*Journal of Biomolecular NMR*, **12:** 231–241, 1998.
*KLUWER/ESCOM*



# Rotational-echo double-resonance in complex biopolymers: a study of *Nephila clavipes* dragline silk

Carl A. Michal* & Lynn W. Jelinski**
*Center for Advanced Technology, 130 Biotechnology Building, Cornell University, Ithaca, NY 14853, U.S.A.*

Received 14 October 1997; Accepted 12 February 1998

*Key words:* Protein structure, REDOR, spider silk, solid-state NMR

**Abstract**

Rotational-Echo Double-Resonance (REDOR) NMR on strategically $^{13}C$ and $^{15}N$ labeled samples is used to study the conformation of the LGXQ (X = S, G, or N) motif in the major ampullate gland dragline silk from the spider *Nephila clavipes*. A method is described for calculating REDOR dephasing curves suitable for background subtractions, using probability distributions of nitrogen atoms surrounding a given carbon site, which are developed from coordinates in the Brookhaven Protein Data Bank. The validity of the method is established by comparison to dephasings observed from natural abundance $^{13}C$ peaks for G and A. Straightforward fitting of universal REDOR dephasing curves to the background corrected peaks of interest provide results which are not self-consistent, and a more sophisticated analysis is developed which better accounts for $^{15}N$ labels which have scrambled from the intended positions. While there is likely some heterogeneity in the structures formed by the LGXQ sequences, the data indicate that they all form compact turn-like structures.

## Introduction

Rotational-Echo Double-Resonance (REDOR) NMR is a powerful method for determining conformations in biopolymers. When samples can be prepared with isolated pairs of isotopic spin labels, the method gives quantitative distance results. The distance information ($r$) derives from the dipolar coupling, which goes as $1/r^3$. Several examples now exist in the literature where REDOR has been used to elucidate structural details of biopolymers, including Gramicidin A, (Hing and Schaefer, 1993), *Saccharomyces cerevisiae* tridecapeptide mating pheromone (Garbow et al., 1994) and EPSP synthase (Li et al., 1994). In general, using $^{13}C$ and $^{15}N$ labels, distances out to ca. 4.5 Å can be determined to an accuracy of 0.1 Å, with even better accuracy for shorter distances.

The structural basis for the remarkable mechanical properties of *Nephila clavipes* drag-line silk (Gosline et al., 1986; Cuniff et al., 1994) has been the subject of several recent publications. A new model for the molecular-level structure of this material emerged from the results of solid-state deuterium NMR experiments on alanine-labeled oriented (Simmons et al., 1996) fibers. This model (Figure 1) is based on the hypothesis that the protein backbone forms a reverse turn at each LGSQ and LGNQ sequence. It should be noted that the LGXQ (X = S, G, or N) motif is remarkably well conserved between repeats in the amino acid sequence and that G-S residues are often found in bends. The model is attractive in that it correctly predicts the size of the crystallites observed by X-ray diffraction experiments. (Grubb and Jelinski, 1997). It furthermore suggests that the irregularities in the sequence (Xu and Lewis, 1990) are required to limit crystallite size. To test the correctness of the model in Figure 1, it is necessary to establish whether the LGSQ sequences do, in fact, assume compact bend structures in the solid silk fibers. To this end, we have measured key distances via REDOR.

In this work, we perform REDOR measurements on isotopically labeled samples of *N. clavipes* (major ampullate gland) dragline silk in order to test the

*Present Address: Bldg 5, Room 406, National Institutes of Health, Bethesda, MD 20892, U.S.A.
**To whom correspondence should be addressed.


*Figure 1.* Packing model proposed for major ampullate gland dragline silk from *N. clavipes*. The shaded regions are highly oriented crystallites of polyalanine in anti-parallel β-sheets. Reprinted with permission from *Science*, **271**, 84-87. Copyright 1996 American Association for the Advancement of Science.

proposed structural model. This paper begins with a description of the preparation and characterization of the samples, followed by a detailed discussion of a method for calculating REDOR dephasing curves suitable for background subtractions. This method is demonstrated by reproducing the dephasings observed on natural abundance carbon peaks of the silk spectra. We then discuss REDOR dephasing of the target carbon sites and show that straightforward fitting of distances to dephasing curves is not an adequately faithful model of the target environment. A more sophisticated simulation model better represents the actual environment of the $^{13}$C labels, and leads to the conclusion that the backbone of all of the LGXQ sequences form tightly packed turn-like structures.

*Labeling strategy*

The molecular degrees of freedom which would define a turn in a four residue segment are the $\phi$, $\psi$ torsion angles of the center two residues. With the choice of $^{13}$C and $^{15}$N as stable isotope labels, the sites most appropriate for labeling are the two carbon sites of leucine and the backbone nitrogens of glutamine and serine. Figure 2 shows the ideal placement of labels, assuming no scrambling occurs (we will later show that some scrambling takes place, and this is quantified and accounted for in the calculations). Labeling of sites in the glycine residues was avoided because glycine composes $\sim$ 45% of the silk (Zemlin, 1968), while leucine and serine residues appear infrequently outside of the LGXQ sequence (Lewis, 1992). The distances between the chosen labels, assuming standard bonding geometry, depend only on the four backbone torsion angles of the glycine and serine residues.

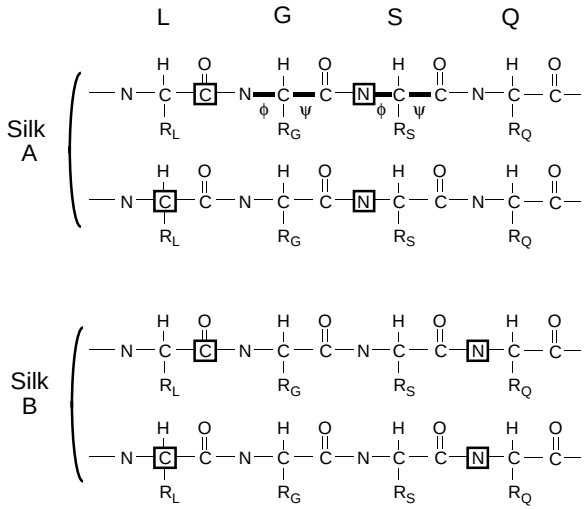


*Figure 2.* Isotope labeling strategy. Positions of the $^{13}$C and $^{15}$N labels are indicated by boxes. Ideally, Silk A should contain equal amounts of protein containing labels as shown by the boxes in the top two structures. Silk B is described by the bottom two structures.

