Assessment of the Higher Order Structure of Formulated Monoclonal Antibody Therapeutics by 2D Methyl Correlated NMR and Principal Component Analysis.

Luke W. Arbogast,[1,2] Frank Delaglio,[1] Robert G. Brinson[1] and John P. Marino[1]

1. National Institute of Standards and Technology, Institute for Bioscience and Biotechnology Research, Rockville, Maryland, United States.
2. Corresponding author: luke.arbogast@nist.gov

ABSTRACT:

Characterization of the higher order structure (HOS) of protein therapeutics and in particular, monoclonal antibodies, by 2D ^{1}H-^{13}C methyl correlated NMR has been demonstrated as precise and robust. Such characterization can be greatly enhanced when collections of spectra are analyzed using multivariate approaches such as principal component analysis (PCA), allowing for the detection and identification of small structural differences in drug substance that may otherwise fall below the limit of detection of conventional spectral analysis. A major limitation to this approach is the presence of aliphatic signals from formulation or excipient components which result in spectral interference with the protein signal of interest; however, the recently described *S*elect*i*ve *E*xcipient *R*eduction and *R*emov*a*l (SIERRA) filter greatly reduces this issue. Here we will outline how basic 2D ^{1}H-^{13}C methyl correlated NMR may be combined with the SIERRA approach to collect 'clean' NMR spectra of formulated monoclonal antibody therapeutics (i.e., drug substance spectra free of interfering component signals) and how series of such spectra may be used for HOS characterization by direct PCA of the series spectral matrix.

Basic Protocol 1: NMR Data Acquisition
Basic Protocol 2: Full Spectral Matrix Data Processing and Analysis
Support Protocol 1: Data Visualization and Cluster Analysis

KEYWORDS:

Monoclonal antibodies, biotherapeutics, nuclear magnetic resonance, methyl correlated, SIERRA, principal component analysis

INTRODUCTION:

Protein therapeutics are complex drug substances whose safety and efficacy are dependent on the critical quality attribute of higher order structure (HOS). In order to maintain the correct HOS over the practical shelf-life for pharmaceutical deployment, often in highly concentrated dosages (~ 100 mg/ml), such therapeutics must be formulated in stabilizing and preserving cosolute excipient molecules. It is therefore necessary to have methods to precisely characterize the HOS of protein therapeutics in the presence of these excipients. Unfortunately, currently employed methods for HOS characterization, namely circular dichroism (CD) and Fourier transform infrared spectroscopy (FT-IR), have limitations that have raised concern with regard to their performance when applied to therapeutics under formulation conditions and their sensitivity to relevant changes in HOS (Lin, Glover, & Sreedhara, 2015; Wen, Batabyal, Knutson, Lord, & Wikström, 2020).

Two dimensional (2D) ^{1}H-^{13}C methyl correlated NMR, on the other hand, has been demonstrated to be a robust and precise means to acquire high-resolution spectral fingerprints of protein therapeutics, including monoclonal antibodies (mAbs) at natural isotopic abundance. Such spectra can be used as reporters of the HOS and can be collected on mid- to high-field spectrometers (≥ 600 MHz) in a matter of hours (Arbogast, Brinson, & Marino, 2015; Brinson et al., 2019). When combined with multivariate chemometric methods, such as principal component analysis, 2D methyl NMR spectra have been demonstrated to be able to distinguish small structural differences with low levels of detection at residue level resolution (Arbogast, Delaglio, Schiel, & Marino, 2017). While the presence of aliphatic excipient molecules has previously been a concern for such methods, the recently introduced *S*elect*i*ve *E*xcipient *R*eduction and *R*emov*a*l (SIERRA) filter (Arbogast, Delaglio, Tolman, & Marino, 2018), a selective pulsed-double resonance element that can be appended to standard 2D ^{1}H-^{13}C methyl NMR experiments, provides a method to selectively reduce the signals for excipient components with minimal losses to the protein therapeutic signal. Further, post-acquisition data processing can remove any residual excipient signal by difference with a synthetically modeled signal using the SMILE spectral reconstruction algorithm (Ying, Delaglio, Torchia, & Bax, 2017). Together, this combined attenuation using pulse-based signal suppression followed by numerical subtraction via spectral modeling allows for the removal of excipient signals to the level of the baseline, without decreases in signal-to-noise of the protein therapeutic signal. When combined with principal component analysis (PCA), high-resolution HOS characterization of formulated protein therapeutics can be achieved to provide

both a test of the structural similarity of analyte samples as well as insight into mechanisms of excipient-protein interactions and stabilization.

In this document, we will detail strategic considerations for parameterization of therapeutic protein spectral space to allow for facile structural interpretation of the data as well as protocols for acquisition and processing of spectral data. These include considerations for sample preparation, optimization of spectral parameters for the SIERRA-filtered 2D ^{1}H-^{13}C methyl heteronuclear single-quantum coherence (HSQC) experiment, methods for data processing and finally multivariate, principal component analysis of the resultant spectral data.

STRATEGIC PLANNING

An important consideration before acquisition of data is the desired product parameter space to be explored. The basis for structural characterization by multivariate analysis of product NMR spectral libraries is the creation of a well-defined spectral space that covers a given product parameter space, ideally covering the breadth of relevant parameter variability. Supervised and unsupervised approaches can then be employed to identify and classify a test spectrum in relation to the well-defined spectral library. Therefore, it is worth considering what reference or challenge materials may be used to adequately define the parameter space of interest and what orthogonal methods can be employed to create a well-characterized parameter space. In this paper, we will examine the effects of concentration of a small molecule excipient, L-alanine, on formulation of NISTmAb as a demonstration of the method.

Sample Considerations:

Intrinsic sample properties will largely be dictated by the considerations outlined above; however, several sample and experimental conditions are important to consider, including sample concentration, solution properties and temperature. Ideally, data will be acquired at formulated concentration, but if concentrations are too low, time requirements to achieve adequate signal to noise and acquire replicate measurements may become prohibitive. Likewise, while, higher concentrations are generally favorable, if increases in solution viscosity result, spectral quality may be degraded. The range of favorable concentrations will be sample dependent, but for NISTmAb, we find that concentrations from 20 mg/mL to 60 mg/mL provide reasonable quality data, though spectra have been successfully acquired as low as 10 mg/mL and as high as 100 mg/mL. Temperature likewise has a large impact on spectral quality. In general, it is best to collect data at as high a temperature as the sample will tolerate to increase molecular rotation and achieve better sensitivity. For NISTmAb, data can be collected at 50 °C without compromising sample integrity, however for many samples this will be too high. Because of the rapid rotation of methyl groups about their local C3 symmetry axis, the effects of global rotation are muted compared to other reporters such as amide groups and adequate data can be obtained at lower temperature, with 37 °C being preferred for many mAb samples. Regardless of the employed experimental temperature, proper calibration and precise control are critical as even small deviations in temperature may significantly influence chemical shifts and line-widths and thus compromise statistical analysis of spectra. Because non-labile methyl groups are used as reporters, experiments tend to be quite tolerant to ionic and pH effects and spectra can readily be acquired under a variety of solution conditions. However, given instrument limitations, high salt concentrations (above ca. 300 mM) will typically degrade the performance of NMR probe hardware and reduce sensitivity. Additionally, although spectra may be successfully acquired, when performing multivariate analysis on spectral data, effects from pH, salt or temperature must be well controlled and understood to avoid assigning structural meaning to trivial spectral differences arising from solution conditions.

Extrinsic sample factors may also impact results, including chemical shift referencing and sample tube architecture. The choice of the appropriate chemical shift referencing scheme will be dictated by sample fidelity. In general, internal referencing to sodium 3-(trimethylsilyl)propane-1-sulfonate (DSS, ~100 µM) is the easiest and most practical solution. However, if concerns over adulteration of the analyte sample by DSS are present, alternatives include external referencing to DSS or to an excipient signal that is known to be invariant to sample conditions. Likewise, the choice of NMR sample tube will be dictated by sample limitations. If sample quantities are not limited, it is recommended to use standard 5mm NMR tubes, slightly overfilled (ca. 600 µL total sample volume), to reduce chances for changes in field homogeneity over the course of sample replicate measurement, to which chemometric methods are sensitive. This is especially true for high temperature data acquisition (37 ^{o}C and higher), where degassing and sample evaporation may lead to changes in optimal shim conditions. When sample conditions are limited and a smaller volume tube is required, Bruker shaped microtubes filled with 300 µL of sample are the most optimal solution, if spectrometer hardware configuration allows. However, if a susceptibility matched tube with plunger is necessary, care must be taken to adequately degas the tube before insertion in the magnet, by placing the tube without a plunger at a temperature 5-10 ^{o}C higher than acquisition temperature for 15-20 minutes. Afterwards, gentle agitation of the tube to remove resultant air bubbles should be accomplished before setting of the plunger. The plunger should likewise be set above the receiver coil window and secured in place by parafilm or similar to avoid changes in sample volume during acquisition.

