Extracting Structural Information Using Time-Frequency Analysis of Protein NMR Data

Christopher James Langmead* and Bruce Randall Donald*†

October 26, 2000

Abstract

High-throughput, data-directed computational protocols for *Structural Genomics* (or *Proteomics*) are required in order to evaluate the protein products of genes for structure and function at rates comparable to current gene-sequencing technology. To develop such methods, new algorithms are required that can quickly extract significantly more structural information from sparse experimental data. This paper presents a new class of signal processing algorithms for nuclear magnetic resonance (NMR) structural biology, based on time-frequency analysis of chemical shift dynamics.

A novel approach to multidimensional NMR analysis is proposed in which the data are interpreted in the time-frequency domain, as opposed to the traditional frequency domain. Time-frequency analysis (TFA) exposes behavior orthogonal to the magnetic coherence transfer pathways, thus affording new avenues of NMR discovery. An implementation yielding new biophysical results is discussed. In particular, we demonstrate the heretofore unknown presence of through-space inter-atomic distance information within ^{15}N-edited heteronuclear single-quantum coherence (^{15}N HSQC) data. A biophysical model explains these results, and is supported by further experiments on simulated spectra.

Submitted to RECOMB'2001 (International Conference on Computational Molecular Biology)

*Dartmouth Computer Science Department, Hanover, NH 03755, USA

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

1 Introduction

Molecular biology is undergoing a transition towards high-throughput methods. Advances in a variety of different technologies are enabling this transformation. Microarray technology, for example, allows massively parallel high-throughput gene-expression experiments. Consequently, microarrays have revolutionized modern genetics. Advances in structural genomics methods would enable a similarly radical change in structural biology and proteomics. Unfortunately, protein structure determination remains a costly and time-consuming endeavor. Nuclear Magnetic Resonance (NMR) is one of two experimental techniques for determining atomic-resolution structures of biological macromolecules. Standard NMR protocols require running many separate experiments. A given experiment can take hours to days of spectrometer time and it can take weeks to months to prepare a protein sample needed for a sophisticated experiment (e.g., residue-specific isotopic labelings). Once the data has been collected it all must be carefully assigned, analyzed and consolidated. This process can take months and requires many tedious, manual steps. Due to the many steps in NMR discovery, advances in many subproblems are required to develop high-throughput methods for NMR structural biology. Automating the manual steps of NMR data assignment and analysis will be one advance [1-5]. Reducing the amount of spectrometer and wet-lab time by reducing the number of required experiments will be another [1, 2, 6]. Our work focuses on developing new algorithms that can quickly extract significantly more structural information from sparse experimental data. In this paper, we introduce and analyze a new class of signal processing algorithms for NMR structural biology, based on time-frequency analysis of chemical shift dynamics.

Our algorithms leverage the time-varying behavior of NMR data to extract useful information. This permits the algorithms to extract more information from NMR data than traditional methods. In particular, we describe how Time-Frequency Analysis (TFA) can be employed to observe and quantitate *Chemical Shift Dynamics* (CSD). We demonstrate that CSD can be analyzed using TFA to extract important, and heretofore unobserved structural information, from NMR experiments. Our algorithm demonstrates the utility of higher-order statistics (in particular, polyspectral analysis and the bicoherence spectrum) for protein NMR, bringing new data analysis tools to the armamentarium of the structural biologist. CSD are rich in structural and dynamic information, and yet they have never been previously exploited. TFA allows us to decode the information locked in CSD. The CSD TFA protocol effectively defines a new class of NMR experiment. Our work shows that the information content of NMR data (in general) and the ^{15}N HSQC (in particular) is much higher than previously believed. Furthermore, since the ^{15}N HSQC is perhaps the simplest, cheapest, and fastest heteronuclear NMR experiment, our method may have applications in high-throughput structural genomics. We present the experimental results of applying our algorithms on two protein NMR data sets from (1) human glutaredoxin, which plays an important role in maintenance of the redox state of the cell as well as in DNA biosynthesis and (2) core-binding factor, a heteordimeric transcription factor involved in hematopoesis. Oncogenic translocations in CBF-α and -β are implicated in acute myelomonocytic leukemia.

We now summarize the potential application of our work in high-throughput NMR methods for structural genomics. Generalizing the JIGSAW protocol [1,2], four spectra (the ^{15}N-edited HSQC, 3D ^{15}N-NOESY, 80 ms. ^{15}N-TOCSY, and HNHA) from a uniformly ^{15}N-labeled protein would be acquired in a few days. JIGSAW would then be employed to perform backbone resonance assignments and calculate secondary structure including β-sheets [1,2]. Next, we wish to constrain and calculate the global fold in a high-throughput manner. The HSQC can then be reanalyzed (as described in this paper) to reveal correlations in the CSD TFA ^{15}N-HSQC. CSD TFA yields structural constraints (distance correlations) that, together with the secondary structure and backbone amide proton assignments from JIGSAW, can be interpreted as distance restraints to calculate an approximate global

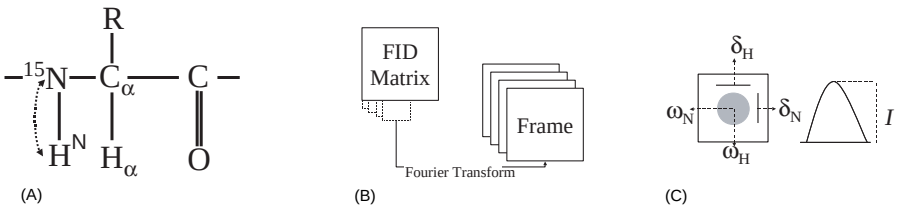

Figure 1: (A) The HSQC Magnetic Transference Pathway (dotted line). Magnetization is transfered, through-bond from the amide proton, to the amide nitrogen and back. (B) Windowing. In Time Frequency Analysis short segments of the full data, called *windows*, are extracted and analyzed separately. In this example, each row is an individual FID. A window is a subset of the columns in the full FID matrix. Each window is Fourier Transformed (after scaling and padding) yielding a single spectral frame. (C) Peak Features. For each peak we extract the frequency of the center of the peak (ω_H and ω_N), the width of the peak (δ_H and δ_N) and the intensity, or height of the peak I.

fold. The above set of four experiments requires only days of spectrometer time, rather than the months required for the traditional suite of dozens of experiments. Furthermore, the proposed protocol only requires a protein to be ^{15}N-labeled, a much cheaper and easier process than ^{13}C labeling. From a computational standpoint, we adopt a minimalist approach, demonstrating the large amount of information available in a few key spectra. While JIGSAW is used as an example, our method for CSD TFA is actually independent of JIGSAW: alternatively, other high-throughput assignment strategies could be employed [e.g., 3-5], along with secondary structure predictors [e.g., 7, 8] or other NMR methods for rapid secondary structure determination [9].

