Supplementary Information

Accompanying: "PR-CALC: A Program for the Reconstruction of NMR Spectra from Projections," by Brian E. Coggins and Pei Zhou

Department of Biochemistry, Duke University Medical Center, Durham, NC 27710

Appendix A: Control Files and Scaling of Input Data

Before the spectrum can be reconstructed, a text "control file" must first be created, describing the experiment and the projection data. The control file has two sections, the first defining the experiment and its parameters, and the second listing the available projection data. An example control file, for a (4,2)-D methyl/amide NOESY experiment, is shown in Figure S1.

At this point, it is important to ensure that the signal levels on the different projections are consistent. The **scale** option in the control file can be used to rescale individual projections to a common level if needed. The proper scaling factors can be calculated *a priori* based upon the pulse sequence, the number of transients used to collect the projections, and knowledge of the processing pipeline.

As an example of how scaling corrections can be determined, consider a typical (4,2)-D HNCACB experiment. In order to achieve equal sensitivity on the different projections, we collect different numbers of transients for the HN-N orthogonal projection, the HN-CA and HN-CB orthogonals, and the tilted projections, at a ratio of 1:2:4 transients for the three categories, respectively. For simplicity in the discussion below, we will assume that the numbers of transients collected are 1, 2 and 4,

respectively, although naturally in practice these numbers must be multiplied by the minimum phase cycle size for the pulse sequence. If we define the signal recorded per transient for the HN-N projection as *s*, and the corresponding noise level as *n*, the signal and noise levels for the orthogonal projections are calculated as follows, for the HN-N:

HN-N Orthogonal	Signal	Noise	SNR
Initial value per transient	s	n	s/n
Sensitivity-enhancement processing	$\times 2$	$\times\sqrt{2}$	
Final level	$2s$	$\sqrt{2}n$	$\sqrt{2}s/n$

and the others:

HN-CA and HN-CB Orthogonals	Signal	Noise	SNR
Initial value per transient	s	n	s/n
2 transients collected per FID	$\times 2$	$\times\sqrt{2}$	
Final level	$2s$	$\sqrt{2}n$	$\sqrt{2}s/n$

Thus the orthogonal projections have matching signal and noise levels after processing the HN-N projection for gradient sensitivity-enhancement. The tilted projections are different in that they begin with four-fold less signal per transient than the orthogonals. This is due to the signal modulation in the additional two dimensions, which splits the peak intensity into four signals at different projected positions. The subsequent separation of intermodulated projections isolates each of the multiplet components on an independent plane, at the same time that it restores the sensitivity through the additional and subtraction of FIDs. The full analysis is given below:

Tilted	Signal	Noise	SNR
Initial value per transient	$s/4$	n	$s/4n$
4 transients collected per FID	$\times 4$	$\times 2$	
Sensitivity-enhancement processing	$\times 2$	$\times\sqrt{2}$	
Separation of intermodulated projections	$\times 4$	$\times 2$	
Final level	$8s$	$4\sqrt{2}n$	$\sqrt{2}s/n$

The final absolute levels for the tilted projections are four-fold higher than for the orthogonal projections, although they share the same SNR. Thus a **scale = 4.0** value is needed in the control file for each of the three orthogonal projections, to bring the absolute levels up to match those on the tilted projections. Although the specific purpose of this correction is to make signal levels consistent, note that as a consequence of the identical SNR, the noise level also becomes consistent.

Appendix B: Calculation Time and Large Datasets

Strategy for Large Datasets. Although Equations 10 and 11 may suggest significant limitations on the kinds of data that could be evaluated routinely by PR-NMR, we have found that with an appropriate choice of strategy, the calculation burdens become quite manageable. First, as a general rule with large datasets, and certainly before starting any large calculation, we highly recommend choosing a handful of expected crosspeaks as test cases for optimizing the reconstruction parameters. For sequential assignment data collected according to the methodology in Venters et al. (2005), one should experiment with different values of the HBLV parameter *k*, to determine the optimum reconstruction settings.