SIERRA Considerations

The SIERRA filter is designed to act on any given target signal in a selective double resonance manner so long as it is resolved by ± $^{1}J_{HC}$ in both dimensions. It has been previously demonstrated on various excipient targets such as acetate, polysorbate-80 and methionine. The target can be generally be of any molecular weight and any spin multiplicity, although ideal outcomes are only possible for AX spin-systems (Tolman & Arbogast, 2019). When considering methyl spectra, AX3 and AX2 type signals will be the most commonly encountered, and for such type of signals, parameter optimization of the ^{13}C contact pulse power level can still yield suppression in excess of 90% the initial intensity. While there is no upper limit of intensity on which the SIERRA filter can act, given that a 10% residual may remain, this suggest a practical upper limit of 100:1 excipient signal to protein signal, as the residual should be on the same order of magnitude as the protein signal. The SIERRA filter can be applied to more than one signal by applying it consecutively at different target frequencies. However, losses in protein signal during the filter are additive and has previously been described, after four applications, spectral quality degrades appreciably (Arbogast et al., 2018).

Instrumentation

The characterization of therapeutic proteins by NMR at ^{13}C natural abundance (1.1%), in general requires moderate to high magnetic field strengths. The recent multi laboratory 2D NMR study on the Fab fragment from the NISTmAb found that precision of measurement, as measured by combined chemical shift deviation, was less than 4.5 ppb for fields as low as 500 MHz. On the other hand, even though the greatest precision was observed at the highest field strengths (*e.g.*, 800 MHz - 900 MHz), gains were modest and comparable results were obtained at the lower fields (Brinson et al., 2019). For intact mAbs, from which there are substantially more methyl signals as well as broader lines, it is recommended to use the highest field available, and fields below 600 MHz are strongly discouraged. In addition to field strength considerations, owing to the low natural abundance of ^{13}C as well as the requisite depth of data necessary for successful chemometric applications, cryogenically cooled probes, which can deliver improvements in signal-to-noise of up to 3-4 times that of conventional probes for low-salt samples (Kovacs, Moskau, & Spraul, 2005), are strongly encouraged. However, natural abundance ^{1}H-^{13}C gsHSQC spectra collected on room temperature probes have been reported, but require 12x – 16x more experimental time than a cryogenic probe (Arbogast, Brinson, & Marino, 2016; Brinson et al., 2017). Before data acquisition, spectrometer and probe performance on all channels should be validated for drift and stability, benchmarked for line-width and signal-to-noise and calibrated for pulse-width and temperature. Temperature calibration is of particular importance as even small offsets in the desired temperature can have significant impacts on spectra to which chemometric methods are quite sensitive (Ghasriani et al., 2016).

BASIC PROTOCOL 1

NMR Data Acquisition

In general, the preferred experiment to record formulated monoclonal antibody methyl spectra at natural isotopic abundance is the gradient-selected, sensitivity-enhanced HSQC (gsHSQC) (Schleucher et al., 1994), given its superior suppression of artifacts resulting from the large abundance of $^{1}H(\ ^{12}C)$ spins (98.9%). The gsHSQC can be tailored for methyl groups by optimizing the forward and reverse INEPT transfer delays for the methyl $^{1}J_{HC}$ (~125 Hz). However, the presence of aliphatic excipient components in the vicinity of the protein methyl spectral window, regardless of spin-multiplicity, can have profound negative impacts on spectral quality through the introduction of baseline distortions and T_1 noise. When such excipients are present, the SIERRA filter can be used to remove the undesired and detrimental signals from a spectrum of interest with high specificity, by appending the SIERRA selective double resonance pulse element to the beginning of a standard pulse sequence such as the gsHSQC. The system-specific parameters for a given standard sequence are unchanged by the SIERRA filter and can be determined beforehand as normal. The SIERRA filter however, requires additional optimization to achieve efficient suppression of targeted signals. In general, it is necessary to identify the 2D frequency position of the signal(s) of interest to be suppressed, to determine the $^{1}J_{HC}$ of the target signal(s) and additionally to optimize the power level of the ^{13}C selective contact pulse in the SIERRA element, which is dependent of the spin multiplicity of the target signal and possible third spin effects (Tolman & Arbogast, 2019). The protocol below is described for the acquisition of a SIERRA-filtered ^{1}H-^{13}C HSQC monoclonal antibody methyl spectrum. It is assumed the user is familiar with the standard gsHSQC experiment and has established appropriate pulse-sequence parameters for the protein system of interest. Example data are demonstrated on a 40 mg/mL sample of NISTmAb (PS-8670) in 25 mM L-histidine, with 20 mM L-alanine, 4 mM sodium acetate, 100 µM DSS and 3% (v/v) D_2O at pH 6.0.

Materials:

High field (≥ 600 MHz) Bruker NMR Spectrometer running TopSpin 3.0 or higher*
ghsqc_SIERRA parameter and pulse sequence files

**It is possible to run SIERRA experiments on other vendor consoles (e.g., JEOL or Agilent/Varian) or with older TopSpin software. However, the described protocol is specific to the listed spectrometer setup.*

Protocol steps

1. Prepare sample in magnet using standard procedures.

   Equilibrate to desired temperature, lock, match/tune, shim, optimize ^{1}H 90 pulse. Data presented here were collected at 50 °C. Values for the ^{1}H 90° pulse ranged from 9.825 µs to 10.65 µs depending on alanine concentration

2. Create a new experiment using the '*edc*' command and load the gsHSQC_SIERRA experiment.

   The SIERRA experiment can be loaded either by setting the pulse program 'pulprog' line under the 'acqupars' tab or alternatively, by loading an appropriate parameter file using the 'rpar' command.

3. Set relevant gsHSQC parameters to optimum values for sample/system of interest (*e.g.*, *d1, O2, SW*, *cnst4*/*d4*).

   Data presented here were collected with a recycle delay, d1 of 1.5 s, a ^{13}C center frequency, O2P of 16 ppm, a sweep width, SW of 14 ppm in ^{1}H and 30 ppm in ^{13}C, acquisition times, AQ of 100 ms and 10 ms in ^{1}H and ^{13}C, and an INEPT transfer constant, cnst4 of 145 Hz corresponding to an INEPT transfer delay, d4 of 1.725 ms.

3a. If the target system contains strong signals downfield of the protein spectrum, such as from excipients like sucrose or glutamate, set the ZGOPTNS flag –DSELECTIVE, to switch to a methyl-selective gsHSQC. Determine the pulsewidths and power levels of the ^{13}C selective 90^{o} and 180^{o} pulses using the '*stdisp*' interface.

   For mAb methyl spectra we recommend Q5.1000 and Q3.1000 selective ^{13}C pulses applied over a bandwidth of ca. 20 ppm centered at approximately 18 ppm.

   Acquire a test 1D spectrum (Figure 1A).

4. Set parameters for a test scan to identify excipient signal and acquire a 2D spectrum; remove the *ZGOPTNS* flag – *DSIERRA* to turn off the SIERRA filter. Set number of scans to 4 *'ns 4'*, steady-state scans to 4 *'ds 4'* and the ^{13}C acquisition time to 5 ms '*1aq 5m'*. If the $^{1}J_{HC}$ of the target signal is not known, set '*pldb12 1000*' to turn off ^{13}C decoupling during acquisition and determined the coupling from the frequency positions of the doublet in the ^{1}H dimension.
5. Process 2D spectrum using the *'xfb'* command and determine frequency position(s) and if necessary, $^{1}J_{HC}$ of target signal(s).

   Peaks can be identified manually from the cursor position or by using the TopSpin peak picking program to create a peak table for later use.

6. Create a new experiment and set the *ZGOPTNS* flag *–DSIERRA* to turn on the SIERRA filter. Set '*cnst20 100*' to set ^{13}C CP power to the theoretical optimum value for an AX system. Edit the FQ1LIST and FQ2LIST entries to the ^{1}H and ^{13}C ppm frequencies respectively of the first excipient signal to be suppressed (make sure the first line reads '*bf ppm,*'). Set '*cnst5*' to the $^{1}J_{HC}$ value. If decoupling was turned off in step 5, reset *pldb12* to appropriate value. In the '*acqupar'* tab, switch to 1D mode and acquire a test spectrum (Figure 1B).