We begin, in Section 2, with a review of the theory and practice of NMR spectroscopy and discuss the implications of protein dynamics on quantum systems. Section 3 details our method for extracting time-varying behavior from NMR data. In section 4 we introduce methods for analyzing time-varying NMR data. Section 5 presents the results of the application of TFA to the raw HSQC data for human glutaredoxin and CBF-β. Finally, section 6 discusses these results and introduces a biophysical model to explain them.

2 NMR Data

Correlations in nuclear spin angular momentum are manifested as resonant peaks in NMR spectra. The location of these peaks in frequency space is measured as *chemical shift*. Multidimensional NMR spectra capture interactions between atoms as peaks in $\mathbb{R}^2$ or $\mathbb{R}^3$, where the axes indicate resonance frequencies or *chemical shifts* of atoms. In a typical ^{15}N spectrum peaks correspond to an ^{15}N atom, an H$^{\mathrm{N}}$ atom, and possibly another ^{1}H atom, of particular resonance frequencies. A peak occurs when atoms *interact*. Atoms interact via quantum magnetic coherence transfer either through covalent bonds, or through space.

Traditional NMR structure-determination protocols call for a number of different experiments. Each experiment gives qualitatively different kinds of information. NMR experiments fall into two categories: those (such as NOESY) that transfer magnetization *through-space* and those (such as HSQC) that transfer magnetization *through-bond*. Through-space interactions are caused by the Nuclear Overhauser Effect (NOE) which falls off with r^{-6} [10] and is essentially zero beyond 6 Å. Consequently, NOEs are typically employed to derive distance restraints among pairs of protons. Through-bond experiments are used to derive several different kinds of information, although in general, not distance restraints. For example, the ^{15}N HSQC is a two-dimensional through-bond experiment correlating the amide proton with the amide ^{15}N of the same residue [11] (Fig. 1 A). The HSQC is typically used to determine and pairwise correlate the chemical shifts of the amide protons and nitrogens along the backbone of the protein. These correlations establish the H$^{\mathrm{N}}$-^{15}N connectivities, and the backbone chemical shifts are subsequently used as reference points within

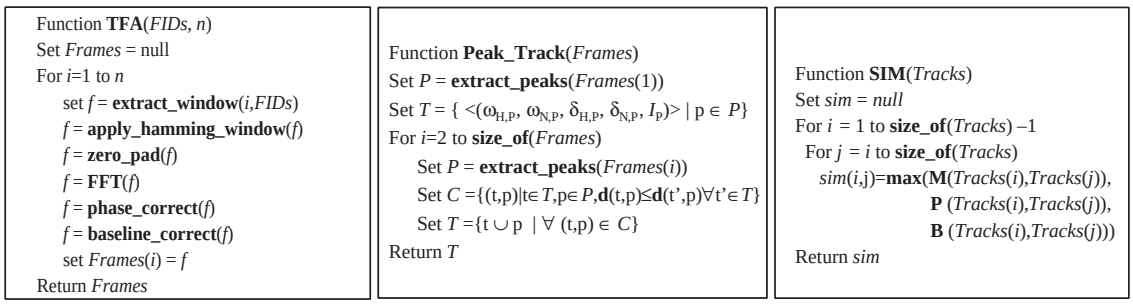

Figure 2: (A) TFA Algorithm. (B) Peak Tracking. A track is a list of temporally sequential peaks. The peak tracking algorithm creates an initial set of tracks from the first frame. Peaks from subsequent frames are appended to the track with a peak that is closest in frequency, shape and intensity. (C) Similarity Measurements. A similarity matrix is generated using the maximum similarity between tracks i and j under the M, P, and B similarity metrics (equations 2-4)

other spectra.

The precise location of an NMR peak in frequency-space is determined by a number of factors. Each atom-type has an inherent chemical shift. For example, in "isolation", all hydrogen atoms would have the same chemical shift. This fundamental frequency is modulated upfield or downfield via shielding by the electron clouds of nearby atoms. Within an amino acid (monopeptide), these shielding interactions are systematic and repeatable. That is, in a test tube of a given amino acid (e.g., Lysine) in solution, the amide proton for each monopeptide will have the same chemical shift. In a large protein, sequential interactions and the shielding of atoms brought into spatial proximity due to secondary and tertiary structure also significantly affect the chemical shift of a given nucleus.

2.1 Chemical Shift Dynamics

Proteins tend to be flexible and in solution, are constantly undergoing small conformational changes. Since chemical shifts are affected by tertiary structure [12-14], we must conclude chemical shifts are in fact dynamic (time-varying). We will refer to the phenomena as *Chemical Shift Dynamics* (CSD).

Molecular motion occurs simultaneously at many different time-scales spanning many orders of magnitude [15]. Some of these time scales are within the Nyquist frequency defined by NMR sampling rates. Consequently, CSDs are, in principle, observable by NMR. Furthermore, it is reasonable to hypothesize that CSD reflect structural properties and are therefore, worthy of examination.

Interestingly, chemical shift is typically viewed as a static property. This is in large part due to tools employed in traditional NMR analysis. A NMR spectrometer records a series of time-domain signals, know as *Free Induction Decays* (FIDs). A given atom's chemical shift is encoded as a periodicity within the FIDs. It is obtained by applying a Fourier Transform to the FIDs. FIDs, being time-domain signals, are capable of encoding CSD. However, it is not possible to observe CSD using the Fourier Transform because it integrates over time. The primary contribution of this paper is application of Time-Frequency Analysis (TFA) to extract CSD from NMR data.

We note that CSD are a different phenomena than traditional NMR dynamics (e.g., ^{15}N- and ^{13}C-relaxation rates for molecular mobility studies, their interpretation via the 'model-free' formalism to obtain generalized order parameters, or amide proton exchange measurements)[16-23]. Hence, we will show that CSD contain complementary information to traditional NMR dynamics protocols.

3 Time-Frequency Analysis

Our algorithm for extracting CSD from NMR data is summarized in Figure 2 A.The details of each step are discussed in the following subsections. We will focus on the application of TFA to the

^{15}N-edited HSQC in our examples, but TFA can be applied to the data from any NMR experiment.

3.1 Data Acquisition and Preprocessing

The data acquisition and preprocessing steps are the same for TFA and traditional methods. A sample is placed in the spectrometer and a series of pulse sequences are applied. At the end of each pulse sequence a signal is recorded —this is the FID. A two-dimensional NMR experiment, such as the ^{15}N-edited HSQC, involves the acquisition of a sequence of FID's with increasing T_1 times, resulting in a two-dimensional FID matrix. Once the data are collected they are subjected to a number of preprocessing steps. Typical transformations include noise-reduction and water-suppression. After the preprocessing, the traditional technique would apply a single 2-dimensional Fourier Transform to the data to obtain the NMR spectrum.