## Materials and Methods

*Silk Samples*

Major ampullate gland dragline silk samples were harvested from *N. clavipes* spiders (central Florida) in a manner similar to that described by Work (Work and Emerson, 1982). Spiders were divided into two

groups, the first of which was fed Dubellco's Modified Eagle Medium (DMEM) (Gibco) which had been fortified with 1.3% (w/v) $^{15}$N serine, 0.56% (w/v) 1-$^{13}$C leucine and 0.56% (w/v) 2-$^{13}$C leucine. Spiders were hand fed before, during, and after silking and once per day on non-silking days. Silk was collected three times per week at an extraction rate of 2 cm/s for 45 min per session, with constant monitoring to ensure inclusion of only dragline fibers. Silk collected at the first two sessions contained little or no label and was discarded. A total of 57 mg of dragline silk was harvested from these 18 spiders and denoted Silk A.

The second sample was obtained similarly from 17 spiders which were fed DMEM fortified with the same two leucine labels, but with the serine replaced with 1.2% amine-$^{15}$N glutamine. This 54 mg sample was denoted Silk B.

*NMR Spectroscopy*

Solid-state $^{13}$C and $^{15}$N cross-polarization/magic-angle spinning (CP/MAS) spectra were acquired on a home-built spectrometer based upon an Oxford 8.4 T magnet, a Tecmag Libra pulse programmer and Doty Scientific narrow bore triple-resonance probe. All spectra in this work were acquired at a spinning speed of 5 kHz with a proton decoupling field strength of 90 kHz and 2.5 ms CP contact time. $^{13}$C spectra were externally referenced to the upfield peak of adamantane at 29.5 ppm while $^{15}$N spectra are referenced to the upfield peak of $NH_4NO_3$ at 0 ppm.

$^{13}$C detected, $^{15}$N dephased REDOR measurements of the two silk samples were made with dephasing times ranging from two rotor cycles (0.4 ms) to 136 rotor cycles (27.2 ms) with a 2 s repetition delay, 14 $\mu$s $^{15}$N $\pi$ pulses and 90 kHz proton decoupling.

To determine the extent of the $^{15}$N scrambling the samples were hydrolyzed in propionic acid/12 N HCl (50/50 v/v) at 130 °C for one hour. As well as degrading the protein to its amino-acid constituents, this process converts glutamine (gln) residues to glutamic acid (glu). The resulting free amino acids were dried and redissolved in $D_2O$ so that solution-state $^{13}$C spectra could be acquired. Nuclear Overhauser Effect (NOE) enhanced solution-state NMR spectra were obtained on a Varian VXR-400 spectrometer with a 1.2 s acquisition delay.

**Results and discussion**

*Characterization of samples*

The solid-state $^{13}$C CP/MAS spectrum of Silk A is shown along with that of a natural abundance sample in Figure 3. The $^{13}$C NMR spectrum of Silk B was indistinguishable from that shown of Silk A. By normalizing the spectra such that the alanine $C_\beta$ and glycine $C_\alpha$ peaks match and subtracting, it is determined that for each peak, 70% of the intensity in the labeled samples arises from the $^{13}$C labels. It is somewhat unexpected that these portions are identical for both the carbonyl and the $C_\alpha$ peak, as the background associated with the carbonyl peak is due to all other residues, while the background associated with the $C_\alpha$ peak does not include a contribution from glycine, as the glycine $C_\alpha$ peak is separately resolved. If we assume that leucine is 50% labeled in each of the two carbon sites, the known amino acid abundances of silk (Work and Young, 1987) predict that leucine should correspond to 67% of the carbonyl peak and 79% of the $C_\alpha$ peak. The discrepancy is most likely due to the fact that the alanine residues, which contribute the largest share of the $C_\alpha$ peak, reside in crystalline regions (Simmons et al., 1996) and are likely to cross-polarize more efficiently than the other residues. The fact that the alanine residues appear to contribute more intensity than their abundance would predict has been noted previously (Simmons et al. 1994). The solid-state $^{13}$C spectra are thus consistent with full (50% in each backbone site) $^{13}$C labeling of leucine residues with no scrambling of the $^{13}$C labels.

Solid-state CP/MAS $^{15}$N spectra of Silk A and Silk B are shown in Figure 4. While $^{15}$N chemical shifts are dependent on conformation (Le and Oldfield, 1994), the amide-$^{15}$N residue shifts are typically found in a narrow 30 ppm region from ~87–117 ppm, with glycine exclusively occupying the 3 ppm farthest upfield (Wishart et al., 1991). Typically the amide shift of each residue type may vary by 3–4 ppm depending on conformation (Wishart et al., 1991; Asakura et al., 1997). The breadth of these spectra thus suggests that some amount of scrambling of the $^{15}$N labels took place within the spider's metabolism.

High-resolution $^{13}$C NMR was used to establish the degree of $^{15}$N scrambling. The $C_\alpha$ regions of proton-decoupled NOE enhanced spectra are displayed in Figure 5. Each amino acid residue displays a main peak and a shoulder which is shifted by an $^{15}$N isotope effect by approximately 5.4 Hz upfield. The

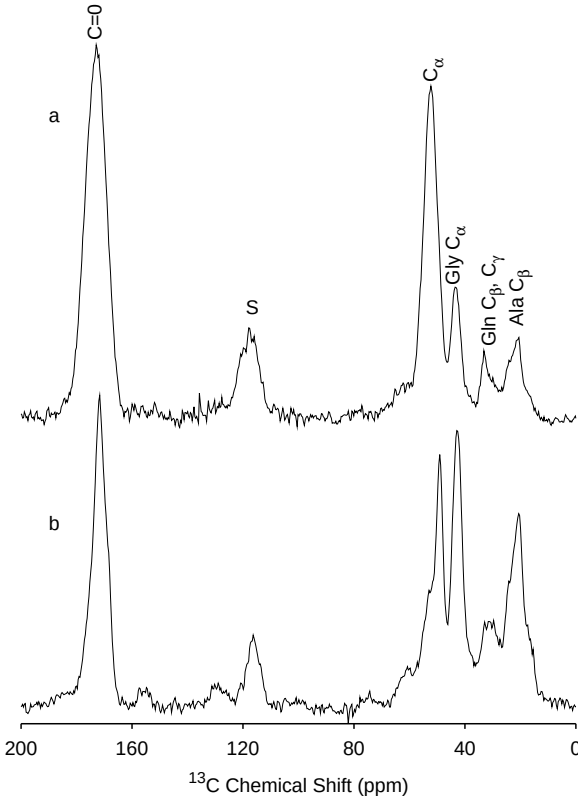


*Figure 3.* $^{13}$C CP/MAS Spectra of (a) leucine labeled Silk A and (b) natural abundance dragline silk. Assignments as in (Simmons et al., 1994), S indicates spinning sidebands.