For large (4,2)-D or (5,2)-D datasets, once the reconstruction parameters have been determined we recommend reconstructing individual slices or regions for each residue. The batch reconstruction feature allows this to be carried out quickly. The result is a set of small spectra for individual residues, each produced at high resolution. The reconstruction parameters may be optimized for each individual slice or region, as needed. An alternative is to reconstruct the full spectrum, which is often feasible, although several days of calculation time may be needed. This option may also require reducing the resolution significantly below what the data would support, because of either the calculation time or the file size (note that most current computing hardware restricts file sizes to less than 2 or 4 GB, directly limiting the reconstruction resolution). Note also that it is not possible, when calculating a single large spectrum, to optimize the reconstruction parameters for different regions of the spectrum.

Calculation Time Examples and Practical Advice. The practical consequences of computational complexity can best be appreciated by considering some commonplace examples. We shall discuss potential approaches and results for three typical PR-NMR experiments: a (3,2)-D HNCO, a (4,2)-D HNCACB and a (4,2)-D methyl/amide-NOESY. In these discussions, we shall concern ourselves primarily with the practical reconstruction calculation issues for each dataset, rather than the relative merits of different data collection approaches or reconstruction algorithms. The scenarios that we have considered, with respect to the algorithms to use for different numbers of projections, are based on the discussions of the algorithms in the literature. All computation times are for a single Intel Xeon 2.5 GHz processor with 1 GB of memory. Timings were measured to one second precision; for the longer calculations marked with

a footnote in Table S1, the full calculation was not timed, but rather the calculation time was extrapolated based on the progress of the calculation over 10 minutes. A "startup time," consisting of the overhead time needed for loading data, creating output files, filtering projections and initializing HBLV reconstruction parameters, was measured separately from the "calculation time," the time needed to evaluate Equation 1 for each reconstruction point. For HBLV reconstructions, the default noise floor option—with the noise floor enabled—was used. The measured timings are given in Table S1.

The results show that the 3-D HNCO dataset can be reconstructed quickly by almost any of the methods described in the literature, at any desired resolution, in as little as one or as many as seven minutes. However, it should be noted that HBLV reconstruction for 36 projections was not possible due to insufficient memory to track the reconstruction bins, but could be completed in 10 minutes with 18 projections, which highlights the combinatorial difficulties that can arise with HBLV, in this case because of memory requirements rather than time requirements. Nevertheless, these results confirm that for a typical 3-D spectrum such as HNCO, reconstruction of the full spectrum is straightforward, and requires minimal time and storage space.

As a second example, we consider the (4,2)-D PR-NMR reconstruction of the TROSY-HNCACB spectrum of human carbonic anhydrase (HCA) II. The collection of the projection data for this experiment was described previously (Venters et al., 2005). Because reconstruction time and disk storage requirements can be a serious issue for high resolution 4-D datasets, we have considered both the strategy of calculating selected planes and regions as well as that of calculating the full spectrum. The former is quite rapid, requiring only seconds to calculate a plane. In fact, if the batch mode is used,

individual planes can be calculated for all 265 residues in the protein, by LV or BP, in less than 5 minutes, and the total disk space needed to store the planes of interest would be less than 5 MB (with the batch or pipe modes, the "startup time" is only needed at the beginning of the first calculation; subsequent planes require only the "calculation time"). Naturally, the time requirements for 4-D regions scale approximately linearly with the increased number of planes. In all of these cases of slices and regions, it would be quite reasonable to increase the reconstruction resolution beyond the example given here. Reconstructing the full spectrum by LV or BP is feasible, although at 17 hours, it certainly is not instantaneous. Note that more than 1.5 GB of disk space is needed to store the full reconstruction.

