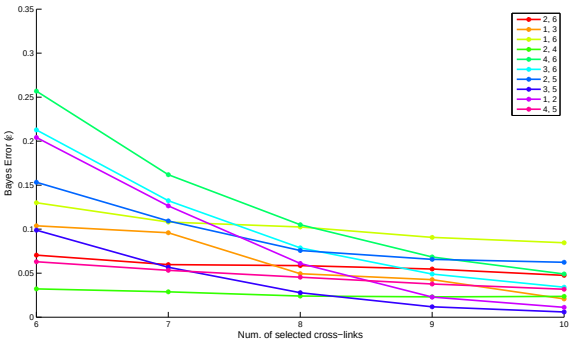
pairs) are included in the experiment plan for each SSE pair in the topological fingerprint. We demonstrate the results by showing only three protein targets in our test data sets, but the rest of them have very similar patterns. There is a clear trend of descending Bayes errors when we choose to test more cross-links in our experiment plans. This is of course expected since we then have more experimental information. However, it is also interesting to observe that different SSE pairs have different slopes to reach a low Bayes error, indicating the relative difficulties for the cross-link experiments to capture the SSE contact information. Part of the reason is because of the length of the SSE. For example, SSE pair $4, 5$ (pink line) of T0283_D1 has over 13 residues for each SSE, thus making it harder to select a limited number of cross-links to properly "sample" the SSE contact. Meanwhile, we also see a overall slowing trend for the reduction in Bayes error, meaning that the marginal benefit of including more cross-links for test is diminishing. Based on the results we have seen so far, our plan can generally achieve a state with low Bayes error (< 0.1, as the default value set for the likelihood computation in topological fingerprint selection), by including just 8 or 9 cross-links for the experiment.



T0283_D1



T0312_D1



T0383

Figure 4.10: Bayes error (ϵ) with addition of cross-links to the experiment plans for three protein targets. x-axis: Number of selected cross links. y-axis: Bayes error (ϵ). Names in the legend stand for the SSE pairs included in the topological fingerprint of corresponding protein target. We perform separate cross-link experiment planning for each SSE pair.

Chapter 5

Summary and Future Work

5.1 Summary

This thesis has introduced two significant new applications, contact replacement for NMR backbone resonance assignment and protein fold determination by disulfide cross-linking, and developed effective criteria and efficient algorithms based on graph representations of protein structure.

For the first application, we defined an approach called "contact replacement", which performs NMR backbone resonance assignment given a 3D structure and a set of relatively sparse ^{15}N-edited NMR data, with the through-space ^{15}N-edited NOESY as the primary source of information. Our approach supports high-throughput solution studies of dynamics and interactions (e.g., ligand binding), when the structure has previously been determined by crystallography or modeled computationally. It employs a graph matching approach by identifying correspondence between a given contact graph and a corrupted version representing the NMR data.

Two complementary algorithms have been developed for contact replacement, addressing different practical requirements. One is the Hierarchical Grow-and-Match (HGM) algorithm, which divides the contact graph into sequential fragments with relatively dense

interactions, and then combines all possible assignments for fragments in a hierarchical manner, searching over the combinations with effective but conservative pruning. It is guaranteed to be complete, returning all solutions consistent with the data within a likelihood threshold of the optimal solution. It also correctly handles missing edges and unassigned NMR peaks, which are quite common. Tests on a number of experimental datasets and simulations with varying noise and sparsity demonstrate that our algorithm can handle significant data corruption (2.5–6.0 noisy edges per correct one) and sparsity (10–40% of the correct edges missing). In addition to the reference solution, the complete ensembles include a number (up to 30) of alternatives. We use these complete ensembles to characterize confidence in parts of an assignment.