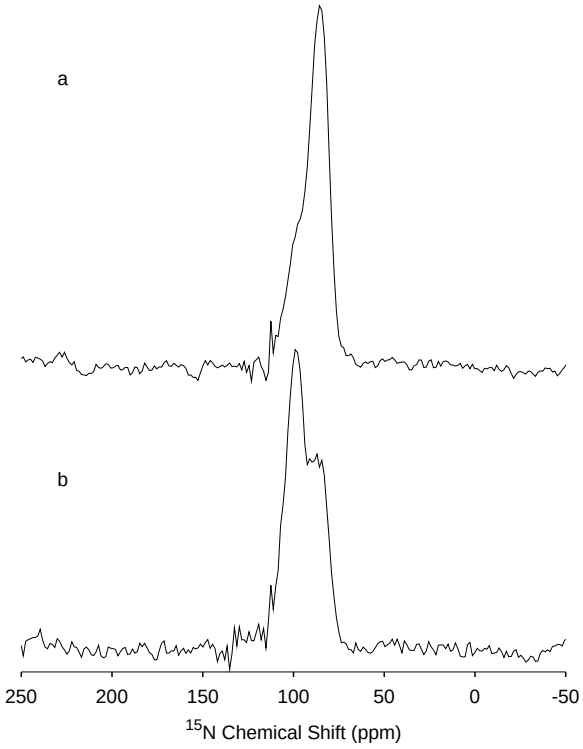


*Figure 4.* $^{15}$N Spectra of (a) Silk A (serine labeled) and (b) Silk B (glutamine labeled).

*Table 1.* $^{15}$N Labeling estimates and amino acid abundances from solution-state $C_\alpha$ peaks

| | % labeling Silk A | | | Silk B | | | abundance (%) from peak intensity Silk A | Silk B |
|---|---|---|---|---|---|---|---|---|
| gly | 30 | ± | 5 | 15 | ± | 7 | 51.7 | 47.6 |
| ala | 11 | ± | 4 | 14 | ± | 7 | 29.0 | 30.0 |
| leu | 7 | ± | 3 | 11 | ± | 7 | 2.8* | 3.0* |
| ser | 30 | ± | 5 | 20 | ± | 15 | 2.2 | 2.7 |
| glu | 12 | ± | 5 | 24 | ± | 7 | 10.7 | 12.1 |
| tyr | | | | | | | 2.1 | 1.6 |
| arg | | | | | | | 1.9 | 2.4 |

* amplitude of leucine peak was multiplied by 1.1/50.0 to adjust for the $^{13}$C labeling.

amplitude of the shifted shoulder in a glycine/$^{15}$N-glycine standard is suppressed by a factor of 2.2, and this correction was applied to obtain estimates of labeling in the silk samples. Although we do not understand this effect in detail, we believe that the suppression is due to differences in relaxation processes affecting $^{13}$Cs due to the different nitrogen isotopes (i.e., $^{14}$N or $^{15}$N), and will be the subject of further study.

$^{15}$N labeling estimates, along with amino acid abundances (assuming full $^{13}$C labeling of leucine, natural abundance of others) determined from these solution-state $C_\alpha$ peaks are compiled in Table 1. The amino acid abundances so determined are in excellent agreement with reports in the literature (Zemlin, 1968; Work and Young, 1987), supporting the conclusion that the leucine residues are fully labeled and that no scrambling of the $^{13}$C labels occurred. Conversely, the $^{15}$N labels have undergone substantial scrambling in the spider's metabolism. In each sample, the most highly $^{15}$N labeled residue was that which was fed with the $^{15}$N label, but in Silk A, the glycine appears as well labeled as the target serine, while in Silk B, there is almost uniform labeling with some enhancement of the glutamine and possibly serine. This is not surprising as glutamine is a principal biochemical source of nitrogen (Lehninger et al., 1993).

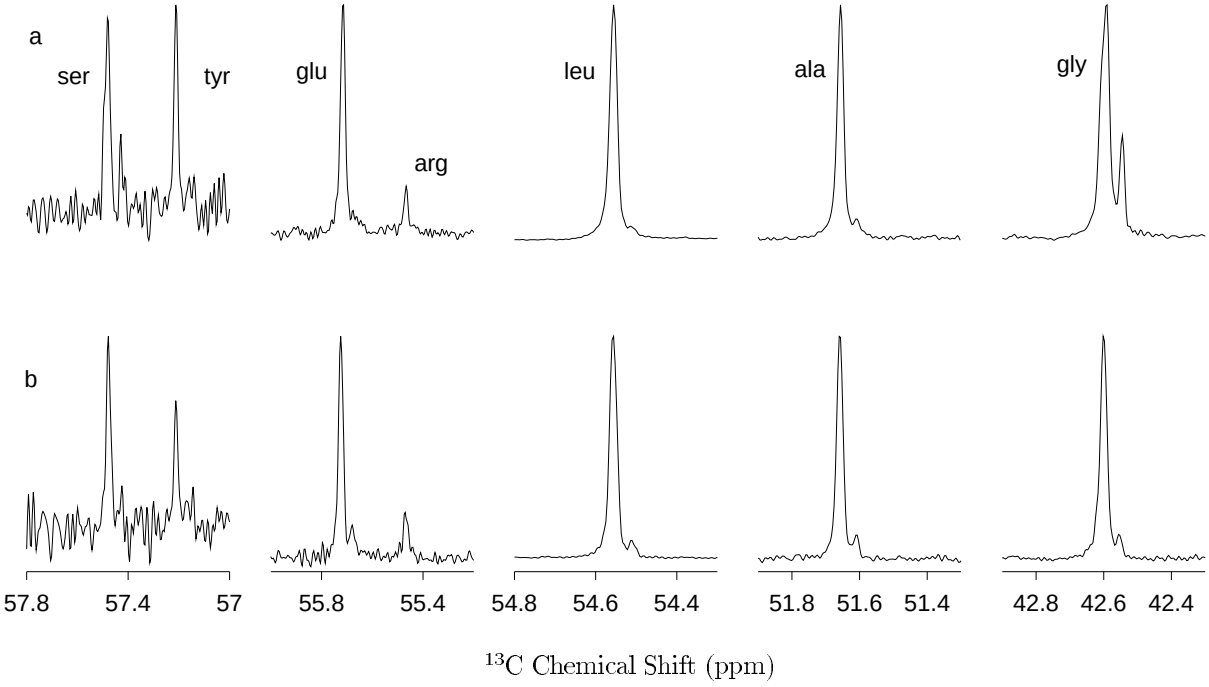


*Figure 5.* $C_\alpha$ regions of $^{13}C$ Solution-state spectra of hydrolyzed (a) Silk A and (b) Silk B.