As one would expect, HBLV requires considerably more calculation time. For this example, there are potentially 490,314 bins to be evaluated for each data point, although the noise floor cutoff significantly reduces the actual number examined for many points. Slices and regions can be determined with HBLV in a reasonable amount of time. We did not attempt to measure the time of a full HBLV reconstruction under controlled conditions, although such calculations were completed previously and required 4-7 days of computer time, depending upon the resolution in the direct dimension.

Finally, we present results for the (4,2)-D methyl/amide NOESY spectrum of HCA II, which we have published previously (Coggins et al., 2005). For the quantitative reconstruction of NOESY spectra, we measure a large number of projections (100 experiments, which yields 391 projections after application of PRSP), and reconstruct using FBP. As with the HNCACB experiment, it is straightforward to reconstruct slices and regions at medium or high resolution, but calculating the entire spectrum requires

several days at medium resolution, and would be unfeasible at high resolution without a multiprocessing or cluster system. With an eight node cluster, the calculation would require two days. On more modest computers, the best option for this experiment is to reconstruct a plane for each of the 265 residues, which could be completed, even at high resolution, in just two hours.

One can see from these examples that 3-D reconstructions are quick and straightforward, and that (4,2)-D spectra can be approached easily through the strategy of computing selected planes and regions, although reconstructing complete 4-D matrices places some demands on computer hardware. It is interesting to note that the cost per reconstruction point per projection is in fact constant, as suggested by Equation 10, and that the cost on this particular CPU comes to approximately 10 μs. The HBLV results for the TROSY-HNCACB spectrum are naturally the exception. The continued advancement of computer technology, including the increasingly widespread availability of high performance cluster systems, should progressively reduce the computational limitations of this method.

References

Coggins, B. E., Venters, R. A. and Zhou, P. (2005) *J. Am. Chem. Soc.,* **127,** 11562-11563.

Venters, R. A., Coggins, B. E., Kojetin, D., Cavanagh, J. and Zhou, P. (2005) *J. Am. Chem. Soc.,* **127,** 8785-8795.

Table S1 – Examples of Calculation Times

Type	Total Proj.	Digital Resolution (data points/axis)	Algorithm	*Nn* (note a)	File Size (MB)	Startup Time[b] (h:min:s)	Calc. Time[c] (h:min:s)	Total Time (h:min:s)	Calc. Time/*Nn* (μs)
(3,2)-D HNCO of Protein G, B1 Domain									
Full	6	32 x 32 x 420[d]	LV	2.6×10^6	1.77	0:00:10	0:00:51	0:01:01	20
Full	6	32 x 32 x 420	HBLV *k*=2	2.6×10^6	1.77	0:00:25	0:01:00	0:01:25	23
Full	6	64 x 64 x 420	LV	1.0×10^7	7.08	0:00:20	0:02:15	0:02:35	13
Full	6	64 x 64 x 420	HBLV *k*=2	1.0×10^7	7.08	0:00:34	0:02:28	0:03:02	14
Full	6	128 x 128 x 420	LV	4.1×10^7	28.31	0:00:56	0:05:41	0:06:37	8
Full	6	128 x 128 x 420	HBLV *k*=2	4.1×10^7	28.31	0:01:14	0:06:47	0:08:01	10
Full	18	64 x 64 x 420	BP	3.1×10^7	7.08	0:00:36	0:05:45	0:06:21	11
Full	18	64 x 64 x 420	HBLV *k*=6	3.1×10^7	7.08	0:01:22	0:10:20	0:11:42	20
Full	36	64 x 64 x 420	BP	6.2×10^7	7.08	0:00:57	0:10:19	0:11:16	10
Full	36	64 x 64 x 420	FBP	6.2×10^7	7.08	0:00:55	0:11:21	0:12:16	11
(4,2)-D TROSY-HNCACB of Human Carbonic Anhydrase II									
2-D Slice[e]	23	64 x 64[f]	LV	9.4×10^4	0.02	0:00:24	0:00:01	0:00:25	11
2-D Slice	23	64 x 64	BP	9.4×10^4	0.02	0:00:24	0:00:01	0:00:25	11
2-D Slice	23	64 x 64	HBLV k=8	9.4×10^4	0.02	0:01:22	0:00:22	0:01:44	234
4-D Region[g]	23	64 x 64 x 5 x 5[h]	LV	2.4×10^6	1.05	0:00:24	0:00:21	0:00:45	9
4-D Region	23	64 x 64 x 5 x 5	BP	2.4×10^6	1.05	0:00:21	0:00:19	0:00:40	8
4-D Region	23	64 x 64 x 5 x 5	HBLV k=8	2.4×10^6	1.05	0:01:19	0:09:27	0:10:46	241
Full	23	64 x 64 x 128 x 738	LV	8.9×10^9	1560.28	0:02:43	17:10:00[i]	17:12:43	7
Full	23	64 x 64 x 128 x 738	BP	8.9×10^9	1560.28	0:02:41	17:10:00[i]	17:12:41	7
(4,2)-D Methyl/Amide NOESY of Human Carbonic Anhydrase II									
2-D Slice[j]	391	64 x 64[k]	FBP	1.6×10^6	0.02	0:01:50	0:00:09	0:01:59	6
2-D Slice	391	128 x 128	FBP	6.4×10^6	0.07	0:01:50	0:00:29	0:02:19	5
4-D Region[l]	391	64 x 64 x 5 x 5[m]	FBP	4.0×10^7	1.05	0:01:53	0:03:43	0:05:36	6
4-D Region	391	128 x 128 x 5 x 5	FBP	1.6×10^8	4.20	0:01:53	0:13:52	0:15:45	5
Full	391	64 x 64 x 64 x 403	FBP	4.1×10^{10}	427.82	0:03:08	64:06:00[i]	64:09:08	6
Full	391	128 x 128 x 128 x 403	FBP	3.3×10^{11}	3422.55	0:06:31	416:40:00[i]	416:46:31	5