However, HGM only performs well on well-formed secondary structures (α-helices or β-strands), and the combinatorics of branch-and-bound search result in the algorithm taking a long time (and no theoretical guarantees) to return the complete ensemble. To better serve the need of large-scale NMR studies, we proposed a second algorithm based on a random graph approach. It does not guarantee completeness, but can assign the entire protein backbone and can provably find the best assignment with high probability in an expected polynomial time. We utilize the random structure of both noisy edges and ambiguous vertex type labels in the NMR graph, compared with the connectivity and type information in the contact graph, to effectively reduce the uncertainty of assignment. Since the NMR interaction graphs we are studying have up to 500% noise edges compared to correct edges, the ability to handle this degree of noise is important. A *reuse* paradigm was also introduced to avoid backtracking when an inconsistency is found during the assignment, and employ local fix-up rules to modify a solution that is mostly correct. Our empirical results show that this approach is quite effective in practice, relatively insensitive both to noise/missing in the NMR graph and structural variation in the contact graph.

For the second application, we defined an approach called "protein fold determina-

tion". Different from earlier methods that focused on probing geometry and selecting a model, it targets a better-defined characterization (overall protein fold). An integrated computational-experimental method has been presented to determine the fold of a target protein by probing it with a set of planned disulfide cross-links. We start with predicted structural models obtained by standard fold recognition techniques. In a first stage, we characterize the fold-level differences between the models in terms of topological (contact) patterns of secondary structure elements (SSEs), and select a small set of SSE pairs that differentiate the folds. In a second stage, we determine a set of residue-level cross-links to probe the selected SSE pairs. Each stage employs an information-theoretic planning algorithm to maximize information gain while minimizing experimental complexity, along with a Bayes error plan assessment framework to characterize the probability of making a correct decision once data for the plan are collected. By focusing on overall topological differences and planning cross-linking experiments to probe them, our fold determination approach is robust to noise and uncertainty in the models (e.g., threading misalignment) and in the actual structure (e.g., flexibility). We demonstrate the effectiveness of our approach in case studies for a number of CASP targets, showing that the optimized plans have low risk of error while testing only a very small portion of the quadratic number of possible cross-link candidates. Simulation studies with these plans further show that they do a very good job of selecting the correct model, according to cross-links simulated from the actual crystal structures. Fold determination can overcome scoring limitations in purely computational fold recognition methods, while requiring less experimental effort than traditional protein structure determination approaches.

5.2 Future work

5.2.1 Contact replacement

We have demonstrated that contact replacement is a powerful and effective approach for *backbone* NMR resonance assignment, able to obtain high-quality results, and robust to noise, missing data, and moderate structure variance. A natural thought is to extend this method to *side-chain* NMR assignment. Each amino acid has a unique side chain, differentiating it from other amino acids (and providing a unique chemistry). Indeed, side-chain NMR data are critical in standard protein structure determination, as they provide substantial information about contacts supporting the overall fold. It would thus be interesting to again see if the flow of information can be inverted, using a reference structure to assign the side-chains. Such a mechanism is also useful in the "inner loop" of structure determination. However, the conformational variability in side-chains is quite different from that in the backbone, with many relatively independent degrees of freedom, resulting in a larger search space for possible matches and requiring more tolerance for mismatches caused by noise or missing.

Another possible extension of our contact replacement approach is to apply it to protein complexes. Again, x-ray structures or computational models (e.g., from docking) could be available, providing contact information for some residue pairs. One concern is the possible difference between unbound and bound forms of the individual proteins in the complex, with possible systematic structural differences that would need to be accounted for. One particular complex type of interest is symmetric homo-oligomers, where similar subunits (protein chains) are arranged symmetrically to form the complex. Some previous research [53, 54] from our lab and our collaborators has introduced algorithms to conduct a complete search for all possible conformations, and evaluate them separately for consistency with NMR experimental data and for quality of packing. We are interested to further

investigate this problem by applying efficient randomized algorithms, like those we developed in contact replacement, to handle unassigned (and ambiguous) data in this approach.