*REDOR measurements and background subtractions*

*Natural Abundance Peaks*

Before addressing the REDOR dephasing curves from the two target (i.e., isotopically labeled) peaks, we first examine the natural abundance carbon signals of the alanine $C_\beta$ and glycine $C_\alpha$ sites. Analysis of the natural abundance peaks of these two most prevalent amino acids was performed (1) to confirm the amount of adjacent $^{15}N$ labeling reported in Table 1; and (2) to ensure that the REDOR experiments and distance calculation routines reproduce the known $C_\beta$-N and $C_\alpha$-N distances.

Dephasing curves constructed from these two peaks are displayed for Silk B in Figure 6. The solid lines are best fits incorporating two distances and associated fractions of the peak intensity. The simulations are actually superpositions of three curves, where the third curve represents $^{13}C$ nuclei which have both $^{15}N$ neighbors present. This third curve is calculated with a fast algorithm (Michal, 1997) which closely approximates the dephasing expected with two labeled neighbors. The curve resembles in shape the universal curve expected with just the nearest of the two neighbors, but does depend weakly on the angle between the two $^{13}C$-$^{15}N$ vectors. This angle is allowed to vary in the fitting routine, but because the fit is very insensitive to its value, and because this curve makes only a small contribution to the overall dephasing, no interpretation is applied.

*Table 2.* Distances and fractions of peak intensity from alanine and glycine peaks

| Distance (Å) | Fraction | Distance (Å) | Fraction |
|---|---|---|---|
| Silk A alanine $C_\beta$ | | Silk B alanine $C_\beta$ | |
| $2.6 \pm 0.2$ | $0.10 \pm 0.04$ | $2.39 \pm 0.1$ | $0.10 \pm 0.02$ |
| $4.0 \pm 0.3$ | $0.60 \pm 0.1$ | $3.77 \pm 0.1$ | $0.39 \pm 0.05$ |
| glycine $C_\alpha$ | | glycine $C_\alpha$ | |
| $1.51 \pm 0.05$ | $0.31 \pm 0.02$ | $1.47 \pm 0.2$ | $0.12 \pm 0.01$ |
| $3.5 \pm 0.3$ | $0.66 \pm 0.2$ | $3.9 \pm 0.3$ | $0.7 \pm 0.3$ |

The parameters found from the fitting routines are given in Table 2. The quoted uncertainties are derived from a Monte-Carlo algorithm where many synthetic data sets were produced using the best fit parameters and the original experimental errors. Each synthetic data set was fit with a new set of parameters and the standard deviation of the set of fits was taken as the uncertainty in the parameter. In each case, the shorter

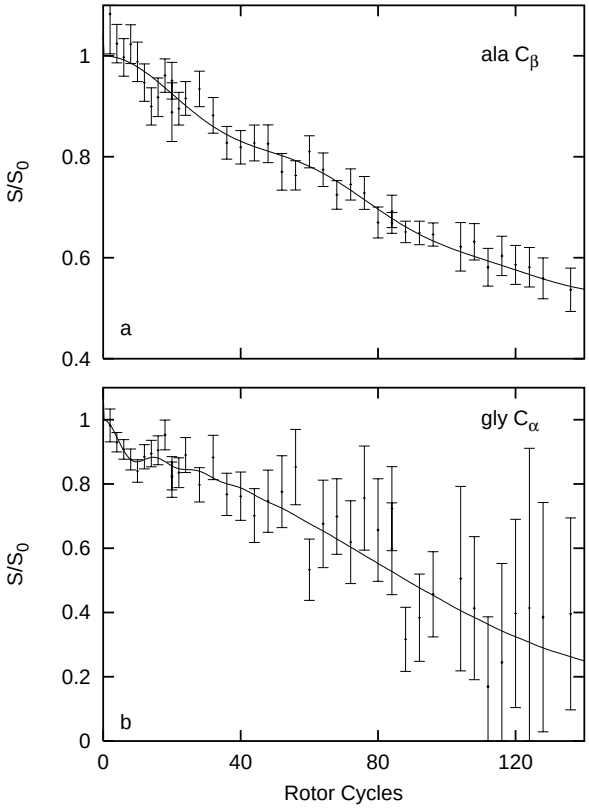


*Figure 6.* REDOR dephasing curves for natural abundance carbon peaks in Silk B. Experimental data shown along with best-fits incorporating nitrogen neighbors at two distances.

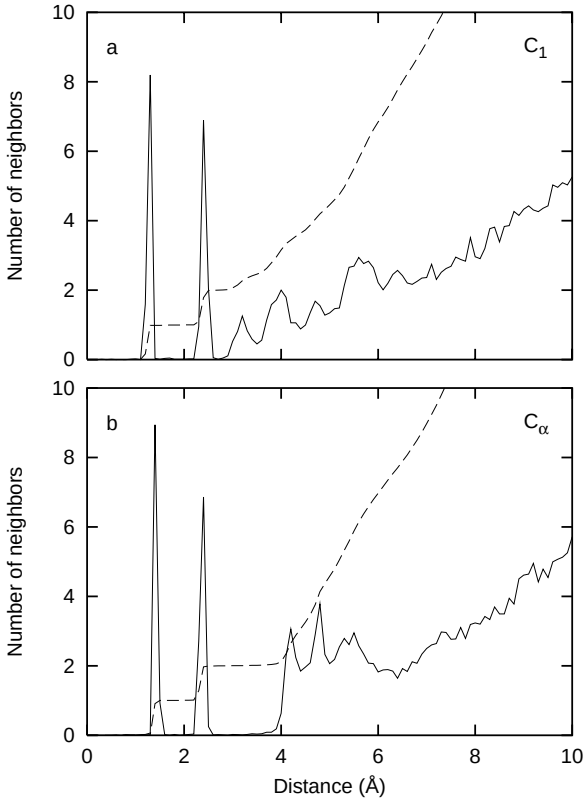


*Figure 7.* Probability distribution of finding nitrogen neighbors calculated from globular protein structures. Dashed curves represent total integrated number of nitrogen neighbors in sphere of the given radius.

distance is in excellent agreement with the distance expected to the residue's own nitrogen (2.45 Å for $C_\beta$, 1.45 Å for $C_\alpha$), and the associated fractions of peak intensity give $^{15}$N labeling estimates for these residues. These estimates are in agreement with those found from solution-state NMR (Table 1), and suggest that the alanine residues are ∼10% labeled in each sample, while the glycine residues are 30% labeled in Silk A and 12% labeled in Silk B. We suspect that these estimates are more reliable that those derived from the solution-state NMR spectra due to the uncertainties associated with overlap and relaxation there.