[a]The number of reconstruction data points (*N*) times the number of projections (*n*), as in Equations 10 and 11.

[b]The time from the program's start until it begins evaluation of Equation 1; includes creation of the output file, loading input data, filtering projections (when applicable) and initializing reconstruction algorithm parameters. All times were measured on a single 2.4 GHz Intel Xeon processor with 1 GB memory, running Microsoft Windows XP. PR-CALC was compiled with the Microsoft Visual Studio 8.0 Optimizing C++ Compiler with full optimization, and the Microsoft Visual Studio 8.0 C++ Standard Library was used.

[c]The time needed for the evaluation of Equation 1 for all reconstruction points, also including overhead tasks that occur during this evaluation, such as the paging of data to and from disk.

[d]For all HNCO reconstructions, corresponds to the N, CO and HN dimensions, respectively.

[e]For all 2-D slice reconstructions of the HNCACB, computed for residue K126, at HN=7.965 ppm and N=124.569 ppm.

[f]For all 2-D slice reconstructions of the HNCACB, corresponds to the CA and CB dimensions.

[g]For all 4-D region reconstructions of the HNCACB, computed for residue K126, centered at HN=7.965 ppm and N=124.569 ppm.

[h]For all 4-D reconstructions of the HNCACB, corresponds to the CA, CB, N and HN dimensions, respectively.

[i]This approximate value was extrapolated based on the progress of the calculation after 10 minutes of calculation time.

[j]For all 2-D slice reconstructions of the methyl/amide NOESY, computed for residue S50, at HN=8.376 ppm and N=122.553 ppm.

[k]For all 2-D slice reconstructions of the methyl/amide NOESY, corresponds to the HM and CM dimensions.

[l]For all 4-D region reconstructions of the methyl/amide NOESY, computed for residue S50, at HN=8.376 ppm and N=122.553 ppm.

[m]For all 4-D reconstructions of the methyl/amide NOESY, corresponds to the HM, CM, N and HN dimensions, respectively.