3.2 Short-Time Fourier Analysis

The primary distinction between traditional NMR analysis and TFA is the use of the Short-Time Fourier Transform (STFT) [24] . The STFT is a standard method for analyzing time-varying signals. Whereas the Fourier Transform takes as input the entire FID data set to produce a single spectrum, the STFT takes as input successive, overlapping temporal *windows* of the FID matrix to produce multiple spectra (Fig. 1 B).

There is an inherent trade-off between frequency and temporal resolution when applying TFA. In summary, the smaller the input window the higher the temporal resolution but the lower the frequency resolution. To a certain extent, one can compensate for lower frequency resolution by zero-padding the data prior to analysis and increasing the amount of temporal overlap (in data points) between windows. Our input window size was 128 data points. We used maximally overlapping windows so that we could generate as many spectral frames as possible. When windowing data, it is crucial to apply a scaling function to the window. Failure to do so results in spectral artifacts. We applied a Hamming window scaling function to each window and then padded the data with zeros just prior to spectral analysis.

The output of TFA is, in essence, a movie —a time-series of spectral *frames.* *Qualitatively*, a single frame from a TFA looks very similar to the traditional ^{15}N HSQC spectrum. *Quantitatively* however, there are differences due to the fact that frames are localized in time.

After spectral analysis, traditional methods usually apply phase and base-line correction to the spectra. We applied both phase and baseline correction to each of the TFA spectra frames [32].

3.3 Peak Picking and Feature Extraction

The next step in either the traditional or the TFA method is to locate and characterize the resonant peaks within the spectra. This can be done manually or automatically. We utilized the automatic peak picking capabilities of the program NMRPipe [25] to locate the peaks in each frame. In addition to locating the position of each peak in frequency space, the NMRPipe peak picker also extracts a number of other *features* from each peak. In our experiments we utilized 5 features: the peak's amide-proton and ^{15}N chemical shifts (ω_H and ω_N), amide-proton and ^{15}N line-widths (δ_H and δ_N), and intensity (I) (Fig. 1 C).

3.4 Peak Tracking

Once the peak picking and feature extraction are completed, the next step is to trace the evolution of each peak through time (Fig. 2 B). We call this trajectory a *track.* The input to the peak tracking algorithm are the individual peak lists, one for each spectral frame. For each frame, a greedy algorithm

matches a peak in frame i with the peak in frame $i+1$ whose 5 features most closely match its own. If no such peak exists then the track is labeled as "terminated". All matchings are unique. That is, no peak from frame $i+1$ is paired with more than one peak from frame i. The output of the peak tracker is a set of tracks. Each track encodes a trajectory in a five dimensional space. Alternatively, one can think of a track as a $5 \times N$ matrix where N is the number of frames. We call this matrix the *track matrix*. Each track corresponds to a single peak in the traditional ^{15}N HSQC spectrum When the assignments of these peaks to specific (H^N, ^{15}N) pairs are known, we can assign each track as well.

4 CSD Analysis

TFA is primarily a means for observing CSD. Analysis of CSD, we will show, yields relevant biological information. We've stated that protein motion gives rise to CSD. Differences in track dynamics may be due to differences in the molecular dynamics of various parts of the protein. If this is true, then there is information encoded in CSD. Specifically, if we can find sets of tracks that are temporally correlated, it might indicate something about the atoms associated with those tracks. For this reason, we chose to explore the notion of *similarity* among pairs of tracks.

4.1 Track Similarity Measurements

Different similarity measurements emphasize different properties of the tracks. The molecular dynamics which give rise to CSD are varied, complex and typically unknown at the time of NMR analysis. For these reasons, we implemented three different track similarity measurements, each targeting a different kind of information [Fig. 2 C]. It is worth introducing and reviewing these metrics, since their application may be unfamiliar in this context. The use of the power spectrum to infer structural constraints from energetic similarity in chemical shift dynamics is novel. Our third similarity metric employs *higher-order statistics* (specifically polyspectral analysis and the bicoherence spectrum) [26] which have not been previously applied to any form of biopolymer NMR.

The first measurement, M, compares track morphology using the correlation coefficient. The second measurement, P, compares periodicities within the tracks using the *power spectrum*. The power spectrum of a signal is the square of the magnitude of its Fourier transform. It reveals the amount of energy present as a function of frequency. Two tracks experiencing similar periodicities will have similar power spectra. The final measurement, B, compares nonlinearities within the tracks using the *bicoherence spectrum* [26]. The bispectrum is a higher order statistic capable of detecting third-order correlations within a signal. It is often used to detect quadratic phase coupling, a specific type of non-linearity. It is defined as $B(\omega_1, \omega_2) = Y(\omega_1)Y(\omega_2)Y^*(\omega_1 + \omega_2)$ where $Y(\omega)$ is the Fourier transform and $Y^*(\omega)$ is its complex conjugate. The functions governing CSD are nonlinear. Thus, it is possible that tracks will exhibit quadratic phase coupling. Two tracks which are the product of the same non-linear process will have similar bispectra. The bicoherence is the normalized bispectrum. It is defined as

$$B_c(\omega_1, \omega_2) = \frac{Y(\omega_1)Y(\omega_2)Y^*(\omega_1 + \omega_2)}{\sqrt{|Y(\omega_1)Y(\omega_2)|^2 \, |Y^*(\omega_1 + \omega_2)|^2}} \tag{1}$$

The bispectrum has previously been utilized in a number of domains to extract information from the higher-order statistics of natural data [e.g., 26, 27].

We say that two tracks are *correlated* if their similarities exceed a chosen threshold under any of the three similarity measurements, otherwise they are *uncorrelated*. Let C denote the set of pairs of

correlated tracks and let U denote the set of pairs of uncorrelated tracks. C and U are disjoint and the set $C \cup U$ is the set of all pairs of tracks. Note that the cardinality of C, and consequently U, is determined by the chosen threshold.

Prior to calculating similarities between pairs of tracks, each profile is normalized to the range $[-1, 1]$. The similarity between two tracks are only computed over temporally coincident frames. The M, P, and B similarity measurements are calculated as follows. Let X and Y be two track matrices. Let x_{ω_H}, x_{ω_N}, x_{δ_H}, x_{δ_H} and x_I denote the rows of X, corresponding to the chemical shift, line-widths and intensity profiles of X, respectively. Note that x_{ω_H}, for example, is a vector of N ω_H-values, one for each frame.