The interpretation of the longer distances in Table 2 is less transparent, and is called into question when it is noticed that the fractions contributing to these longer distances are larger than the fractions of $^{15}$N labeled of any one nitrogen site.

This issue is addressed by consideration of the spatial distribution of nitrogen atoms surrounding a given carbon site. Such probability distributions are shown for carbonyl and $C_\alpha$ sites in Figure 7. These distributions were derived from the structures of six globular proteins. Coordinates were obtained from the Brookhaven Protein Data Bank (PDB) (Abola et al. 1987; Bernstein et al., 1977) and the PDB identifiers of the proteins used are 1CSE (Bode et al., 1987), 1ARB (Oda et al. 1996), 2SN3 (Zhao et al., 1992), 7RSA (Svensson et al., 1986), 1FNC (Bruns and Karplus, 1995) and 1PTX (Housset et al., 1994). These proteins were chosen randomly from those used by Karplus (Karplus, 1996) and were used there because they met stringent criteria on the quality of the coordinate data. Specifically, each structure has ≤1.75 Å resolution, implying atomic positions accurate to better than 0.15 Å. The first two sharp peaks in each plot correspond to the C – nearest nitrogen (directly bonded distance) and the C – second neighbor nitrogen distances, respectively. The additional structure in these plots may be due to prevalent secondary structures.

Because the REDOR dephasing curve for a carbon with two $^{15}$N neighbors is in form very similar to that with only the nearest of those neighbors

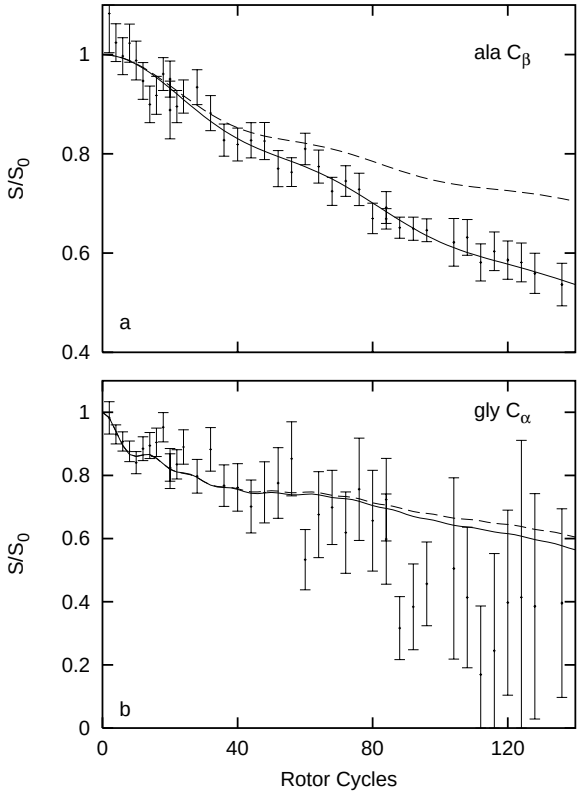


*Figure 8.* REDOR dephasing curves for natural abundance carbon peaks in Silk B. Data as in Figure 6, solid curves here are parameter-free dephasing curves as described in text.

(Michal, 1997), appropriate $^{15}$N labeling levels were used to calculate distributions of *nearest* $^{15}$N neighbors from these structures. From these distributions, parameter-free dephasing curves may be calculated. Such dephasing curves are shown along with the alanine methyl and glycine $C_\alpha$ dephasing data again in Figure 8.

The dashed lines in this figure are simulations made using nearest neighbor distributions derived from the globular protein structures. The nearest neighbor distributions assume labeling levels from the best estimates from the earlier fits (12% for glycine, 10% for alanine). The solid lines are similar calculations, using nearest neighbor distributions derived from a regular β-sheet crystal structure (Arnott et al., 1967). While the two curves are very similar for the backbone $C_\alpha$ (true for carbonyl carbons also, not shown), they are dramatically different for the alanine $C_\beta$. This is reasonable, as it is to be expected that the side chain, which is known to be in the β-sheet conformation, will be influenced more by the details of inter-chain packing, while nearby (in sequence) residues will be most important for the backbone sites.

The crystal structure for β-sheets (Arnott et al., 1967) mentions two possibilities for sheet layering, and includes a parameter which adjusts the height of adjacent layers. The neighbor distribution, and thus the resulting dephasing curve, appears to be very insensitive to these possible adjustments. The match between the alanine simulation and experimental data is extremely good, and corroborates other evidence that alanine residues reside in β-sheets.

Such parameter-free dephasing curves may also be calculated for the Silk A natural abundance sites (Michal, 1997), but their description is somewhat more complicated due to the much greater heterogeneity of the $^{15}$N labeling there.

*Leucine Peaks*

Analysis of the natural abundance alanine and glycine peaks (prior section) provides confidence that these parameter free dephasing curves may be accurately calculated. Similar curves appropriate to 30% of each target peak (which corresponds to natural abundance $^{13}$C) were calculated. The dephasing so contributed is subtracted off the data, and what remains is scaled to amount to 100%. The resulting data are displayed along with best fits in Figure 9. The fits shown were performed in a fashion similar to that described above for the natural abundance sites in alanine and glycine. In this case, each data set requires the incorporation of three distances. Each curve is the superposition of five dephasing curves, one for each of the three distances, and two curves accounting for the two more common occurrences of two nitrogen neighbors. Distances and fractions of leucine carbons determined by this fitting algorithm are shown in Table 3.