   A significant reduction in the intensity of the target peak should be observed.

7. Process 1D spectrum. If excipient suppression is adequate, skip to step 9. Otherwise, zoom in on excipient signal; then right click, select '*save display region to,*' and click '*ok*'.

   With this few scans. the protein signal should only be at or below the limit of detection of 3:1 signal to noise. If the excipient signal is not observed above this level, suppression is adequate. If the excipient signal exceeds the limit of quantification of 10:1 signal to noise, further optimization is required.

8. Optimize the ^{13}C contact pulse power level used by arraying the '*cnst20*' parameter with the '*popt*' interface. For AX_3 type signals, *cnst20* should be approximately 200. Perform an array around this value to find the value that is closest to achieving the null crossing. Then set *cnst2* to it (Figure 1C).

   Typically, a course array is performed from -/+ 50 of theoretical value in increments of 10, followed by a fine array from -/+ 10 of the nearest value to the null crossing in increments of 2.

9. Create a new 2D experiment and set all relevant gsHSQC parameters for your system (ns, ds, d1, O2, SW, AQ, cnst4/d4). Make sure the ^{13}C contact pulse power level '*cnst20*' is set to the optimum value. If more than one signal is being targeted, edit the FQ1LIST and FQ2LIST lists to include the frequencies of all target signals.
10. Acquire a 2D spectrum and make sure all targeted signals are adequately suppressed (Figure 1D).

    Target signals do not need to be perfectly suppressed. A small residual positive negative mode or mixed-phase signal can be removed in post-processing.

11. Create additional replicate experiments as needed for statistical assessment and queue them up for acquisition. Between each acquisition, queue a '*topshim tune tunea'* command to touch up the shims between each experiment.

    It is important to optimize shims between each duplicate measurement to improve the precision during PCA, which is sensitive to small changes in signal lineshape.

BASIC PROTOCOL 2

Full Spectral Matrix Data Processing and Analysis

Data to be analyzed by multivariate methods such as PCA should be batch processed with common processing parameters so as not to encompass variation from processing artifacts. After processing of raw spectral data, spectra need to be aligned using a chemical shift referencing scheme, and all data need to be of a uniform size; therefore, interpolation may be necessary. In order to accomplish this data processing on a large dataset, a number of C-shell scripts have been developed for batch processing of data using NMRPipe functions. Scripts are given below or provided as tables and additionally available online (see Online Resources). Example data are demonstrated on a series of SIERRA-filtered ^{1}H-^{13}C gsHSQC methyl spectra collected on a series of NISTmAb samples with varying concentrations of L-alanine. Example data are processed with a cosine squared apodization function, zero-filling in both dimensions and 0^{th} order polynomial baseline correction in the detected dimension over an extracted region from 2.5 ppm to -0.75 ppm in ^{1}H using NMRPipe processes (Delaglio et al., 1995) Additional resources concerning data processing presented here can be found on the NMRPipe webpage provided below in *Internet Resources.*

Materials:

Linux workstation with C or T-shell running NMRPipe v9.8 or higher

NMRPipe C-Shell batch processing scripts

Protocol steps

1. Import all raw spectrometer data and parameter files into a directory tree on a data processing workstation.
2. Create a folder called '*com'* in the parent directory of the imported data and copy batch processing scripts into it. *all.com, clean.com, title.com, fid.com, proc.com, ref.com, interp.com, show.com, meta.tab, sierra.tab*

3. Create a meta table (meta.tab) for the data to be processed as shown below.

File meta.tab:

```
VARS    DIR_NAME   TITLE PROC_FLAG SIERRA_FLAG   XP0    CLIP_THRESH   COLOR
FORMAT %s             %s    %d        %d            %5.4f %.2f           %s

       ../25mpm/1/    25-1  1         1             -31.8   2.0           #ffff00
       ../25mpm/2/    25-2  1         1             -31.8   2.0           #ffff00
       ../25mpm/3/    25-3  1         1             -31.8   2.0           #ffff00
       ../25mpm/4/    25-4  1         1             -31.8   2.0           #ffff00
```

The meta table is a Linux ASCII text file table with header information defining parameters for downstream processing scripts to extract relevant information about the sample data. Columns in the table are delimited by one or more whitespace characters. The 'DIR_NAME' field specifies the location of the data. The 'TITLE' field allows for the data to be given unique identifier, with the '- 'character serving as a field delimiter. The 'PROC_FLAG' specifies whether the data should be processed (1) or left untreated (0), which can be used when adding new data to previously processed collections. The SIERRA_FLAG indicates whether to use SIERRA processing (1) or normal processing (0). The XP0 field specifies the 0th order phasing to be applied to the direct dimension; if a value of 0.0 is entered, autophasing will be applied. The CLIP_THRESH is an optional field for applying noise thresholding below a given percentage of the maximum signal amplitude. The COLOR field is an optional field for choosing a display color for a given spectrum based upon a color name or RGB hex code.

4. For SIERRA data, provisionally process a representative spectrum using the 'basicFT2.com' NMRPipe function to identify residual excipient signals in need of additional SMILE based suppression during data processing. For gsHSQC data without excipient signals, skip to step 6.

```
basicFT2.com -xP0 Auto -yP0 -90 -xEXTX1 2.5ppm -xEXTXN -1.0ppm
nmrDraw -in test.ft2
```

5. Edit the *REG* entry in the *'sierra.tab'* input table to include the regions for SMILE signal suppression. If the residual excipient signal includes a distorted baseline, set the *TDN* parameter greater than zero (typically 6) to apply a local baseline correction during processing.

File sierra.tab:

VARS	REG	TDN
FORMAT	%s	%d
	'1.45 1.36 *18.3 19.9* **1.44 1.37** *23 24*.**2 1.87 1.8** *25 27.15*'	0

In the example three signals are set as targets, input as sets of ^{1}H (bold) and ^{13}C (italics) resonance positions (in ppm) that define a box around the signal used by the SMILE algorithm to identify the signals to remove. None of the signals require a baseline correction so the tdN parameter is set to 0.

6. Confirm the 'ref.com' script is correct for identifying the DSS signal, or set it to another desired signal for alignment.

File ref.com:

```
#!/bin/csh

set dirList = (`getTabCol.tcl -in meta.tab -var DIR_NAME`)
set origDir = `pwd`

@ i = 0

foreach d ($dirList)
   @ i++

   cd $d
   echo $d

   ref2D.tcl -in test.ft2 \
             -x1 -0.05ppm -xn -0.08ppm -y1 29.5ppm -yn 30.5ppm \
             -xRef 0.0 -yRef 30.0

   cd $oridDir
end
```

In this example, the DSS signal has been folded into the methyl window so its ^{13}C frequency is set based on the experimental spectral width (here, 30.0 ppm), assuming a true position of 0.0 ppm.

7. Run the 'all.com' script with the metadata table file as a command-line argument.

File all.com:

```
#!/bin/csh

setenv meta $1
clean.com
title.com
fid.com
proc.com
ref.com
interp.com
show.com
```

```
all.com meta.tab
```

8. After the interactive *specView* window opens (Figure 2A), zoom in on the region of interest for PCA.

   For methyl correlated spectra of mAbs this is typically from 2.2ppm to -0.75ppm in ^{1}H and 10 ppm to 28 ppm in ^{13}C.

9. Click the 'PCA icon' to perform PCA on the dataset (Figure 2B).

   Note that the data is not centered due to the sparsity of NMR spectral data.

10. In the PCA window, set the desired number of clusters based on your sample set, and click 'cluster' to perform complete linkage clustering (Figure 2B).

    Refer to the 'Statistical Analysis' section for details on complete linkage clustering.

11. To view loading spectra, click on the 'PCA Spectra icon' to open an NMRDraw window of the spectra. Use the 'Z' field to scroll through the different principal component loadings (Figure 3).

    The 1st PC loading (default display) is the average series spectrum. The 2nd and 3rd PC loadings generally contain the bulk of the variation of interest. Higher order components contain mostly noise.

Support Protocol 1

Data Visualization and Cluster Analysis

For many applications, it may be useful to perform additional cluster and statistical analyses on the PCA results from NMRPipe. Therefore, it may be desirable to export data into another program for further analysis. The results of the NMRPipe PCA are conveniently output as tab separated data files containing PC coordinates and cluster identities. This format may be imported into programs such as Matlab or other programming interfaces such as R or Python. Below, we demonstrate how to import PC data into Matlab, calculate cluster confidence integrals, evaluate cluster performance and plot the results using the example L-alanine concentration series from the Basic Protocol 2 section.