5.2.2 Protein fold determination

One of the important questions in protein fold determination is how to evaluate an SSE pair's contact state. We treat the state as a binary value, determined from the number of residue contacts between the SSE pairs. However, in practice we found that this definition may be somewhat over-simplified, since it does not include the potentially crucial information of where those residue contacts are located, and their relative "strength" given their spatial distances. For example, as in Figure 4.2 (right), if most of the residue contacts are located near a "hinge" formed by the two SSEs, they will be less sensitive to structural flexibility (and more likely to exist) than those located further away. How to include location information to address such scenarios will then be critical to more precisely determining the SSE contact. A possible approach is to divide every SSE into *head*, *center* and *tail* regions, and mark each residue contact with the location label such as *head-to-center*, *head-to-tail*, etc. Alternatively, we could introduce an additional metric to measure the strength of a residue contact (e.g., by pairwise distance or number of constituent atoms in contact), and mark it as a "strong" contact or "weak" contact. Based on how many strong or weak residue contacts exist for the SSE pair, we would determine the SSE contact as "strong" or "weak" accordingly.

Likewise, location information could be very important for a cross-link experiment plan's quality in term of its "coverage" of the target SSE contact. Our current implementation uses a sigmoidal function which only considers the number of cross-links when computing the likelihood of cross-link data being generated from an expected SSE contact type. If the selected cross-links are more broadly distributed throughout the SSEs, we will have more confidence in their representative power to capture the true SSE contact information.

We have demonstrated preliminary results in using cross-linking for protein fold determination, accommodating moderate noise from either threading misalignment or structural flexibility. However, the determined fold may still be subject to significant errors in the case of substantial noise. Therefore, it is worthwhile to introduce complementary experimental techniques, such as mutagenesis [78], to further support our evaluation of models. A multimodal approach that incorporates data from both cross-linking and mutagenesis will be particularly interesting, since it reduces the risk of choosing a wrong model (fold) unless it receives consistent support from multiple modes of experimental data. The challenge is then to combine the two approaches in a quantitative decision framework, with illustrative parameters to manage each experimental approach's impact and the overall joint contributions. We expect more promising results from the multimodal framework when the parameters can be estimated more accurately based on the accumulated experimental data.

Bibliography

[1] R. Apweiler, A. Bairoch, and C. H. Wu. Protein sequence databases. *Current Opinion in Chemical Biology*, 8:76–80, 2004.

[2] L.V. Avramova, J. Desai, S. Weaver, A.M. Friedman, and C. Bailey-Kellogg. Robotic hierarchical mixing for the production of combinatorial libraries of proteins and small molecules. *J. Comb. Chem.*, 10:63–68, 2008.

[3] C. Bailey-Kellogg, S. Chainraj, and G. Pandurangan. A random graph approach to NMR sequential assignment. *J. Comp. Bio.*, 12:569–583, 2005. Conference version: Proc. RECOMB 2004, pp. 58-67.

[4] C. Bailey-Kellogg, A. Widge, J.J. Kelley III, M.J. Berardi, J.H. Bushweller, and B.R. Donald. The NOESY Jigsaw: Automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comp. Bio.*, 7:537–558, 2000.

[5] D. Bandyopadhyay, J. Huan, J. Liu, J. Prins, J. Snoeyink, W. Wang, and A. Tropsha. Structure-based function inference using protein family-specific fingerprints. *Protein Science*, 15(6):1537–1543, 2006.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[7] A. Bjorklund, T. Husfeldt, and S. Khanna. Approximating longest directed path. *Electronic Colloquium on Computational Complexity*, 32, 2003.

[8] B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

[9] C. Careaga and J. Falke. Thermal motions of surface -helices in the d-galactose chemosensory receptor detection by disulfide trapping. *J. Mol. Biol.*, 226:1219–1235, 1992.

[10] T. Chen, J. Jaffe, and G. Church. Algorithms for identifying protein cross-links via tandem mass spectrometry. *J. Comp. Biol.*, 8:571–583, 2001.

[11] Y. Chen, J. Reizer, M.H. Saier Jr., W.J. Fairbrother, and P. E. Wright. Mapping of the binding interfaces of the proteins of the bacterial phosphotransferase system, HPr and IIAglc. *Biochemistry*, 32:32–37, 1993.

[12] J.F. Doreleijers, M.L. Raves, T. Rullmann, and R. Kaptein. Completeness of NOEs in protein structures: A statistical analysis of NMR data. *J. Biomol. NMR*, 14:123–132, 1999.