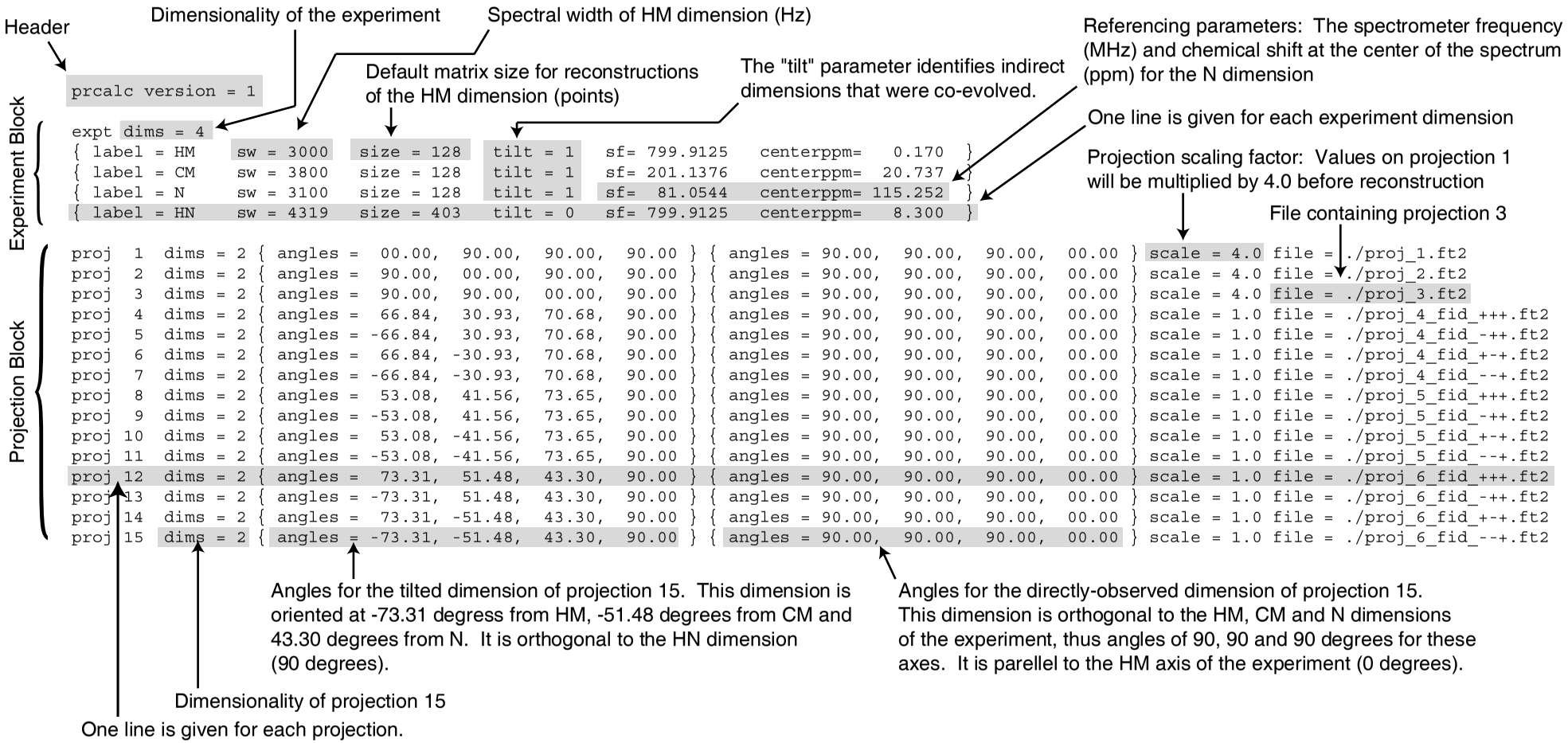

Figure S1 – PR-CALC Control Files. A portion is shown, with annotations, from a control file defining a (4,2)-D methyl/amide NOESY experiment. The full control file lists 391 projections.