Materials:

Workstation running Matlab

Protocol steps

1. Open Matlab and navigate to the appropriate 'pca' subdirectory of the 'com' folder of the relevant data directory.
2. Import the PC coordinate information using the *importdata* function;

```
data=importdata('clust.tab');
```

3. Parse the input data from NMR Pipe to extract the desired PC coordinates (*e.g.*, 1 and 2), the cluster identities and the sample identities.

```
clusterID=sortrows(data.data(:,13));
noCls=max(clusterID);
pcCoord=sortrows(clusterID,data.data(:,7:8)]);
specID=sortrows([clusterID,string(data.textdata(6:size(data.textdata,1),3)]);
```

4. Parse data by cluster and determine confidence ellipses.

```
for idx=1:max(data.data(:,13));

   sample_no(idx)=sum((pcCoord(:,1)==idx));

   if idx==1
      parseD=pcCoord(1:sample_no(idx),2:3);
   else
      lowerB=sum(sample_no(1:idx-1))+1;
      upperB=lowerB+sample_no(idx)-1[;
      parseD=pcCoord(lowerB:upperB,2:3);
   end

   clsCent(idx,1)=sum(parseD(:,1))/sample_no(idx);
   clsCent(idx,2)=sum(parseD(:,2))/sample_no(idx);

   if size(parseD,1) == 1
      dEllx(idx)=0;
      dElly(idx)=0;
      thetaC(idx)=0;
   else
     pdistX=fitdist(parseD(:,1),'Normal');
     pdistY=fitdist(parseD(:,2),'Normal');

     clsIntX=(paramci(pdistX,'Alpha',intV));
     clsIntY=(paramci(pdistY,'Alpha',intV));

     dEllx(idx)=abs(clsIntX(1)-clsIntX(2));
     dElly(idx)=abs(clsIntY(1)-clsIntY(2));

    [maxY,Imax]=max(parseD(:,2));
    [minY,Imin]=min(parseD(:,2));
    minX=min(parseD(:,1));
    maxX=max(parseD(:,1));

    opp=(maxY-minY);
    adj=(maxX-minX);

     if parseD(Imax,1)<parseD(Imin,1)
        thetaC(idx)=(atan(opp/adj));
     else
        thetaC(idx)=(atan(opp/adj))*-1;
     end
   end
end
```

5. Perform silhouette analysis on clusters and plot results (Figure 4A).

```
pad_cell=[' ';' '];

silho_lab=specID(1:sample_no(1),2);

for idx=2:(cls)
     lab_append=specID((sum(sample_no(1:(idx-1)))+1):(sum(sample_no(1:idx))),2);
     silho_lab=vertcat(silho_lab,pad_cell,lab_append);
end

tick_space=(3:1:(size(clusterID)+2*cls));

silho=figure;
[silho_val,silho]=silhouette(pcCoord(:,2:3),clusterID);

hold on
set(gca,'ytick',tick_space,'yticklabel',silho_lab,'FontSize',16,'fontweight','bold',
    'fontname','times');

sum_silho=sum(silho_val);
```

6. Plot PCA results with confidence intervals (Figure 4B).

```
pca_full=figure;

el=-pi:0.01:pi;

scatter(pcCoord(:,2),pcCoord(:,3),100,clusterID);
    set(gca,'FontSize',20,'fontweight','bold','fontname','times');
    xlabel(['PC ' num2str(pComps(1))])
    ylabel(['PC ' num2str(pComps(2))])

hold on
scatter(clsCent(:,1),clsCent(:,2),50,'MarkerFaceColor','r');

hold on

for idx=1:cls
    cCos(idx,:)=clsCent(idx)+dEllx(idx)*cos(el);
    cSin(idx,:)=clsCent(idx,2)+dElly(idx)*sin(el-thetaC(idx));

    plot(cCos(idx,:),cSin(idx,:),'k:');
end

xlabel=(['PC' num2str(pComps(1)];
ylabel=(['PC' num2str(pComps(2)];
```

7. Optionally, export figures as encapsulated post script files and edit further in vector graphics software (Figure 4C).

COMMENTARY

BACKGROUND INFORMATION:

NMR spectroscopy has been a premier technique for studying protein higher order structure for over three decades. However, traditional NMR structural studies are predicated on isotopic enrichment with ^{13}C and ^{15}N, and for larger systems ^{2}H at non-labile sites to achieve adequate sensitivity. Furthermore, structural characterization by NMR has historically been limited to smaller molecular weight systems, with the vast majority of NMR structures in the PDB below 25 kDa (Jiang & Kalodimos, 2017). Given these limitations, conventional wisdom in the field discounted NMR as a suitable technique for higher order structural characterization of large protein therapeutics such as monoclonal antibodies, given both the lack of isotopic enrichment in drug products, as well as the large molecular weight of mAbs (150 kDa) (Berkowitz, Engen, Mazzeo, & Jones, 2012).

As such, HOS characterization of mAb drug products has relied heavily on FT-IR for secondary structure analysis as well as circular dichroism (CD) for secondary and tertiary structural analysis. While these methods are sensitive and suitable to provide structural information on mAbs, they have several drawbacks for HOS characterization of protein therapeutics. Practical hardware limitations of CD require dilute protein concentrations relative to typical mAb formulations (Kelly, Jess, & Price, 2005). Moreover, the degree to which results obtained at lower concentration are predicative of behavior at those higher formulation concentrations remains an open question. Likewise, formulation components can interfere with protein absorbance in CD, requiring deformulation (Lin et al., 2015). FT-IR, on the other hand, can operate over a wide protein concentration range and formulation conditions, but is a limited single-attribute method (2^o structure only) and thus does not provide a complete measure of protein therapeutic structure. The low spectroscopic and structural resolution of both techniques also makes it difficult to correlate spectral variance to structural variance. Perhaps, most importantly, recent studies have called into question the sensitivity of both methods to report on clinically relevant HOS variance (Wen et al., 2020).

In the last decade, with continued improvements in NMR hardware and methodology, sensitivity has been greatly improved, allowing for structural studies of systems as large as 1 MDa (Mainz et al., 2013) and of proteins at natural isotopic abundance (1.1% ^{13}C, 0.3% ^{15}N). Accordingly, there has been increasing appreciation that NMR can play a role in structural characterization of mAb and other protein drug products (Aubin, Gingras, & Sauvé, 2008; Freedberg, 2005; Marino et al., 2015) and can potentially overcome the limitations of traditional HOS methods such as CD and FT-IR. A number of NMR techniques have been proposed and demonstrated, broadly categorized as 1D ^{1}H methods (Chen et al., 2016; Franks et al., 2016; Kheddo, Cliff, Uddin, van der Walle, & Golovanov, 2016; Poppe et al., 2013), 2D ^{1}H-^{1}H correlated methods (Brinson & Marino, 2019; Japelj et al., 2016; Župerl, Pristovšek, Menart, Gaberc-Porekar, & Novič, 2007) and 2D ^{1}H-X heteronuclear correlated methods (Arbogast et al., 2016; Chen, Freedberg, & Keire, 2015; Singh, Bandi, Jones, & Mallela, 2017), such as the 2D ^{1}H-^{13}C methyl correlated method presented here. Due to the multi-modal nature of NMR, these methods allow for the HOS characterization of protein therapeutics, in formulation and at high concentrations. Furthermore, when high resolution ^{1}H-X correlated methods can be employed, HOS variance can potentially be described with high structural resolution. In addition to the NMR methods themselves, a number of chemometric methods for analyzing NMR spectra in terms of HOS variability have been put forth and are actively being explored allowing for quantitative assessment of spectral and structural similarity (Arbogast et al., 2017; Chen, Park, Li, Patil, & Keire, 2018; Japelj et al., 2016; Wang et al., 2020; Župerl et al., 2007).