In each case, the two shorter distances correspond to distances determined by bonding geometry, and are compared with the standard distances in Figure 10. In general, the agreement of the measured distances with those expected from bonding geometry is excellent. The exception is the N-$C_\alpha$ single bond length determined for Silk A. The fit overestimates this distance slightly, which is not surprising due to the very small fraction of target carbon nuclei which contribute. The fractions found for the shorter distances are all consistent with the labeling estimates used in the background correction. In Silk A, the carbonyl one-bond distance and the $C_\alpha$ two-bond distance represent dephasing due to the nitrogen in the glycine which immediately follows each leucine. Both fractions agree well with the

*Table 3.* Distances and fractions of leucine signal from fitting to background corrected REDOR dephasing curves

| Distance (Å) | Fraction | Distance (Å) | Fraction |
|---|---|---|---|
| Silk A carbonyl carbon | | Silk B carbonyl carbon | |
| $1.39 \pm 0.01$ | $0.262 \pm 0.01$ | $1.36 \pm 0.02$ | $0.109 \pm 0.01$ |
| $2.59 \pm 0.1$ | $0.088 \pm 0.02$ | $2.31 \pm 0.1$ | $0.104 \pm 0.01$ |
| $3.73 \pm 0.07$ | $0.84 \pm 0.05$ | $3.49 \pm 0.04$ | $0.565 \pm 0.02$ |
| $C_\alpha$ | | $C_\alpha$ | |
| $1.57 \pm 0.2$ | $0.039 \pm 0.01$ | $1.48 \pm 0.06$ | $0.12 \pm 0.01$ |
| $2.36 \pm 0.04$ | $0.29 \pm 0.02$ | $2.59 \pm 0.2$ | $0.09 \pm 0.03$ |
| $3.87 \pm 0.2$ | $0.753 \pm 0.3$ | $3.70 \pm 0.2$ | $0.54 \pm 0.1$ |

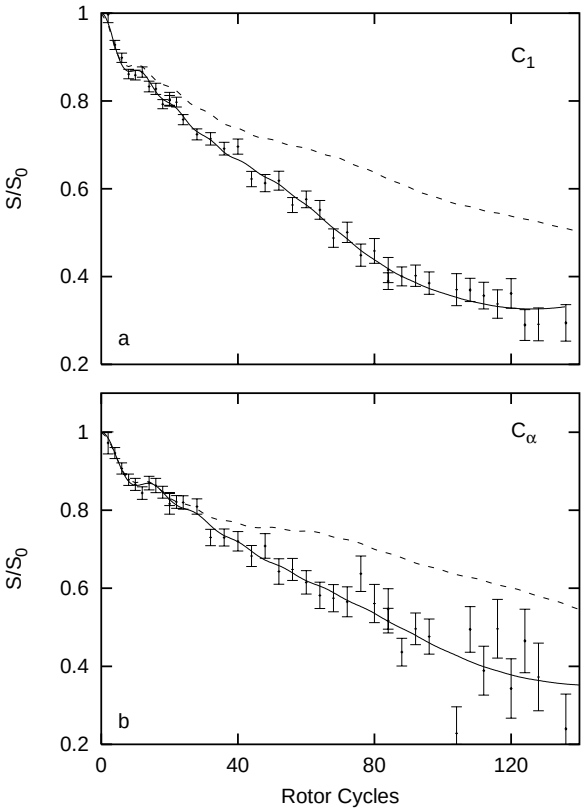


*Figure 9.* Background corrected dephasing curves for leucine peaks. Solid curves are best-fits incorporating nitrogen neighbors at three distances. The dashed curves are calculated using the known $^{15}N$ labeling levels and the nitrogen neighbor distributions derived from globular protein structures and demonstrate the fact that random protein structures yield very different dephasing curves from those observed here.

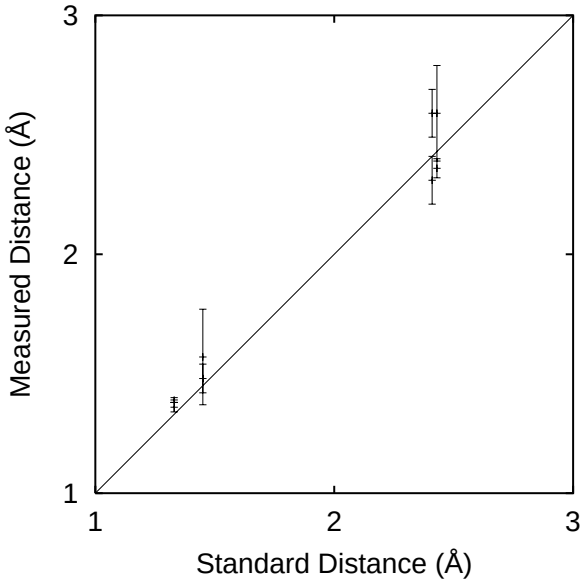


*Figure 10.* Correlation of one and two-bond distances determined from fits to leucine REDOR dephasing curves with standard geometry predictions.

$30 \pm 5\%$ expected labeling for glycine residues in this sample. The one-bond $C_\alpha$ and two-bond carbonyl distance are from $^{15}N$ nuclei within leucine residues. The fit values of 8.8 and 3.9% agree with the $7 \pm 3\%$ found from solution state spectra. In Silk B, all fractions for these close distances are in the neighborhood of 10%, in good agreement with the estimates used in the background subtraction.

### *More Sophisticated Fitting Using the 6-Residue Segment: GLGXQG*

The fact that the fractions of leucines which see nitrogen neighbors at the longest distances are greater than the fraction $^{15}N$ labeled of the most labeled residue (Table 3) implies that these distances are averages to more than one nitrogen neighbor. This observation leads to further consideration of the nitrogen probability distributions of Figure 7. These curves demonstrate that within a sphere of radius 5 Å of each backbone carbon site, there are, on average, about 4.5 nitrogen neighbors. Most of these 4.5 appear nearby in sequence. The fitting algorithm accounts carefully for the closest two of these, but gives a weighted average for all more distant neighbors. The simulation model used in this fitting algorithm thus does not do an adequate job of representing the actual environment of the carbon sites.

*Table 4.* Best-fit GLGXQG torsion angles

| | glycine | leucine | glycine | serine | glutamine | glycine |
|---|---|---|---|---|---|---|
| $\phi$ (°) | | −121 | −47 | −92 | −172 | |
| $\psi$ (°) | −23 | | −63 | −1 | 98 | |

To better account for the true environment of the carbon nuclei, we now consider the six-residue segment: GLGXQG (X = S, G or N). Of the 4.5 neighbors within 5 Å, all but an average 0.6 (carbonyl site, 0.8 for $C_\alpha$ site) occur within this six-residue segment. Using known bonding geometry and the known $^{15}N$ labeling levels, dephasing curves for the two carbon sites in each of the two samples may be calculated as a function of just eight backbone torsion angles.