Our M similarity measurement is defined as

$$M(X,Y) = \left(r\left(x_{\omega_H}, y_{\omega_H}\right), r\left(x_{\omega_N}, y_{\omega_N}\right), r\left(x_{\delta_H}, y_{\delta_H}\right), r\left(x_{\delta_N}, y_{\delta_N}\right), r\left(x_I, y_I\right)\right) \tag{2}$$

where $r(x, y)$ is the correlation coefficient of vectors x and y.

Our P similarity measurement is defined as

$$\begin{aligned} P(X,Y) \quad = \quad & \left(r\left(H\left(x_{\omega_H}\right), H\left(y_{\omega_H}\right)\right), r\left(H\left(x_{\omega_N}\right), H\left(y_{\omega_N}\right)\right),\right. \\ & r\left(H\left(x_{\delta_H}\right), H\left(y_{\delta_H}\right)\right), r\left(H\left(x_{\delta_N}\right), H\left(y_{\delta_N}\right)\right), \\ & \left. r\left(H\left(x_I\right), H\left(y_I\right)\right)\right) \end{aligned} \tag{3}$$

where $H(x)$ is the power spectrum of the vector x, and $r(H_1, H_2)$ is the correlation coefficient of the power spectra H_1 and H_2.

Our B similarity measurement is defined as

$$\begin{aligned} B(X,Y) \quad = \quad & \left(r\left(B_c\left(x_{\omega_H}\right), B_c\left(y_{\omega_H}\right)\right), r\left(B_c\left(x_{\omega_N}\right), B_c\left(y_{\omega_N}\right)\right),\right. \\ & r\left(B_c\left(x_{\delta_H}\right), B_c\left(y_{\delta_H}\right)\right), r\left(B_c\left(x_{\delta_N}\right), B_c\left(y_{\delta_N}\right)\right), \\ & \left. r\left(B_c\left(x_I\right), B_c\left(y_I\right)\right)\right) \end{aligned} \tag{4}$$

where $B_c(x)$ is the bicoherence of the vector x, and $r(B_{c1}, B_{c2})$ is the correlation coefficient of the bicoherences B_{c1} and B_{c2}.

The similarity measurements are in the range $[-1, 1]$. Each similarity measurement (M, P, B) is multidimensional (one dimension per feature) and a separate threshold was selected for each dimension. The master threshold for a given similarity measurement is adjusted by maintaining the relative positions of the thresholds for the individual dimensions. The global similarity measurement takes the maximum similarity under M, P and B. The correlated pairs from each of M, P and B are combined to create the the final, correlated set. We are presently exploring analytical methods for determining thresholds based on the distributions of similarities observed under a given measurement/dimension.

5 Results

Our technique has been applied to the raw, two-dimensional ^{15}N HSQC FID matrices from the two proteins Human Glutaredoxin (huGrx) [*28*] and Core Binding Factor Beta (CBF-β) [*29*]. The sizes of the the two proteins are 106 and 143 residues respectively. We were provided the original ^{15}N HSQC FID data, signal processing parameters, and original peak lists for each protein by Dr. John Bushweller. ^{15}N HSQC spectra were recorded at Dartmouth on a 500 MHz Varian UnityPlus spectrometer with an actively shielded gradient triple resonance probe and pulsed field gradients at 20°C and at 30°C for CBF-β and huGrx, respectively, in 5% D_2O. In our experiments we utilized signal processing parameters similar or identical to those used in [*28*, 29] when possible.

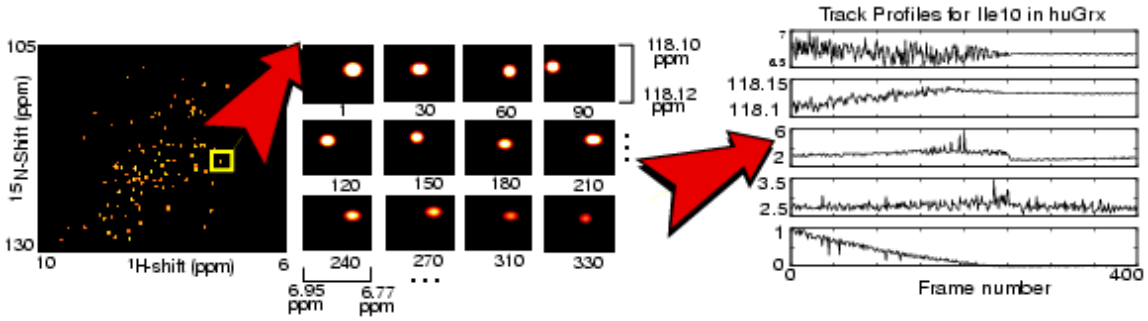

Figure 3: The track corresponding to the amide proton and ^{15}N of Ile10 in huGrx. A single frame from the TFA is seen on the left. The peak corresponding to Ile10 is outlined. The twelve smaller frames detail the behavior of that peak through time. The numbers under each image indicate the frame number it was taken from. Each of these details were taken from the same region in frequency space (118.1 ppm to 118.12 ppm on the ω_N axis, 6.77 ppm to 6.95 ppm on the ω_H axis). The full profiles for this track are seen on the right hand side of the figure. Each panel depicts the profile of a single feature (From top to bottom: ω_H, ω_N, δ_H, δ_N and I). In each panel the x-axis is the frame number.

(A) Track Statistics

	Protein	
	CBF-β(ppm)	huGrx (ppm)
Mean Δ Chem. Shift	0.16	0.17
Max Δ Chem. Shift	0.78	0.63
Min Δ Chem. Shift	0.07	0.07
St. Dev. Δ Chem Shift	0.09	0.07

(B) Inter Atomic Distance Statistics

	huGrx		CBF-β	
	C (Å)	U (Å)	C (Å)	U (Å)
Mean	11.02	17.07	11.90	22.26
Median	9.59	16.56	12.09	21.34
Max	23.45	40.76	21.27	53.68
Min	3.45	1.85	1.91	1.80
Pairs	23	8187	21	19001
t-test	$p < 1.8 \times 10^{-5}$		$p < 7.6 \times 10^{-7}$	

Table 1: (A) Summary of track statistics for CBF-β and huGrx. Δ chemical shift is calculated as the difference between the highest and lowest proton chemical shift value in each track. (B) Inter-atomic distance statistics for the distribution of temporally correlated peaks (C) vs. uncorrelated peaks (U) in huGrx and CBF-β. The number of pairs of protons in each distribution is also reported. Student's t-test confidence scores (p-values) reflect the probability the differences in means are due to chance.