[13] B. Dukka, E. Tomita, J. Suzuki, K. Horimoto, and T. Akutsu. Protein threading with profiles and distance constraints using clique based algorithms. *J Bioinform Comput Biol*, 4(1):19–42, 2006.

[14] M. Erdmann and G. Rule. Rapid protein structure detection and assignment using residual dipolar couplings. Technical Report CMU-CS-02-195, School of Computer Science, Carnegie Mellon University, 2002.

[15] T. Feder and R. Motwani. Finding large cycles in Hamiltonian graphs. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 166–175, 2005.

[16] T. Feder, R. Motwani, and C. Subi. Approximating the longest cycle problem in sparse graphs. *SIAM J. Comput.*, 31:1596–1607, 2002.

[17] H. N. Gabow. Finding paths and cycles of superpolylogarithmic length. In *Proceedings of the 36th ACM Symposium on the Theory of Computing (STOC)*, pages 407–416, 2004.

[18] M.R. Garey, D.S. Johnson, and R.E. Tarjan. The planar Hamiltonian circuit problem is NP-complete. *SIAM J. Comput.*, pages 704–714, 1976.

[19] A. Godzik. Fold recognition methods. *Methods Biochem. Anal.*, 44:525–546, 2003.

[20] S. Govindarajan, R. Recabarren, and R. A. Goldstein. Estimating the total number of protein folds. *Proteins*, 35:408–414, 1999.

[21] V. Grantcharova, D. Riddle, and D. Baker. Long-range order in the src SH3 folding transition state. *Proc Natl Acad Sci USA (PNAS)*, 97(13):7084–7089, 2000.

[22] P. Güntert, M. Saltzmann, D. Braun, and K. Wüthrich. Sequence-specific NMR assignment of proteins by global fragment mapping with program Mapper. *J. Biomol. NMR*, 17:129–137, 2000.

[23] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[24] P.J. Hajduk, R.P. Meadows, and S.W. Fesik. Drug design: Discovering high-affinity ligands for proteins. *Science*, 278:497–499, 1997.

[25] M. Haniu, L. O. Narhi, T. Arakawa, S. Elliott, and M. F. Rohde. Recombinant human erythropoietin (rHuEPO): cross-linking with disuccinimidyl esters and identification of the interfacing domains in EPO. *Protein Sci.*, 9:1441–1451, 1993.

[26] B. Hendriksen. Conditions for unique graph realizations. *SIAM Journal of Computing*, 21:65–84, 1992.

[27] J. Huan, W. Wang, D. Bandyopadhyay, J. Snoeyink, J. Prins, and A. Tropsha. Mining protein family specific residue packing patterns from protein structure graphs. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology (RECOMB)*, pages 308–315, 2004.

[28] R. Hughes, P. Rice, T. Steitz, and N. Grindley. Protein-protein interactions directing resolvase site-specific recombination: A structure-function analysis. *EMBO J.*, 12:1447–1458, 1993.

[29] D. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2):195–202, 1999.

[30] J.-S. Jung and M. Zweckstetter. MARS - robust automatic backbone assignment of proteins. *J. Biomol. NMR*, 30:11–32, 2004.

[31] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[32] H. Kamisetty, C. Bailey-Kellogg, and G. Pandurangan. An efficient randomized algorithm for contact-based NMR backbone resonance assignment. *Bioinformatics*, 22:172–180, 2006.

[33] L.E. Kay. Protein dynamics from NMR. *Nat. Struct. Biol.*, 5 Suppl:513–517, 1998.

[34] F. Khatib, C. Rohl, and K. Karplus. Pokefind: a novel topological filter for use with protein structure prediction. *Bioinformatics*, 25(12):281–288, 2009.

[35] G. H. Kruppa, J. Schoeniger, and M. M. Young. A top down approach to protein structural studies using chemical cross-linking and Fourier transform mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(2):155–62, 2003.

[36] I. Kwaw, J. Sun, and H. Kaback. Thiol cross-linking of cytoplasmic loops in lactose permease of escherichia coli. *Biochemistry*, 39:3134–3140, 2000.

[37] C.J. Langmead and B.R. Donald. An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Biomol. NMR*, 29:111–138, 2004.