CRITICAL PARAMETERS:

While the SIERRA filter depends on five parameters (^{1}H target frequency, ^{13}C target frequency, SIERRA contact pulse width, ^{1}H contact pulse power and ^{13}C contact pulse power), the target frequencies can easily be determined and sensitivity to frequency offsets is much smaller than uncertainty in peak position. Likewise, both the contact pulse widths and ^{1}H contact pulse power are determined directly from the $^1J_{HC}$ of the target, which is readily measured. This leaves only the ^{13}C contact pulse power to be determined and is the critical parameter for success of SIERRA. In addition to the $^1J_{HC}$, the optimum ^{13}C contact pulse power also depends on spin multiplicity of the target and possible third spin effects. In order to achieve the best results from SIERRA, a simple single parameter array of the ^{13}C contact pulse power is required. As coded in the pulse sequence, the ^{13}C contact pulse power includes a scaling-factor parameter (*cnst20*), set as a percentage from the theoretical optimum value for an AX spin-system (*i.e*. cnst20 = 100). For mAb methyl spectra, interfering excipient signals in the protein region of interest will most likely also be methyl-type AX_3 spin-systems, therefore the optimum *cnst20* value will be approximately 200. A coarse parameter array from 150-250 in increments of 10, followed by a fine parameters array of +/- 20 in increments of 2 around the estimated null crossing from the coarse parameter array should yield an acceptable setting for *cnst20* to achieve adequate signal suppression. It is not necessary to completely remove the signals as small residuals can be removed in post-processing by the SMILE filter.

In post-processing, it is critical that all spectra be processed with the same basic parameters (*e.g.*, window function, zero-filling, baseline correction), as any parameter that effects lineshape will influence PCA results. In addition to processing parameters, proper spectral alignment is absolutely critical as even frequency small offsets, typically observed with slight miscalibration of the water frequency, can have significant impact on PCA results. On the other hand, PCA is reasonably tolerant to small phase offsets, so either auto-phasing or manual phasing can be employed with reasonable confidence. PCA is also robust to experimental signal to noise and as previously demonstrated, data can be recorded with a limit of detection of only 3:1 S/N and still yield good results (Arbogast et al., 2017). However, if additional, peak-based analyses are desired it is recommended to collect data at or above the limit of quantification of 10:1 S/N. The effects of spectral resolution on PCA results has not been fully explored, however, it is essential that all data be collected with an identical acquisition times; thus far the majority of data used to demonstrate the method have been collected with 10 ms in the indirect ^{13}C dimension. For the SMILE filter, the bounds should be set tight around the residual signal(s) and any associated distortions. If the baseline is distorted locally around the residual signal, the 'tdN' parameters should be set to a small number, typically six, to correct this.

TROUBLESHOOTING:

If the SIERRA filter is not working as expected, it is important to check the optimization of the hard power levels and 90^{o} pulse widths of the ^{1}H and ^{13}C channels, since the ^{1}H and ^{13}C contact pulse power levels are determined from these values. Likewise, the tuning and matching on both channels should be checked. If the SIERRA filter results in loss of adequate water suppression, increase the length of the gradient, *p19*.

During batch processing, if the auto alignment fails, it is possible to manually correct for misalignment of spectra. To do so, the fid.com script needs to be edited as shown below.

To the declaration of variables add:

```
set dx   = (`getTabCol.tcl -in $meta -var DX_PPM`)
set dy   = (`getTabCol.tcl -in $meta -var DY_PPM`)
```

Edit the call to the Bruker *conversion script as follows:*

```
bruker -dxCAR $dx[$i] -dyCAR $dy[$i] -auto -nosleep -notk -exit >& conv.out
```

The meta.tab script can be edited to include the variables DX_PPM and DY_PPM with %5.4f format, and individual offsets can be entered for specific spectra.

STATISTICAL ANALYSIS:

Although one possible approach to statistical analysis of NMR data is outlined here, the use of PCA on the total spectral matrix with Complete Linkage Clustering, myriad potential approaches exist, and the field of chemometric analysis of biotherapeutic NMR spectra is only just emerging. For example, an alternative to inputting the total spectral matrix has been to use peak tables (Brinson 2017 & 2019). In general, peak position is invariant to both experimental approach and field strength; indeed, the precision of each peak was within 6 ppb regardless of pulse sequence, sampling strategy, and field in the multi-national interlaboratory study (Brinson 2019). For a controlled acquisition strategy, peak precision increased to 3 ppb. To implement this approach, a peak table is created in a reference spectrum and then copied to all other test spectrum. This initial step requires an expert analyst, is time intensive, and is possibly subject to analyst bias. Peak tables alone will also miss HOS perturbations as measured by changed in line width or appearances of new cross peaks. However, for protein drug substances for which HOS variations are read out with chemical shift changes, implementing PCA on peak tables could be a very useful tool.

By and large, the PCA method used in this paper, in which the total spectral matrix is employed as input, affords many advantages over peak tables. New cross peaks or any changes linewidths will be detected. Analyst subjectivity is removed by the implementation of automated methods. While the total spectral matrix is very sensitive to acquisition strategy and field strength, these concerns are mitigated, since highly controlled acquisition protocols are implemented at one field. If a new field is implemented at a later date, a bridging study could be performed to transfer a measurement protocol to the new instrument.

Pretreatment of the total spectral matrix needs to be minimal. In general, spectra should be normalized to the most intense cross peak in the spectrum unless absolute quantification is needed. With a controlled experimental strategy, this will remove the effects of a mis-set parameter, such as receiver gain. For the L-alanine results, an arbitrary value of 100 was used for the most intense peak. Other spectral pre-treatments have negligible or even deleterious effects. Normalization of intensity to a reference spectrum affords very similar results as compared to untreated spectra. Another data treatment, spectral binning, which is common for 1D ^{1}H spectra in the field of metabolomics, affords negligible benefits for 2D spectra.

The best practices for clustering of PCA results remains an open question. *k-Means* and *k*-medoids are partitional algorithms, which seek to minimize the intra-cluster distance from a centroid or medoid, respectively, have afforded mixed performance of peak tables from 2D spectra (Arbogast et al., 2017; Brinson et al., 2019). Hierarchical algorithms, such as the UPGMA and min-max group pairing algorithm, have given better performance. The complete linkage clustering (CLC) used herein represents the first report of this specific algorithm on NMR data and has been implemented within NMRPipe V9.8, and as such a brief explanation of the method is warranted. CLC, also called the farthest neighbor method, is an agglomerative hierarchical clustering method, whereby each spectrum begins in its own cluster. At each iteration the two clusters closest to each other are merged until all spectra are in one cluster. The salient feature of CLC is the cluster distance metric:

$$D(I,J) = \max d(i,j)$$

where I and J represent two clusters and i and j represent individual members of those respective clusters. Thus the cluster distance metric is defined as the farthest distance between members of any two given clusters. As such, CLC tends to produce small compact clusters (Massaro, 2005). As implemented in the NMRPipe, CLC will return the results of the iteration that contains the user-specified number of clusters. Here we have evaluated cluster performance using the silhouette value metric

(Rousseeuw, 1987). Values close to 1 indicate that a given observation fits well in the assigned cluster and poorly in the other identified clusters, i.e. is more similar to other members of its own cluster than to those of any other cluster. Values near zero indicate ambiguous cluster assignment or that it is equally similar to members of multiple clusters. Negative values approaching -1.0 indicate that the object is more similar to members of an alternate cluster. For both clustering algorithms and cluster performance metrics, there are very few reports applied to 2D-NMR spectra within a biopharmaceutical setting. Thus more work is needed to benchmark performance of the various methods as well as to determine best practices, including how to implement distance metrics to establish tolerance intervals (Brinson, Arbogast, Marino, & Delaglio, 2020).

UNDERSTANDING RESULTS:

If successfully employed on a series of data with a systematic variation (*e.g.*, concentration of excipient component), PCA should yield data arrayed sequentially in one or more principal components, and 2D plots of specific PCs will yield spectral response lines or curves. The exact nature of the response in PC space, however, will depend on the specific spectral response to the variation. Regardless, because data are not mean-centered, the 1st PC will always represent a continuum of spectra about the series average and will also account for the bulk of the explained variation (typically 85-99%). This value can be regarded as a measure of spectral similarity across the series (the higher the variance in the 1st PC, the more similar the spectra in the series). Note that in general, separation in PC1 is not in and of itself indicative of meaningful spectral differences. For a series where only a few peaks change in a correlated manner, it is expected that meaningful variance will be found primarily in the 2nd and perhaps 3rd PC. As increasing distinct sets of peaks have correlated changes in the series, additional PCs will contain variance of relevance.

Because the 1st principal component accounts for such a great magnitude of the total variation, scree plots are not particularly useful for assessing which higher PCs contain relevant spectra variation. Therefore, to assess the variance in each PC, it is necessary to view the loading plots (Figure 3). These plots will contain positive, negative and mixed phase mode signals at the frequencies of resonances that were perturbed across the series. In the 2nd PC, positive and negative mode signals represent an increase or decrease in intensity of a peak with respect to the series average, while mixed phase signals represent a shift in peak position or change in linewidth. If resonances are assigned, it is then possible to pinpoint this variation to specific residues in the protein, allowing a more in-depth understanding of the structural variation underlying the spectral perturbation. Once a PC loading plot is found to be dominated by noise and with no more than small residual signals from a large amplitude peak (Figure 3C), then the limit of relevant spectral variation has been reached. Only lower order PCs are then relevant for assessing structural (dis)similarity using clustering and distance metrics (see statistical analysis).