5.1 Observability of CSD

A representative track is presented in Fig. 3 A number of reasons suggest that the dynamics exhibited in the tracks are not merely spectral artifacts. First, we note that each track's intensity (I) exhibits the expected decay predicted by the Bloch equations [30]. Second, the measured length of each track closely matches the published T2 times (within 4% for CBF-β, within 5% for huGrx). Third, a typical peak moves in a range of about 0.2 ppm [Table 1 A] which is small enough to be consistent with the change in chemical shift due to structural flexibility [31] yet too large to be explained by errors in estimating a peak's position —NMRPipe estimates the numerical error in localizing a peak in frequency space. In our experiments that error is, on average 0.01 ppm —an order of magnitude smaller than the changes we observe in the chemical shift profiles of our tracks. Thus, CSDs cannot be attributed to measurement error alone.

Of course, an FID is a composite of the individual signals emitted from the atoms in solution. A track exhibits an aggregate of individual behaviors rather than the behavior of a single atom. However, it is reasonable to assume that each molecule in solution has roughly the same structure and therefore the same capacity for motion. Consequently, corresponding atoms from different molecules experience similar variations in their electronic environment. Averaged over all the molecules in solution, the tracks associated with atoms in the vicinity of especially mobile regions of the protein should have

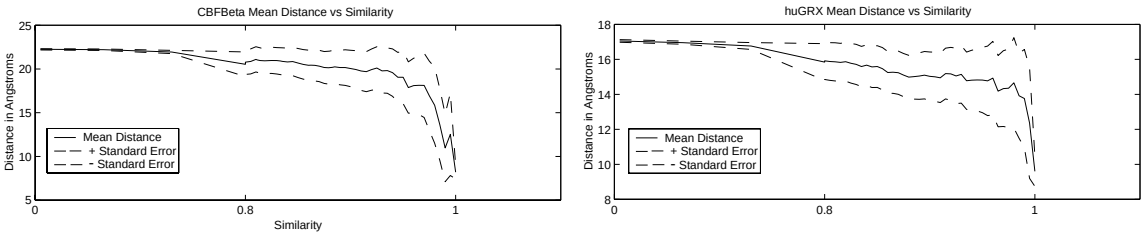

Figure 4: Mean inter-atomic distance vs. similarity for CBF-β(left panel) and huGrx (right panel) . The data-points were obtained by sweeping the similarity threshold from 0 to 1 and computing the mean inter-atomic distance for the set C corresponding to that threshold. The data-point at the far left comprises all pairs of protons. The point at the far right comprises only those pairs of protons with highest similarity. To avoid an unfair skew in the mean, a proton and itself (i.e., similarity =1.0, distance = 0.0 Å) are *not* included in any computed C. The similarity scale is non-linear to highlight the drop in distance at high levels of similarity. Above 0.8 we observe a steeper drop-off. The dashed lines are positive and negative standard error measurements.

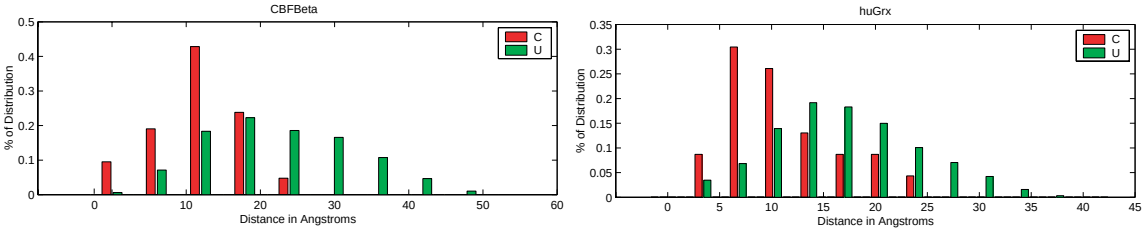

Figure 5: Normalized histograms of distances observed for computed C's and U's for CBF-β (left panel) and huGrx (right panel) as reported in Table 1 B. In both panels, C is shifted significantly towards 0. The height of a bar indicates the percentage of the total population within that range.

characteristics different from those associated with relatively rigid regions. By extension, two flexible regions undergoing different kinds of motion (e.g., periodic, but at different frequencies) will give rise to tracks with different properties.

5.2 Information Content of CSD

We know the peak assignments for each protein, so it is possible to identify the amide proton associated with each track. Furthermore, the 3-dimensional structures of the two proteins are known, so it is possible to validate the calculated similarities in terms of biophysical properties. We calculated the track similarity for all pairs of tracks. We discovered that the graph of cumulative means of inter-atomic, ^{1}H-^{1}H, distances, sorted by increasing track similarities, exhibits a negative correlation (Fig. 4) . That is, for sufficiently high similarity thresholds, the mean inter-atomic distance of set C is smaller than the mean inter-atomic distance of set U.

As the means of the distributions C and U diverge (with higher and higher thresholds) they reach a point where the difference becomes statistically significant according to Student's t-test. Above this range we can adjust the cardinality of C while maintaining a statistically significant difference in the means. Detailed statistics are included in the appendix (Fig. 7) for the interested reader. The t-test assumes that the two distributions are normal with equal variances. Our variances were not equal so we applied the standard log-transformation to the distributions to equalize the variances.

We then selected a threshold that maximizes the distance between the means of C and U when the cardinality of C is between twenty and thirty (Table 1 B, Fig. 5). The statistical significance of these differences in means was verified (Student's t-test, huGrx: $p < 1.8 \times 10^{-5}$; CBF-β: $p < 7.6 \times 10^{-7}$). We conclude that our track similarity measurement has a significant bias towards picking proton pairs that are close in space. Of particular interest is that the distributions reported in Table 1 B. and Fig. 5 include a high percentage of long-range interactions. These long range interactions are analogous to the NOE distance restraints obtained from NOESY spectroscopy [32]. A graphical

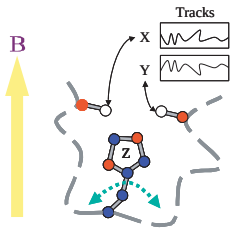

Figure 6: Molecular motion-based model for observed relationship between spatial proximity and temporal correlation of CSD. The dark-gray dashed line represents an arbitrary sequence of residues in the peptide chain. Circles and lines represent atoms and bonds. Protons X and Y are both proximal to Z, a flexible substructure of the protein. As Z moves (in this example back and forth along an arc) its influence on the respective chemical shifts of X and Y changes. The result is a coordinated change (in this example, anti-correlated) of the chemical shifts of X and Y. B is the applied magnetic field.

depiction of these restraints along the backbones of the two proteins are included in the appendix (Fig. 8) for the interested reader.

Under our M, P, and B similarity measurements not all spatially proximate proton pairs are found to be in C. This behavior also parallels NOESY spectroscopy in which not all close ^{1}H-^{1}H pairs appear. However, NOESY peak intensity is correlated with inter-atomic distance in a roughly r^{-6} fashion. In contrast, the degree of M, P, and B-similarity cannot currently be used to quantitate through-space distance. We are presently exploring more sophisticated similarity measurements that may yield quantitative bounds on distance.