All four leucine dephasing curves are shown along with the best-fits in Figure 11 and the torsion angles providing these fits are listed in Table 4. These torsion angles correspond to the structure displayed in Figure 12. While the dephasing curves presented here do not appear to fit as well as those of Figure 9, the model is much more faithful to the actual environment of the carbon sites. The fact that the simulations all rise above the experimental data at the longest dephasing times is due to the finite length of the amino acid sequence considered in this model. The addition of more $^{15}N$ neighbors at further distances will cause more dephasing at the longer times. The fact that the fit gives torsion angles for the glutamine and initial glycine residues which are slightly outside of the 'allowed' regions of the Ramachandran plot is not disturbing as the fitting routine has chosen conformations which place the nitrogens from the edges of the six-residue segment artificially nearer to the leucine carbons to make up for those belonging to residues from outside of this six-residue sequence.

The goodness of fit of these final simulations is striking, when compared to dephasing curves calculated with the globular protein neighbor distributions with known labeling levels (dashed curves in Figure 9).

*Comparison of REDOR results with known secondary structures*

The best-fit structure determined in the previous section is compared with a variety of standard secondary structures in Table 5. The figure of merit listed is the rms difference between the standard structure and

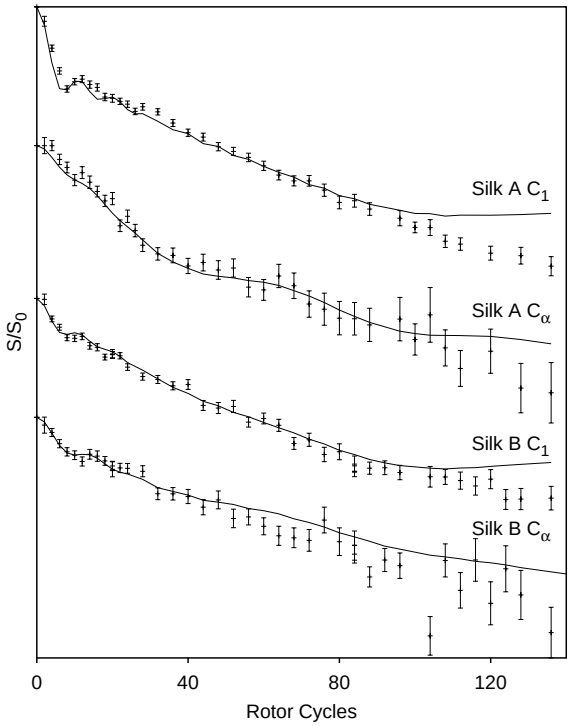


*Figure 11.* Background corrected dephasing curves for leucine peaks shown with best-fits from six-residue fitting.

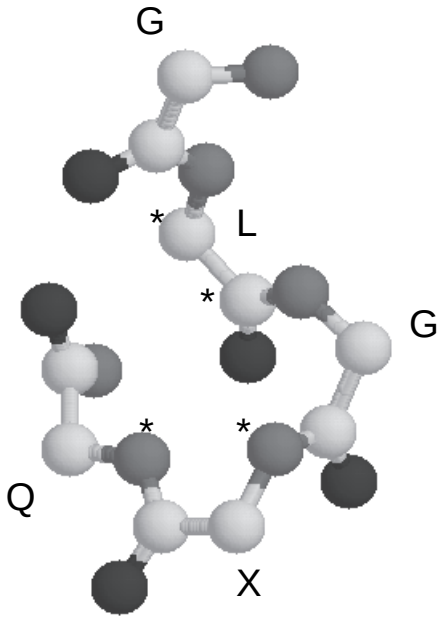


*Figure 12.* Best fit structure (side-chains suppressed for clarity). * indicates the target labeling sites.

*Table 5.* Comparison of REDOR results to standard structures

| Structure | rms deviation (Å) |
|---|---|
| anti-parallel β-sheet | 3.6 |
| parallel β-sheet | 3.1 |
| $3_1$ helix | 2.9 |
| Type II β-turn | 1.9 |
| π helix | 1.9 |
| $3_{10}$ helix | 0.63 |
| α-helix | 0.33 |
| Type I β-turn | 0.27 |
| distances from Table 3 | 0.41 |

the best-fit structure, calculated using the four target distances, as the fits are most sensitive to these parameters.

The extended sheet structures give a very poor fit to the data, while the more compact helices do a much better job. The best fit amongst the standard structures is that of a type I β-turn, which is often found connecting two strands of anti-parallel β-sheets (Creighton, 1993). The distinction between a segment of a helix and a turn is a fine one, as the $\phi$, $\psi$ torsion angles of the α-helix and those of the first residue of the type I turn are almost identical.

The final line of Table 5 compares the structure determined in the previous section with the distances reported in Table 3. The distances being compared here were derived from the same data sets with very different approaches to the analysis. The fact that the results agree as well as they do suggests that the data are sensitive to the desired distances, despite the isotopic scrambling.

The large portion of the signals contributing to the longest distances of Table 3 requires the participation of the LGGQ and LGNQ along with the LGSQ sequences. No separation between these various sequences can be made, and the data sets are weighted averages of the three. The conclusion that the LGSQ sequences form compact turn or helix-like structures thus applies equally to these other sequences as well, indicating that the suggested folding model (Simmons et al., 1996) is too simplistic. This conclusion directly refutes the suggestion that the GGX sequences form β-sheet crystallites (Thiel et al., 1994), for the LGGQ sequences at least. This is in fact consistent with the recent X-ray diffraction study (Grubb and Jelinski, 1997), where it was concluded that the alanine-rich portions of the sequence compose the majority of crystalline regions. The suggestion that the GGX sequences form $3_1$ helices (Kümmerlen et al., 1996) is also not consistent with these data, again for the LGGQ sequences at least.

By concluding that all of the LGXQ sequences form compact turn-like structures, we do not mean to imply that all of these sequences form identical structures. The three possible X residues (S, G, and N) are quite different from each other and it is likely that this identity plays a role. It is further possible that there is structural heterogeneity within each sequence. While this kind of structural heterogeneity is not usually considered in protein structural studies, it is important in the study of synthetic polymer fibers (Bovey, 1982). What the dephasing data from these samples show however, is that all of the LGXQ sequences are tightly folded, as the observed dephasings are not consistent with an appreciable fraction of these sequences in anything but a compact conformation. The question of structural heterogeneity may be qualitatively addressed by consideration of the solid-state NMR linewidths. The alanine $C_\alpha$ peak (Figure 3b) is the sharpest peak in the $^{13}C$ spectrum, consistent with the conclusion (Simmons et al., 1996, 1994) that all of the alanine residues reside in β sheets. After subtracting out the $^{13}C$ natural abundance spectrum, the leucine $C_\alpha$ peak remains almost 30% broader than that of the alanine, suggesting that the leucines occur in a greater variety of environments than do the alanines.