In the case study presented herein, spectral variance was limited to the first two PCs (Figure 3). Examination the 2nd PC loading plot (Figure 3B) reveals several resonances that vary across the series. To highlight how such loading spectra can be used to interpret data in terms of HOS, we consider one area of particular interest, -0.2 ppm to -0.4 ppm in ^{1}H and 19 ppm to 21 ppm in ^{13}C. As shown in Figure 5, a coalescence of two residues at increasing L-alanine concentration is observed. Although resonance assignments are not yet available for NISTmAb methyl spectra, it is known that these resonances emanate from the Fab domain. Using the SHIFTX2 chemical shift prediction software (Han, Liu, Ginzinger, & Wishart, 2011) with the 5K8A NISTmAb Fab fragment crystal structure, (Gallagher, Karageorgos, Hudgens, & Galvin, 2018), the signals have been putatively assigned to threonine-84 of the light chain and threonine-94 of the heavy chain. These two residues are proximal in HOS and are near to a cluster of tyrosine residues. Such a local structural environment could be indicative of a hot-spot for protein-excipient interactions. While the exact nature of these HOS perturbations in this region is unknown, one possible explanation could be changes in sidechain orientation and dynamics as a function of L-alanine concentration. As a result, differential aromatic shielding of the two residues would then be averaged out, leading to a convergence of frequency position of both threonines. This is supported by a similar observation for these resonances as a function of temperature (Brinson et al., 2019). However, we must note that this interpretation is presented only as an illustrative example and not as a verified structural characterization.

Using complete linkage clustering from the coordinates in the first two PCs, spectra were assigned to distinct groupings, which in this case correspond to the L-alanine concentration in each sample. Using the cluster identity, 95% and 99% confidence intervals were calculated from the distribution of the members of each cluster about the cluster center. As shown in Figure 4C, while each cluster is well separated in the 95% confidence interval (CI), there is little differentiation in the 2nd PC, with significant overlap of the 99% CI for the 0 mM, 5 mM, 10 mM alanine and 15 mM samples, while the 20 mM and 25 mM samples are better distinguished. To better contextualize these results, it is necessary to consider orthogonal information. Using the NanoTemper Tycho instrument, the three melting transitions (T_m) of NISTmAb were measured for at each concentration of L-alanine (Figure 6). As the concentration was increased to 15 mM alanine, the NISTmAb increased in thermal stability. As opposed to the NMR results, above this concentration, negligible changes in T_m were observed. Therefore, assuming the goal of a hypothetical formulation study was to maximize thermal stability, one might conclude that 15mM L-alanine was an appropriate formulation given the increases in T_m and, based on the PCA results, the minimal changes in HOS.

TIME CONSIDERATIONS:

Setup and optimization of the sample specific parameters (*p1* and *o1*) as well as SIERRA filter should take no more than one hour. Individual ^{1}H-^{13}C gsHSQC or SIERRA-gsHSQC spectra typically take 3-4 hours to run at 600 MHz using a cryogenically cooled probe for a sample at 40 mg/mL. For cases of lower sensitivity (*e.g.,* lower concentration), the total experimental time may take 2-4 times longer. If chemometrics will be applied, a series of replicate measurements takes approximately 12-20 hours; a data series for PCA typically requires at least 3, but preferably 4-5 distinct sample sets. Thus, the entire experimental spectrometer acquisition time for a PCA series is a minimum of 36 hours, but more typically 72-120 hours. If data acquisition time is limited and statistical analyses are planned, it is preferable to collect more data at lower signal-to noise than fewer data at higher signal-to-noise. Data processing setup (*e.g.*, importing data to workstation, setting up processing table files) should take 30-60 minutes. Data processing of standard gsHSQC or SIERRA-gsHSQC data generally takes <1 min per spectrum with auto-phasing and up to 5 targeted signals by SMILE. A typical data set therefore should take at most, 10-15 minutes to process.

ACKNOWLEDGEMENTS:

We acknowledge the support by NIST Biomanufacturing Initiative as well as NIST and W.M. Keck for support of Biomolecular NMR instrumentation at the Institute for Bioscience and Biotechnology Research.

DISCLAIMER

Certain commercial equipment, instruments, and materials are identified in this paper to specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the material or equipment identified is necessarily the best available for the purpose.

LITERATURE CITED:

Arbogast, L. W., Brinson, R. G., & Marino, J. P. (2015). Mapping Monoclonal Antibody Structure by 2D ^{13}C NMR at Natural Abundance. *Analytical Chemistry*, *87*(7), 3556–3561. https://doi.org/10.1021/ac504804m

Arbogast, L. W., Brinson, R. G., & Marino, J. P. (2016). Application of Natural Isotopic Abundance ^{1}H–^{13}C- and ^{1}H–^{15}N-Correlated Two-Dimensional NMR for Evaluation of the Structure of Protein Therapeutics. In *Methods in Enzymology* (Vol. 566, pp. 3–34). https://doi.org/10.1016/bs.mie.2015.09.037

Arbogast, L. W., Delaglio, F., Schiel, J. E., & Marino, J. P. (2017). Multivariate Analysis of Two-Dimensional ^{1}H, ^{13}C Methyl NMR Spectra of Monoclonal Antibody Therapeutics To Facilitate Assessment of Higher Order Structure. *Analytical Chemistry*, *89*(21), 11839–11845. https://doi.org/10.1021/acs.analchem.7b03571

Arbogast, L. W., Delaglio, F., Tolman, J. R., & Marino, J. P. (2018). Selective suppression of excipient signals in 2D ^{1}H–^{13}C methyl spectra of biopharmaceutical products. *Journal of Biomolecular NMR*, *72*(3–4), 149–161. https://doi.org/10.1007/s10858-018-0214-1

Aubin, Y., Gingras, G., & Sauvé, S. (2008). Assessment of the Three-Dimensional Structure of Recombinant Protein Therapeutics by NMR Fingerprinting: Demonstration on Recombinant Human Granulocyte Macrophage-Colony Stimulation Factor. *Analytical Chemistry*, *80*(7), 2623–2627. https://doi.org/10.1021/ac7026222

Berkowitz, S. A., Engen, J. R., Mazzeo, J. R., & Jones, G. B. (2012). Analytical tools for characterizing biopharmaceuticals and the implications for biosimilars. *Nature Reviews Drug Discovery*, *11*(7), 527–540. https://doi.org/10.1038/nrd3746

Brinson, R. G., Arbogast, L. W., Marino, J. P., & Delaglio, F. (2020). Best Practices in Utilization of 2D-NMR Spectral Data as Input for Chemometric Analysis in Biopharmaceutical Applications. *Journal of Chemical Information and Modeling*, *Just Accep*. https://doi.org/10.1021/acs.jcim.0c00081

Brinson, R. G., Ghasriani, H., Hodgson, D. J., Adams, K. M., McEwen, I., Freedberg, D. I., … Marino, J. P. (2017). Application of 2D-NMR with room temperature NMR probes for the assessment of the higher order structure of filgrastim. *Journal of Pharmaceutical and Biomedical Analysis*, *141*, 229–233. https://doi.org/10.1016/j.jpba.2017.03.063

Brinson, R. G., & Marino, J. P. (2019). 2D J-correlated proton NMR experiments for structural fingerprinting of biotherapeutics. *Journal of Magnetic Resonance*, *307*, 106581. https://doi.org/10.1016/j.jmr.2019.106581

Brinson, R. G., Marino, J. P., Delaglio, F., Arbogast, L. W., Evans, R. M., Kearsley, A., … Wikström, M. (2019). Enabling adoption of 2D-NMR for the higher order structure assessment of monoclonal antibody therapeutics. *MAbs*, *11*(1), 94–105. https://doi.org/10.1080/19420862.2018.1544454

Chen, K., Freedberg, D. I., & Keire, D. A. (2015). NMR profiling of biomolecules at natural abundance using 2D ^{1}H–^{15}N and ^{1}H–^{13}C multiplicity-separated (MS) HSQC spectra. *Journal of Magnetic Resonance*, *251*, 65–70. https://doi.org/10.1016/j.jmr.2014.11.011