It is worth emphasizing that the distance information obtained in our experiments is unexpected. The ^{15}N HSQC, unlike a NOESY, is not supposed to contain *any* inter-atomic distance information. Indeed, it is specifically designed to *prevent* transference of magnetization between anything but the ^{15}N and ^{1}H from each amide group. The key advantage to TFA, however, is that it reveals atomic properties unrelated to transference pathways. We also note that because the mean distance correlations we observe are considerably larger than the 6 Å, they are not explainable by any residual or unsupressed NOE.

TFA of HSQC data is not intended to replace standard NOESY experiments. Rather, it demonstrates that there is more information in NMR data than previously believed. Indeed, TFA can be applied to NOESY data as well. TFA may be used to supplement traditional NMR spectra. Several common problems, such as peak overlap and peak matching within and across spectra, may benefit from an analysis of the time-varying behavior of the data.

6 Comparison of experimental results to theoretical models

Consider the following simplified model in Fig. 6 . Suppose protons X and Y are both near some region of the protein Z. Z is close enough to X and Y to have some influence on their chemical shifts (e.g. via electronic shielding). Now suppose that Z is part of a flexible region. As Z moves, X and Y's chemical shifts will change. If Z's motion has similar influence (i.e., upfield or downfield) on X and Y, then their tracks will have morphological similarities. Furthermore, if Z's motion is periodic, X and Y's tracks will be periodic and therefore have similar power spectra and/or bispectra. Of course, X and Y may themselves be part of (independent) flexible sub-domains. Their individual chemical shifts may reflect the combined influence of multiple Z's plus tumbling and solvent interactions. However, insofar as our model is concerned, these additional factors will yield more complex CSD but the possibility of detecting correlations remains. In such cases, a multi-dimensional similarity measurement, such as the one presented here, will increase the chances of finding correlations.

Z can only influence the chemical shifts of atoms within a fixed radius [33]. When this radius can be estimated, upper bounds on the distance between temporally-correlated tracks can be calculated and

applied quantitatively in a manner analogous to NOE's. Note that under this model, the conditions necessary to produce temporal correlations between pairs of tracks are quite restrictive. In particular, it does not predict that all pairs of close atoms will be temporally correlated.

6.1 Comparison to Simulation of Chemical Shifts in Mobile Protein Domains

We tested our model with simulated spectra of proteins in which we simulate the molecular dynamics over time. In the first simulation we created a time-series from twenty PDB files describing distinct, but similar, low-energy conformations of CBF-β derived from traditional NMR structural techniques. In the second simulation we used the time-series generated from the ten PDB files of hemoglobin (HGN) and Che-Y protein (CHY) as obtained from the database of molecular motions [15]. Using the program SHIFTS [34] we simulated the chemical shifts for each proton in each of the PDB files describing the motion of the molecules. SHIFTS takes as input a PDB file and estimates proton chemical shifts from empirical formulas. The result is analogous to TFA of real NMR data but not identical. A key distinction is the length of the simulated tracks. Ten and twenty data-point tracks are too sparse to perform meaningful spectral analysis so we only considered the morphological similarity (M) of the tracks. The pairwise track similarities under the M similarity measurement were calculated. Two filters were applied to the similarity matrix. The first filter ignores any single track whose chemical shift profile range is below a minimum threshold. In other words, we ignored tracks that were essentially flat. The second filter ignored pairs of tracks whose respective CSD ranges were wildly different. That is, we did not compare a track with high activity with one with low activity. The reason is that atoms experiencing wildly different ranges of CSD are unlikely to be nearby. A threshold was applied to the filtered matrix. The inter-atomic distances of the tracks above the threshold were examined. As in the experiments on real NMR data, there is a statistically significant difference (Student's t-test, CBF-β: $p < 1.8 \times 10^{-9}$; HGN: $p < 4.0 \times 10^{-3}$; CHY: $p < 5.1 \times 10^{-7}$) between the means of correlated and uncorrelated tracks. Detailed statistics are included in the appendix (Table 2) for the interested reader.

7 Conclusion

The application of TFA to NMR data is appropriate because 1) NMR data are inherently time-varying, and 2) CSD have the potential to yield more information about the local electronic environment than the corresponding time-averaged chemical shift. We have shown that it is possible to observe CSD in one class of NMR experiment. The chemical shifts of the atoms in any flexible protein are dynamic. Therefore, TFA is applicable to any NMR experiment with suitable time-resolution. The ^{15}N HSQC is one such experiment. Applying the techniques presented here to other experiments is an obvious extension. One can imagine further enhancing the observability of CSD by manipulating the factors affecting protein flexibility (e.g., solution temperature).

We have also shown that CSD contain structural information. In particular, our results demonstrate that ^{15}N HSQC CSD contain through-space inter-atomic distance information. The model we used to explain the relationship between temporal correlation and inter-atomic distance does not preclude finding this information in other NMR experiments as well. Adapting the techniques presented here to other NMR experiments will permit the kind of cross-validation typical in NMR discovery.

The extraction of inter-atomic distance is not the only potential application of TFA. It might be used to confirm, or provide an alternative means for obtaining, standard NMR measurements (e.g., T2 times). The identification and classification of flexible regions within biological macromolecules, peak separation in dense NMR spectra, and peak matching across spectra are all exciting directions for future work.

References

1. C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. Beradi, J. Bushweller, B. Donald, "The NOESY Jigsaw: Automated Protein Secondary Structure and Main-Chain Assignment from Sparse, Unassigned NMR Data, *4th Annual Intl. Conf. on Comp. Mol. Biol.,* Tokyo, Japan, pp. 33-44, (2000)

2. C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. Beradi, J. Bushweller, B. Donald, "The NOESY Jigsaw: Automated Protein Secondary Structure and Main-Chain Assignment from Sparse, Unassigned NMR Data, J. Computational Biology, 2000 (accepted; in press).

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

4. C. Bartels, P. Guntert, M. Bileter, and K. Wuthrich. GARANT— a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comp. Chem.*, 18:139-149, 1997

5. D. Croft, J. Kemmink, K.-P. Neidig, and H. Oschkinat. Tools for the automated assignment of high-resolution three-dimensional protein NMR specrta based on pattern recognition techniques. *J. Biomol. NMR*, 10:207-219, 1997

6. Y. X. Lin and G. Wagner, Efficient side-chain and backbone assignment in large proteins: Application to tGCN5, *J. Biomol. NMR*, **15** 227-239, 1999.