[38] C.J. Langmead and B.R. Donald. High-throughput 3D structural homology detection via NMR resonance assignment. In *Proc. CSB*, pages 278–289, 2004.

[39] C.J. Langmead, A. Yan, R. Lilien, L. Wang, and B.R. Donald. A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Comp. Bio.*, 11:277–298, 2004.

[40] G. Lin, D. Xu, Z.-Z. Chen, T. Jiang, and Y. Xu. A branch-and-bound algorithm for assignment of protein backbone NMR peaks. In *Proc. CSB*, pages 165–174, 2002.

[41] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Springer, 1998.

[42] I. Michalopoulos, G. Torrance, D. Gilbert, and D. Westhead. TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res.*, 32:251–254, 2004.

[43] G.T. Montelione, D. Zheng, Y.J. Huang, K. Gunsalus, and T. Szyperski. Protein NMR spectroscopy in structural genomics. *Nat. Struct. Biol.*, 7 Suppl:982–985, 2000.

[44] H.N.B. Moseley and G.T. Montelione. Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.*, 9:635–642, 1999.

[45] J. Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.*, 15(3):285–289, 2005.

[46] S. Nelson, D. Schneider, and A.J. Wand. Implementation of the main chain directed assignment strategy. *Biophys. J.*, 59:1113–1122, 1991.

[47] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997.

[48] A.G. Palmer III, J. Williams, and A. McDermott. Nuclear magnetic resonance studies of biopolymer dynamics. *J. Phys. Chem.*, 100:13293–13310, 1996.

[49] G. Pandurangan. On a simple randomized algorithm for finding a 2-factor in sparse graphs. *Information Processing Letters*, 95(1):321–327, 2005.

[50] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Patt. Anal. Machi. Intel.*, 27(8):1226–1238, 2005.

[51] J. Plesnik. The NP-completeness of the Hamiltonian cycle problem in planar digraphs with degree bound two. *Information Processing Letters*, 8(4):199–201, 1979.

[52] J.L. Pons and M.A. Delsuc. RESCUE: An artificial neural network tool for the NMR spectral assignment of proteins. *J. Biomol. NMR*, 15:15–26, 1999.

[53] S. Potluri, A. Yan, J. Chou, B. Donald, and C. Bailey-Kellogg. Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space using NMR restraints and van der Waals packing. *Proteins*, 65:203–219, 2006.

[54] S. Potluri, A. Yan, B. Donald, and C. Bailey-Kellogg. A complete algorithm to resolve ambiguity for inter-subunit NOE assignment in structure determination of symmetric homo-oligomers. *Protein Science*, 16:69–81, 2007.

[55] J.H. Prestegard, H. Valafar, J. Glushka, and F. Tian. Nuclear magnetic resonance in the era of structural genomics. *Biochemistry*, 40:8677–8685, 2001.

[56] P. Pristovek, H. Ruterjans, and R. Jerala. Semiautomatic sequence-specific assignment of proteins based on the tertiary structure–the program st2nmr. *J. Comp. Chem.*, 23:335–340, 2002.

[57] M.G. Rossman and D.M. BLow. The detection of sub-units within the crystallographic asymmetric unit. *Acta. Cryst.*, 15:24–31, 1962.

[58] L. Saftalov, P.A. Smith, A.M. Friedman, and C. Bailey-Kellogg. Site-directed combinatorial construction of chimaeric genes: General method for optimizing assembly of gene fragments. *Proteins*, 64:629–642, 2006.

[59] R. Samudrala and J. Moult. A graph-theoretic algorithm for comparative modeling of protein structure. *J Mol. Biol.*, 279(1):287–302, 1998.

[60] B.R. Seavey, E.A. Farr, W.M. Westler, and J. Markley. A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, 1:217–236, 1991.

[61] 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.

[62] D.L. Di Stefano and A.J. Wand. Two-dimensional ^{1}H NMR study of human ubiquitin: a main-chain directed assignment and structure analysis. *Biochemistry*, 26:7272–7281, 1987.

[63] K. Suhre and YH. Sanejouand. ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res.*, 32:610–614, 2004.