Chen, K., Long, D. S., Lute, S. C., Levy, M. J., Brorson, K. A., & Keire, D. A. (2016). Simple NMR methods for evaluating higher order structures of monoclonal antibody therapeutics with quinary structure. *Journal of Pharmaceutical and Biomedical Analysis*, *128*, 398–407. https://doi.org/10.1016/j.jpba.2016.06.007

Chen, K., Park, J., Li, F., Patil, S. M., & Keire, D. A. (2018). Chemometric Methods to Quantify 1D and 2D NMR Spectral Differences Among Similar Protein Therapeutics. *AAPS PharmSciTech*, *19*(3), 1011–1019. https://doi.org/10.1208/s12249-017-0911-1

Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, J., & Bax, A. (1995). NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, *6*(3), 277–293. https://doi.org/10.1007/BF00197809

Franks, J., Glushka, J. N., Jones, M. T., Live, D. H., Zou, Q., & Prestegard, J. H. (2016). Spin Diffusion Editing for Structural Fingerprints of Therapeutic Antibodies. *Analytical Chemistry*, *88*(2), 1320–1327. https://doi.org/10.1021/acs.analchem.5b03777

Freedberg, D. I. (2005). Using nuclear magnetic resonance spectroscopy to characterize biologicals. *Developments in Biologicals*, *122*, 77–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16375252

Gallagher, D. T., Karageorgos, I., Hudgens, J. W., & Galvin, C. V. (2018). Data on crystal organization in the structure of the Fab fragment from the NIST reference antibody, RM 8671. *Data in Brief*, *16*, 29–36. https://doi.org/10.1016/j.dib.2017.11.013

Ghasriani, H., Hodgson, D. J., Brinson, R. G., McEwen, I., Buhse, L. F., Kozlowski, S., … Keire, D. A. (2016). Precision and robustness of 2D-NMR for structure assessment of filgrastim biosimilars. *Nature Biotechnology*, *34*(2), 139–141. https://doi.org/10.1038/nbt.3474

Han, B., Liu, Y., Ginzinger, S. W., & Wishart, D. S. (2011). SHIFTX2: Significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, *50*(1), 43–57. https://doi.org/10.1007/s10858-011-9478-4

Japelj, B., Ilc, G., Marušič, J., Senčar, J., Kuzman, D., & Plavec, J. (2016). Biosimilar structural comparability assessment by NMR: from small proteins to monoclonal antibodies. *Scientific Reports*, *6*(1), 32201. https://doi.org/10.1038/srep32201

Jiang, Y., & Kalodimos, C. G. (2017). NMR Studies of Large Proteins. *Journal of Molecular Biology*, *429*(17), 2667–2676. https://doi.org/10.1016/j.jmb.2017.07.007

Kelly, S. M., Jess, T. J., & Price, N. C. (2005). How to study proteins by circular dichroism. *Biochimica et Biophysica Acta*, *1751*(2), 119–139. https://doi.org/10.1016/j.bbapap.2005.06.005

Kheddo, P., Cliff, M. J., Uddin, S., van der Walle, C. F., & Golovanov, A. P. (2016). Characterizing monoclonal antibody formulations in arginine glutamate solutions using 1 H NMR spectroscopy. *MAbs*, *8*(7), 1245–1258. https://doi.org/10.1080/19420862.2016.1214786

Kovacs, H., Moskau, D., & Spraul, M. (2005). Cryogenically cooled probes—a leap in NMR technology. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *46*(2–3), 131–155. https://doi.org/10.1016/j.pnmrs.2005.03.001

Lin, J. C., Glover, Z. K., & Sreedhara, A. (2015). Assessing the Utility of Circular Dichroism and FTIR Spectroscopy in Monoclonal-Antibody Comparability Studies. *Journal of Pharmaceutical Sciences*, *104*(12), 4459–4466. https://doi.org/10.1002/jps.24683

Mainz, A., Religa, T. L., Sprangers, R., Linser, R., Kay, L. E., & Reif, B. (2013). NMR Spectroscopy of Soluble Protein Complexes at One Mega-Dalton and Beyond. *Angewandte Chemie International Edition*, *52*(33), 8746–8751. https://doi.org/10.1002/anie.201301215

Marino, J. P., Brinson, R. G., Hudgens, J. W., Ladner, J. E., Gallagher, D. T., Gallagher, E. S., … Huang, R. Y.-C. (2015). Emerging Technologies To Assess the Higher Order Structure of Monoclonal Antibodies. In J. E. Schiel, D. D. Davis, & O. V. Borisov (Eds.), *State-of-the-Art and Emerging Technologies for Therapeutic Monoclonal Antibody Characterization Volume 3.* (pp. 17–43). https://doi.org/10.1021/bk-2015-1202.ch002

Massaro, J. M. (2005). Clustering, Complete Linkage. In *Encyclopedia of Biostatistics*. https://doi.org/10.1002/0470011815.b2a13013

Poppe, L., Jordan, J. B., Lawson, K., Jerums, M., Apostol, I., & Schnier, P. D. (2013). Profiling Formulated Monoclonal Antibodies by 1 H NMR Spectroscopy. *Analytical Chemistry*, *85*(20), 9623–9629. https://doi.org/10.1021/ac401867f

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schleucher, J., Schwendinger, M., Sattler, M., Schmidt, P., Schedletzky, O., Glaser, S. J., … Griesinger, C. (1994). A general enhancement scheme in heteronuclear multidimensional NMR employing pulsed field gradients. *Journal of Biomolecular NMR*, *4*(2), 301–306. https://doi.org/10.1007/bf00175254

Singh, S. M., Bandi, S., Jones, D. N. M., & Mallela, K. M. G. (2017). Effect of Polysorbate 20 and Polysorbate 80 on the Higher-Order Structure of a Monoclonal Antibody and Its Fab and Fc Fragments Probed Using 2D Nuclear Magnetic Resonance Spectroscopy. *Journal of Pharmaceutical Sciences*, *106*(12), 3486–3498. https://doi.org/10.1016/j.xphs.2017.08.011

Tolman, J. R., & Arbogast, L. W. (2019). Selective spin inversion in solution by magic field cross polarization. *Journal of Magnetic Resonance*, *308*, 106588. https://doi.org/10.1016/j.jmr.2019.106588

Wang, D., Park, J., Patil, S. M., Smith, C. J., Leazer, J. L., Keire, D. A., & Chen, K. (2020). An NMR Based Similarity Metric for Higher Order Structure Quality Assessment among U.S. Marketed Insulin Therapeutics. *Journal of Pharmaceutical Sciences*. https://doi.org/10.1016/j.xphs.2020.01.002

Wen, J., Batabyal, D., Knutson, N., Lord, H., & Wikström, M. (2020). A Comparison Between Emerging and Current Biophysical Methods for the Assessment of Higher-Order Structure of Biopharmaceuticals. *Journal of Pharmaceutical Sciences*, *109*(1), 247–253. https://doi.org/10.1016/j.xphs.2019.10.026

Ying, J., Delaglio, F., Torchia, D. A., & Bax, A. (2017). Sparse multidimensional iterative lineshape-enhanced (SMILE) reconstruction of both non-uniformly sampled and conventional NMR data. *Journal of Biomolecular NMR*, *68*(2), 101–118. https://doi.org/10.1007/s10858-016-0072-7

Župerl, Š., Pristovšek, P., Menart, V., Gaberc-Porekar, V., & Novič, M. (2007). Chemometric Approach in Quantification of Structural Identity/Similarity of Proteins in Biopharmaceuticals. *Journal of Chemical Information and Modeling*, *47*(3), 737–743. https://doi.org/10.1021/ci6005273

INTERNET RESOURCES:

https://ibbr.umd.edu/sierra/home - repository for SIERRA pulse-sequence programs, processing scripts and protocols

https://www.ibbr.umd.edu/nmrpipe/index.html - NMRPipe home page with files and instructions for installation as well as protocols and demo data for usage.

https://www.ibbr.umd.edu/groups/nistmab-nmr - repository of NMR data, processing scripts and protocols related to NMR studies of NISTmAb.