7. J.A. Cuff, M.E. Clamp, A.S. Siddiqui, M. Finlay, and G.J. Barton. JPRED: A consensus secondary structure prediction server. *Bioinformatics,* 14:892-893. 1998.

8. G. Dealeage, B. Tinland, and B. Roux. A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins. *Analytical Biochemistry*, 163(2):292-297, June 1987

9. D.S. Wishart, B.D. Sykes, and F.M. Richards. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647-1651, February 1992

10. K. Wüthrich, *NMR of Proteins and Nucleic Acids.* (John Wiley & Sons, 1986).

11. J. Cavanagh, W.J. Fairbrother, A.G. Palmer, N.J. Skelton, *Protein NMR Spectroscopy: Principles and Practice.* 416-418 (Academic Press Inc., 1996).

12. Sitkoff, D., & Case, D. Density Functional Calculations of Proton Chemical Shifts in Model Peptides. *J. Am. Chem Soc. 119,* 12262-12273 (1997)

13. Dejaegere, A. P., Case, D. Density Functional Study of Ribose and Deoxyribose Chemical Shifts, *J. Phys. Chem. A **102**,* 5280-5289 (1998)

14. Moravetski, V., Hill, J. R., Eichler, U., Sauer, J. ^{29}Si NMR Chemical Shifts of Silicate Species: Ab Initio Study of Environment and Structure Effects, *J. Am. Chem. Soc.* **118**, 13015-13020 (1996)

15. M. Gerstein and W. Krebs, *Nucleic Acids Research* **26** (1998).

16. Wagner, G., NMR relaxation and protein mobility, *Current Opinion in Struct. Biol.*, **3**, 748-754 (1993)

17. Palmer, A. G., Williams, J., & McDermott, A. Nuclear Magnetic Resonance Studies of Biopolymer Dynamics, *J. Phys. Chem.*, **100**, 13293-13310 (1996)

18. Lipari, G. & Szabo, A., Model-Free approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity, *J. Am. Chem. Soc.*, **104**, 4546-4559 (1982)

19. Palmer, A. G., Dynamic properties of proteins from NMR spectroscopy, *Current Opinion in Biotechnology.*, 4, 385-391 (1993)

20. Palmer, A. G., Probing molecular motion by NMR, *Current Opinion in Struct. Biol.*, **7**, 732-737 (1997)

21. Lipari, G. & Szabo, A., Model-Free approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results, *J. Am. Chem. Soc.*, **104**, 4560-4570 (1982)

22. Kay, L., Protein Dynamics from NMR, *Nature Struct. Biol.* NMR Supplement, July (1998)

23. Palmer, A. G. & Bracken, C. *Spin Relaxation Methods for Characterizing Picosecond-nanosecond and microsecond-millisecond motions in Proteins*, in NMR in Supramolecular Chemistry, Pons, M. ed. 171-190, (1999 Kluwer Academic Publishers, Netherlands)

24. A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing* (Prentice Hall, 1989)

25. Delaglio, F. *et al,* NMRPipe: a multidimensional spectral processing system based on UNIX Pipes. *J. Biomol. NMR.* **6** (1995).

26. Mendel, J. M., Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications, *Proc. IEEE,* **79** 278-305 (1996).

27. H. Farid, Blind Inverse Gamma Correction, *IEEE Trans. on Image Processing* (to appear)

28. Sun, C., Holmgren, A. & Bushweller, J., Complete ^{1}H, ^{13}C, and ^{15}N NMR resonance assignments and secondary structure of human glutaredoxin in the fully reduced form, *Protein Science*, **6**, 383-390 (1997)

29. Huang, X., Speck, N. A. & Bushweller, J., Complete Resonance Assignments and secondary structure of core binding factor β, *J. Biom. NMR* **12**, 459-460 (1997).

30. Bloch, F., Hansen, W.W. & Packard, M., Nuclear Induction, *Phys. Rev.* **69**,127 (1946).

31. Wijmenga, S. S., Kruithof, M. & Hilbers, C. W., Analysis of ^{1}H chemical shifts in DNA: Assessment of the reliability of ^{1}H chemical shifts for use in structure refinement. *J. Bio NMR* **10,** 337-350 (1997)

32. J. Cavanagh, W.J. Fairbrother, A.G. Palmer, N.J. Skelton, *Protein NMR Spectroscopy: Principles and Practice.* 384-394 (Academic Press Inc., 1996).

33. Pearson, J. G., *et al* Predicting Chemical Shifts in Proteins: Structure Refinement of Valine Residues by Using *ab Initio* and Empirical Geometry Optimizations, *J. Am. Chem. Soc.* **119,** 11941-11950 (1997)

34. Osapay, K. & Case, D.A., Peptides, Chemistry, Structure and Biology, R.S. Hodges and J.A. Smith, eds. (Leiden: ESCOM, 1994), pp. 911-913.

35. Xu, Y. *et al,* Protein Structure Determination using Protein Threading and Sparse NMR Data, *4th Annual Intl. Conf. on Comp. Mol. Biol.,* Tokyo, Japan, pp. 299-307, (2000)

Acknowledgments

We thank Hany Farid and Chris Bailey-Kellogg for their assistance on the algorithmic and signal processing aspects of this work, John Bushweller, Chaohong Sun, and Xuemei Huang for generously providing us their NMR data, Marcelo Berardi, John Bushweller and Jack Kelley for their insights into NMR, Deborah Chiavelli and Jennifer Groh for feedback on the statistical aspects of this work and Fred Henle for his assistance in preparing this paper. This work is supported by the following grants from the National Since Foundation to B.R.D.: NSF IIS-9906790, NSF EIA-9901407, NSF EIA-9802068, NSF CDA-9726389, NSF EIA-9818299, NSF CISE/CDA-9805548, NSF IRI-9896020, NSF IRI-9530785, and by an equipment grant from Microsoft Research.

Appendix

The additional, optional information in this appendix is provided for the interested reader. Table 2 shows the detailed results of temporal similarity measurements on simulated NMR data. Figure 7 shows the relationship between the differences in inter-atomic mean distances between the sets C and U and the size of the correlated set C. Figure 8 shows ribbon diagrams of both test proteins and the distance restraints derived via TFA projected onto cartoons of those proteins.