## Conclusions

The data show unambiguously that all of the LGXQ sequences form compact turn-like structures, but that the proposed folding model (Figure 1) is too simplistic, as this conclusion applies to the LGGQ as well as LGSQ sequences. The structures formed by these sequences are not necessarily all identical, and in fact some structural heterogeneity is suggested by the breadth of the leucine $C_\alpha$ peak. While the proposed model cannot be correct in detail, it is still a useful visual tool for describing what we believe are important features of the silk structure. The detailed variations in amino acid sequence affect local structure, which in turn influences the large scale packing within the silk fibers. The fact that all of these LGXQ sequences form compact structures emphasizes the heterogeneity of the amino-acid backbone in the silk structure. In contrast to the compact nature of the leucine containing region, the alanines are well de-

scribed by an extended β-sheet structure. The variation in the length of the sequence between alanine blocks must play a role in constraining the alanine sheets. It is possible that these irregularities prevent sharp phase-separation boundaries, as could be envisaged with a uniform repeating sequence. Thus we see these connecting sequences playing a crucial role in interfacing the alanine crystals, and preventing sharp, dramatic changes in structure. Rather, we believe the transition from crystalline to amorphous occurs on the scale of several amino-acid residues.

The experiments and data analysis reported here demonstrate that it is possible to obtain meaningful REDOR results in complex biopolymers, even in the event that isotopic scrambling takes place. Nevertheless, careful labeling strategies must be employed. For example, the one used here, in which the $^{13}C$ was placed in an essential amino acid, avoids possible migration of the $^{13}C$ label, while still being able to tolerate migration of the $^{15}N$ label.

This analysis also sets forth a framework for dealing with backgrounds contributed by the sea of $^{15}N$ neighbors in the vicinity of the target site.

## Acknowledgements

The authors acknowledge the support of NSF grants DMR-9708062 and MCB-9601018. Hoon Jung, Amy Blye, Neeral Shah, Zhitong Yang, and Sung Bae Lee assisted in collecting samples and caring for the spiders.

## References

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J. (1987) in *Crystallographic Databases–Information Content, Software Systems, Scientific Applications* (Eds. Allen, F.H., Bergerhoff, G. and Sievers, R.), Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.

Arnott, S., Dover, S.D. and Elliott, A. (1967) *J. Mol. Biol.*, **30**, 201–208.

Asakura, T., Demura, M., Date, T., Miyashita, N., Ogawa, K. and Williamson, M.P. (1997) *Biopolymers*, **41**, 193–203.

Bernstein, F.C., Koetzle, T.F., Williams, G. J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Bode, W., Papamokos, E. and Musil, D. (1987) *Eur. J. Biochem.*, **166**, 673–692.

Bovey, F.A. (1982) *Chain Structure and Conformation of Macromolecules*, Academic Press, New York, NY.

Bruns, C.M. and Karplus, P.A. (1995) *J. Mol. Biol.*, **247**, 125–145.

Creighton, T.E. (1993) *Proteins: Structures and Molecular Properties*, W. H. Freeman and Company, New York, NY.

Cunniff, P.M., Fossey, S.A., Auerbach, M.A., Song, J.W., Kaplan, D.L., Adams, W.W., Eby, R.K., Mahoney, D. and Vezie, D.L. (1994) *Polymers for Advanced Technologies*, **5**, 401–410.

Garbow, J.R., Breslav, M., Antohi, O. and Naider, F. (1994) *Biochemistry*, **33**, 10094–10099.

Gosline, J.M., DeMont, M.E. and Denny, M.W. (1986) *Endeavour*, **10**, 37–43.

Grubb, D.T. and Jelinski, L.W. (1997) *Macromol.*, **30**, 2860–2867.

Gullion, T. and Schaefer, J. (1989a) *Adv. Magn. Reson.*, **13**, 57–83.

Gullion, T. and Schaefer, J. (1989b) *J. Magn. Reson.*, **81**, 196–200.

Hing, A.W. and Schaefer, J. (1993) *Biochemistry*, **32**, 7593–7604.

Housset, D., Habersetzer-Rochat, C., Astier, J.-P. and Fontecilla-Camps, J.C. (1994) *J. Mol. Biol.*, **238**, 88–103.

Karplus, P.A. (1996) *Protein Sci.*, **5**, 1406–1420.

Kümmerlen, J., van Beek, J.D., Vollrath, F. and Meier, B.H. (1996) *Macromol.*, **29**, 2920–2928.

Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.

Lehninger, A.L., Nelson, D.L. and Cox, M.M. (1993) *Principles of Biochemistry*, Worth Publishers, New York, NY, second edition.

Lewis, R.V. (1992) *Acc. Chem. Res.*, **25**, 392–398.

Li, Y., Appleyard, R.J., Shuttleworth, W.A. and Evans, J.N.S. (1994) *J. Am. Chem. Soc.*, **116**, 10799–10800.

Michal, C.A. (1997) *Solid-State Deuterium and REDOR NMR Structural Studies of Spider Dragline Silk and A New Experiment for Measuring Multiple Dipolar Couplings*, Ph.D. thesis, Cornell University.

Oda, Y., Matsuura, S., Norioka, S. and Sakiyama, F. (1996) *Acta Cryst. D*, **52**, 1027–1029.

Simmons, A., Michal, C.A. and Jelinski, L.W. (1996) *Science*, **271**, 84–87.

Simmons, A., Ray, E. and Jelinski, L.W. (1994) *Macromol.*, **27**, 5235–5237.

Stauffer, S.L., Coguill, S.L. and Lewis, R.V. (1994) *J. Arachnol.*, **22**, 5–11.

Svensson, L.A., Sjolin, L., Gilliland, G.L., Finzel, B.C. and Wlodawer, A. (1986) *Proteins Struct. Funct. Genet.*, **1**, 370–375.

Thiel, B.L., Kunkel, D.D. and Viney, C. (1994) *Biopolymers*, **34**, 1089–1097.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.

Work, R.W. and Emerson, P.D. (1982) *J. Arachnol.*, **10**, 1–10.

Work, R.W. and Young, C.T. (1987) *J. Arachnol.*, **15**, 65–80.

Xu, M. and Lewis, R.V. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 7120–7124.

Zemlin, J.C. (1968) A study of the mechanical behavior of spider silks, Technical Report TR69-29-CM (AD 684333), US Army Natick Laboratories, Natick, MA.

Zhao, B., Carson, M., Ealick, S.E. and Bugg, C.E. (1992) *J. Mol. Biol.*, **227**, 239–252.