[64] O. Vitek, C. Bailey-Kellogg, B. Craig, P. Kuliniewicz, and J. Vitek. Reconsidering complete search algorithms for protein backbone NMR Assignment. *Bioinformatics*, 21:ii230–236, 2005.

[65] O. Vitek, C. Bailey-Kellogg, B. Craig, and J. Vitek. Inferential backbone assignment for sparse data. *J. Biomol. NMR*, 35:187–208, 2006.

[66] O. Vitek, J. Vitek, B. Craig, and C. Bailey-Kellogg. Model-based assignment and inference of protein backbone nuclear magnetic resonances. *Statistical Applications in Genetics and Molecular Biology*, 3:article 6, 1–33, 2004. `http://www.bepress.com/sagmb/vol3/iss1/art6/`.

[67] L. Wang and B.R. Donald. An efficient and accurate algorithm for assigning nuclear overhauser effect restraints using a rotamer library ensemble and residual dipolar couplings. In *Proc. CSB*, pages 189–202, 2005.

[68] C. H. Wu, R. Apweiler, A. N. A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research*, 34(suppl_1):D187–191, 2006.

[69] F. Xiong and C. Bailey-Kellogg. A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. In *Proc. IEEE BIBE*, pages 403–410, 2007.

[70] F. Xiong, A. Friedman, and C. Bailey-Kellogg. Planning combinatorial disulfide cross-links for protein fold determination. In *Proc. BIOKDD*, 2010. to appear.

[71] F. Xiong, G. Pandurangan, and C. Bailey-Kellogg. Contact replacement for NMR resonance assignment. In *Proc. ISMB*, pages 205–213, 2008.

[72] J. Xu and B. Berger. Fast and accurate algorithms for protein side-chain packing. *J. of ACM*, 53(4):533–557, 2006.

[73] J. Xu, F. Jiao, and B. Berger. A tree-decomposition approach to protein structure prediction. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, pages 247–256, 2005.

[74] J. Xu, M. Li, D. Kim, and Y. Xu. RAPTOR: Optimal protein threading by linear programming, the inaugural issue. *J Bioinform Comput Biol*, 1(1):95–117, 2003.

[75] Y. Xu and D. Xu. Protein threading using PROSPECT: Design and evaluation. *Proteins*, 40:343–354, 2000.

[76] Y. Xu, D. Xu, O.H. Crawford, J.R. Einstein, and E. Serpersu. Protein structure determination using protein threading and sparse NMR data. In *Proc. RECOMB*, pages 299–307, 2000.

[77] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proc. of Int. Conf. on Data Mining (ICDM'02)*, pages 721–724, 2002.

[78] X. Ye, A. Friedman, and C. Bailey-Kellogg. Optimizing Bayes error for protein structure model selection by stability mutagenesis. In *Proc. CSB*, pages 99–108, 2008.

[79] X. Ye, P. O'Neil, A. Foster, M. Gajda, J. Kosinski, M. Kurowski, A. Friedman, and C. Bailey-Kellogg. Probabilistic cross-link analysis and experiment planning for high-throughput elucidation of protein structure. *Protein Science*, 13:3298–3313, 2004.

[80] M.M. Young, N. Tang, J.C. Hempel, C.M. Oshiro, E.W. Taylor, I.D. Kuntz, B.W. Gibson, and G. Dollinger. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *PNAS*, 97:5802–5806, 2000.

[81] M. Zaki, J. Hu, and C. Bystroff. *Data Mining: Next Generation Challenges and Future Directions*, chapter Methods for Mining Protein Contact Maps, pages 291–314. AAAI/MIT Press, 2004.

[82] A. Zemla, C. Venclovas, J. Moult, and K. Fidelis. Processing and analysis of CASP3 protein structure predictions. *Proteins*, Suppl. 3:22–29, 1999.

[83] Y. Zhang and J. Skolnick. The protein structure prediciton problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences (PNAS)*, January 25:1029–1034, 2005.

[84] D.E. Zimmerman, C.A. Kulikowsi, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J. Mol. Biol.*, 269:592–610, 1997.