	CBF-β		CHY		HGN	
	C (Å)	U (Å)	C (Å)	U (Å)	C (Å)	U (Å)
Mean	9.78	21.70	13.78	17.35	13.30	18.67
Median	4.84	20.46	13.58	17.40	13.21	18.65
Max	26.26	55.96	23.10	34.02	22.66	40.34
Min	2.96	2.58	2.87	2.63	2.83	2.60
Pairs	24	18608	22	7728	44	8867
t-test	$p < 1.8\times 10^{-9}$		$p < 4.0\times 10^{-3}$		$p < 5.1\times 10^{-7}$	

Table: 2 Inter-atomic distance statistics for the distribution of temporally correlated protons (C) vs. uncorrelated protons (U) in *simulated* CBF-β, CHY and HGN spectra. The simulation of CBF-β spectrum is based on twenty PDB files encoding NMR-derived low-energy conformations. These twenty low-energy conformations are derived from NMR data and were averaged to obtain the final, published structure of CBF-β. We consider these conformations to form an ergodic ensemble. That is, each conformation is drawn from some low-energy well in conformation space and any path through these conformations is equally likely. Consequently, we generated a time series by using each conformation once, in random order. We report the results of one random permutation of the conformations but tests with 100 other random permutations yield similar results. The simulation of the CHY and HGN spectra is based on ten PDB files (for each protein) depicting conformations generated by molecular dynamics simulation. Student's t-test confidence scores (p-values) reflect the probability the differences in means are due to chance. The effects of the smaller sample size for CHY and HGN relative to CBF-β are seen in the difference in means between C and U. The shorter series have less power for discrimination, but the statistical significance remains.

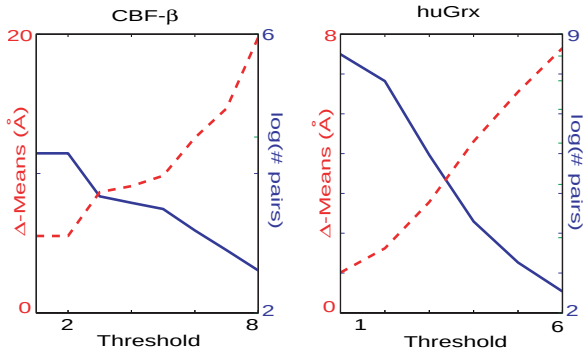

Figure: 7 Similarity Threshold vs. difference in means of C and U (solid line) and cardinality of C (dashed line) for CBF-β and huGrx. The x-axis is a normalized threshold over the multidimensional M, P, and B-similarity measurements. Within the range of thresholds presented here, the distributions C and U are statistically different (i.e., they pass a t-test). As the similarity threshold increases, the cardinality of C decreases and the difference in means increases.

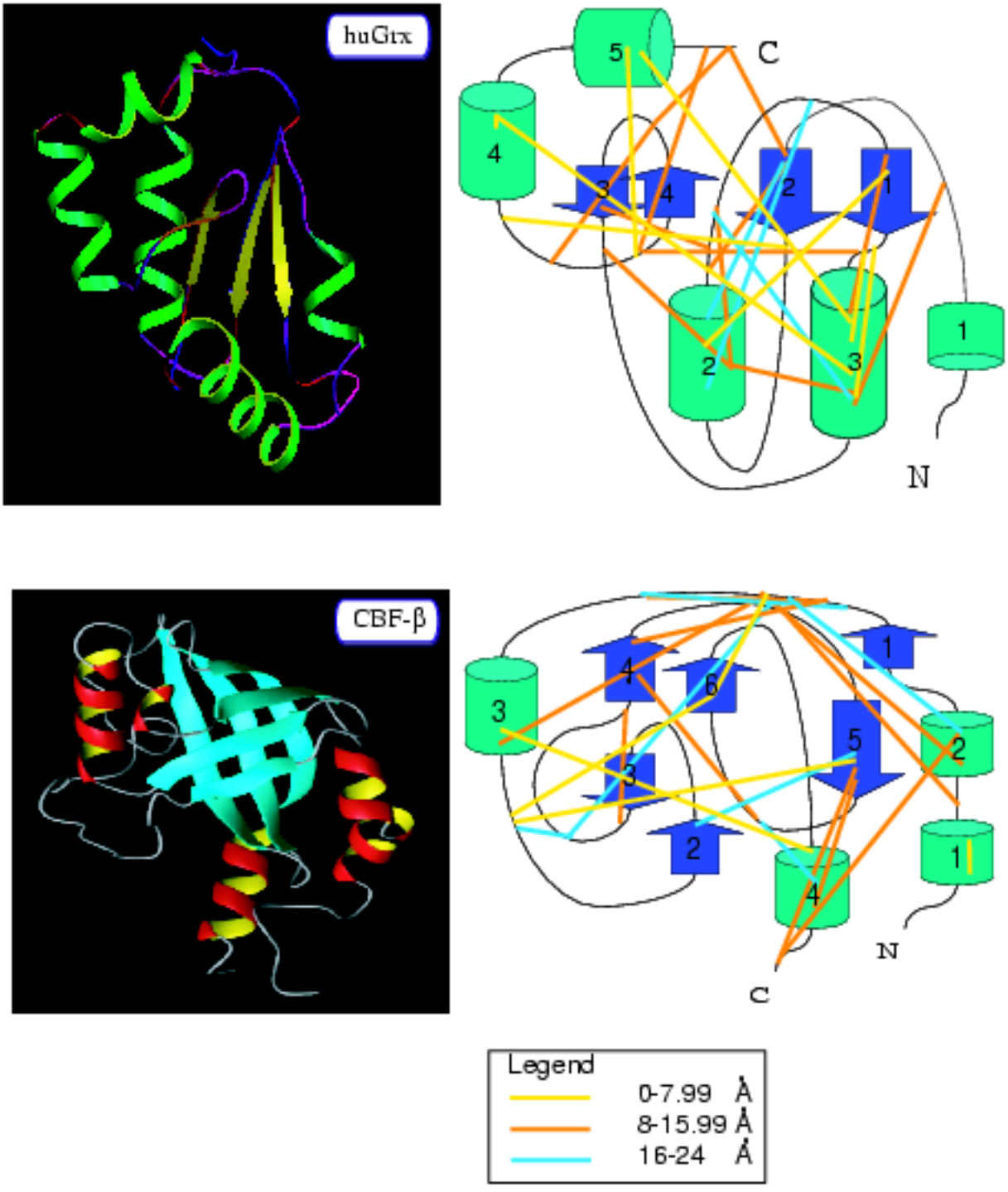

Figure: 8 This figure is best viewed in color. Please see `http://www.cs.dartmouth.edu/~langmead/recomb/fig8.jpg` Similarity pairings for huGrx and CBF-β. Lines connect pairs of atoms whose tracks exhibit temporal correlation. The color of the line indicates the actual distance between the two endpoints. The tertiary structure of each protein is shown on the left for reference in a similar spatial projection. These similarity pairings indicate long-range distance restraints and reflect the spatial proximity of different parts of the proteins. When coupled with a high-throughput assay for secondary structure determination [1,2] and a structure refinement algorithm designed for sparse distance constraints [35], an estimate of the protein's global fold can be obtained.