FIGURE LEGENDS:

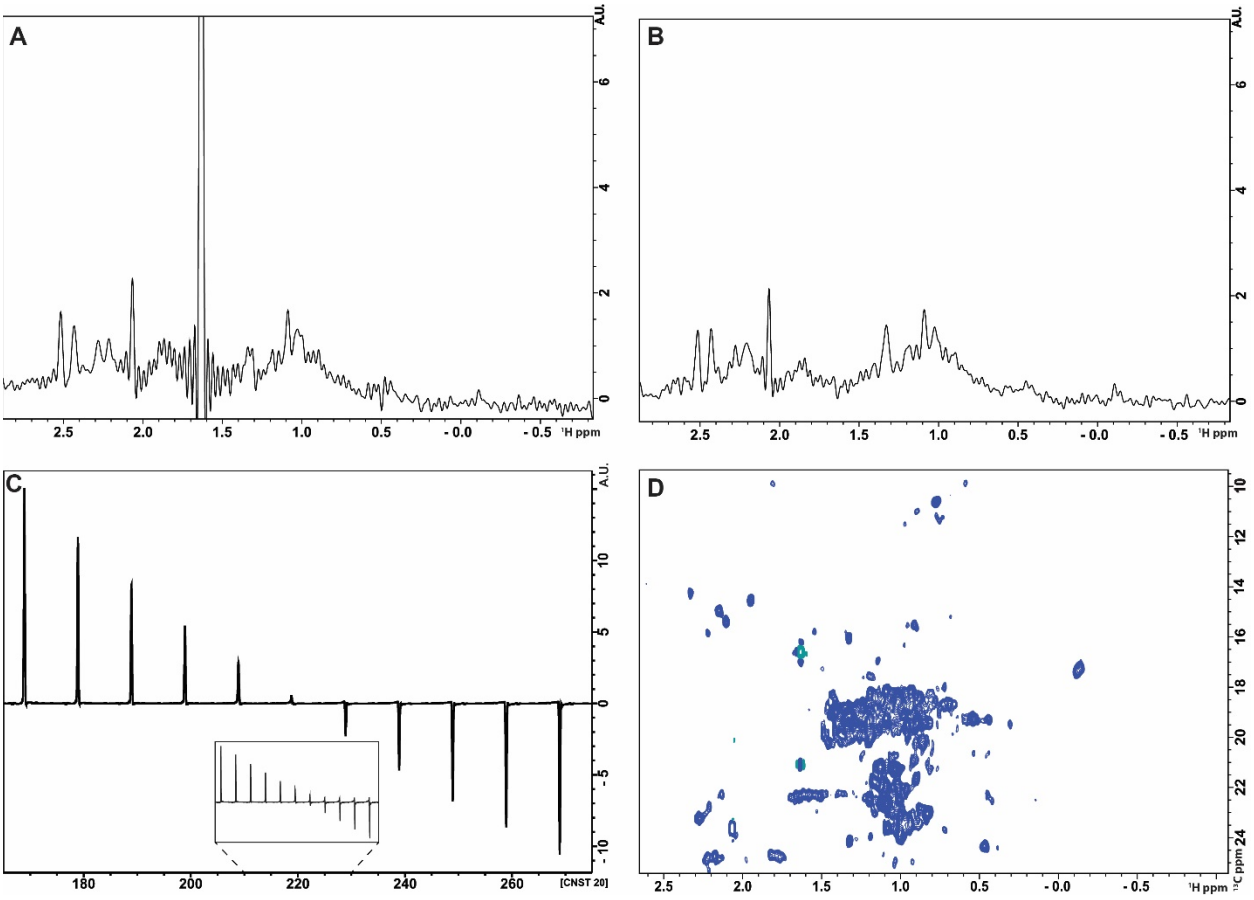

Figure 1. Experimental output during SIERRA optimization A. Test 1D ^{1}H-^{13}C gsHSQC spectrum without SIERRA filter of NISTmAb with 20 mM L-alanine (step 3). B. Test 1D ^{1}H-^{13}C gsHSQC spectrum with SIERRA Filter of NISTmAb with 20 mM L-alanine (step 6). C. Result of *popt* optimization of parameter CNST20. Values were arrayed from 160 % to 270 % in increments of 10 %. In the inset a fine array from 210 % to 230 % in increments of 2 % was used to determine the final experimental value of 223 (step 8). D. 2D ^{1}H-^{13}C gsHSQC spectrum with SIERRA Filter of NISTmAb with 20 mM L-alanine after optimization of the SIERRA parameters (step 10). A small residual L-alanine signal is observed. Additionally, a second, multi-bond correlation peak from alanine has been folded into the methyl window as well as the unsuppressed sodium acetate signal. These peaks, which are circled in panel D, will be removed by SMILE processing in Basic Protocol 2.

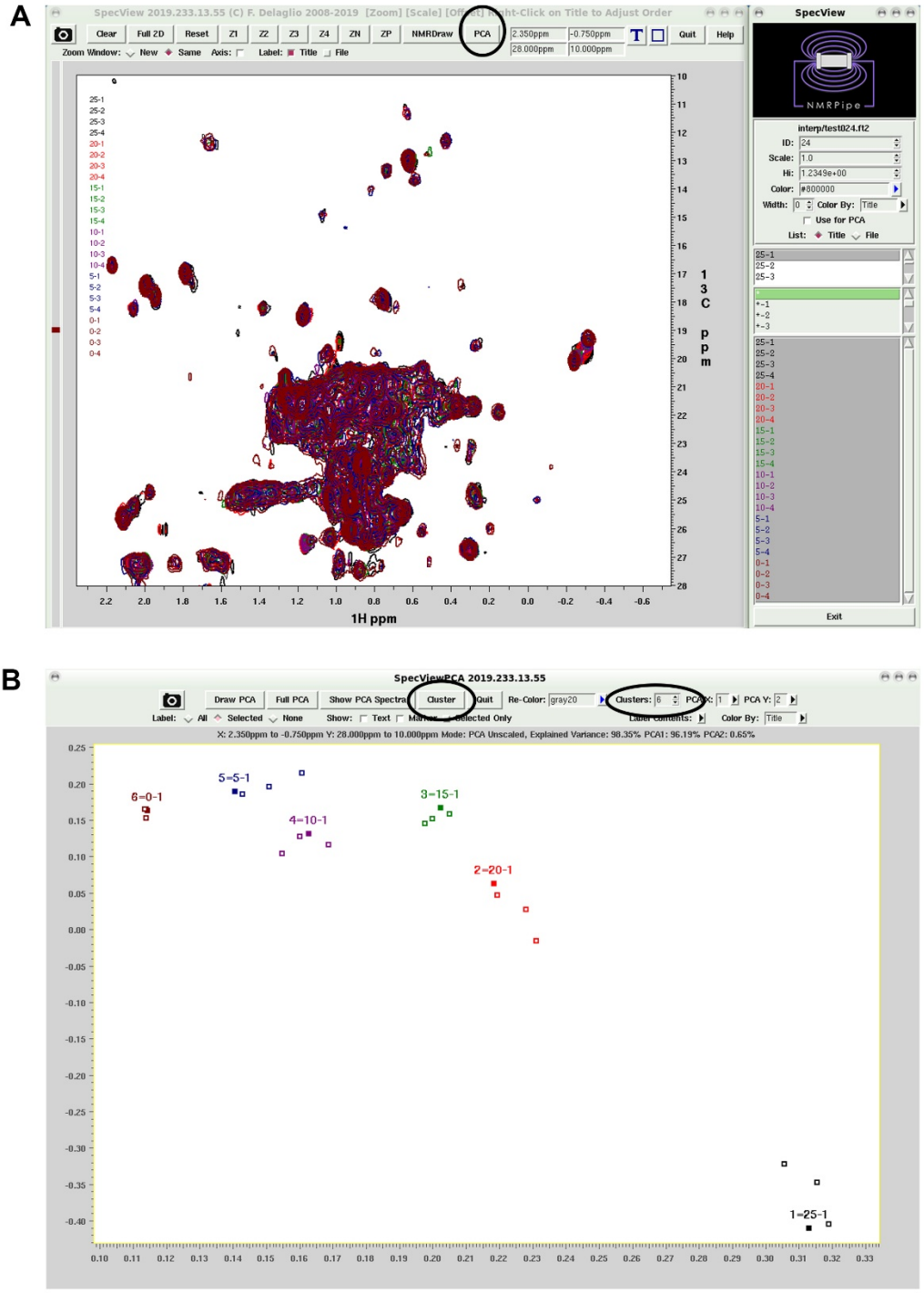

Figure 2 PCA using the NMRPipe SpecView function. A. The SpecView user interface. Selection of the spectral series data can be accomplished in the side pane on the right. The PCA button (circled) performs PCA on the displayed spectral region which can be adjusted interactively or preset in the *show.com* script and applied using the z-buttons. B. Plot of the first two principal components following PCA on the series of NISTmAb in six different L-alanine concentrations (0 mM, 5 mM 10 mM 15 mM, 20 mM and 25 mM). Data have been clustered by the complete linkage clustering method on using the built-in cluster function (circled), with the specified number of clusters.

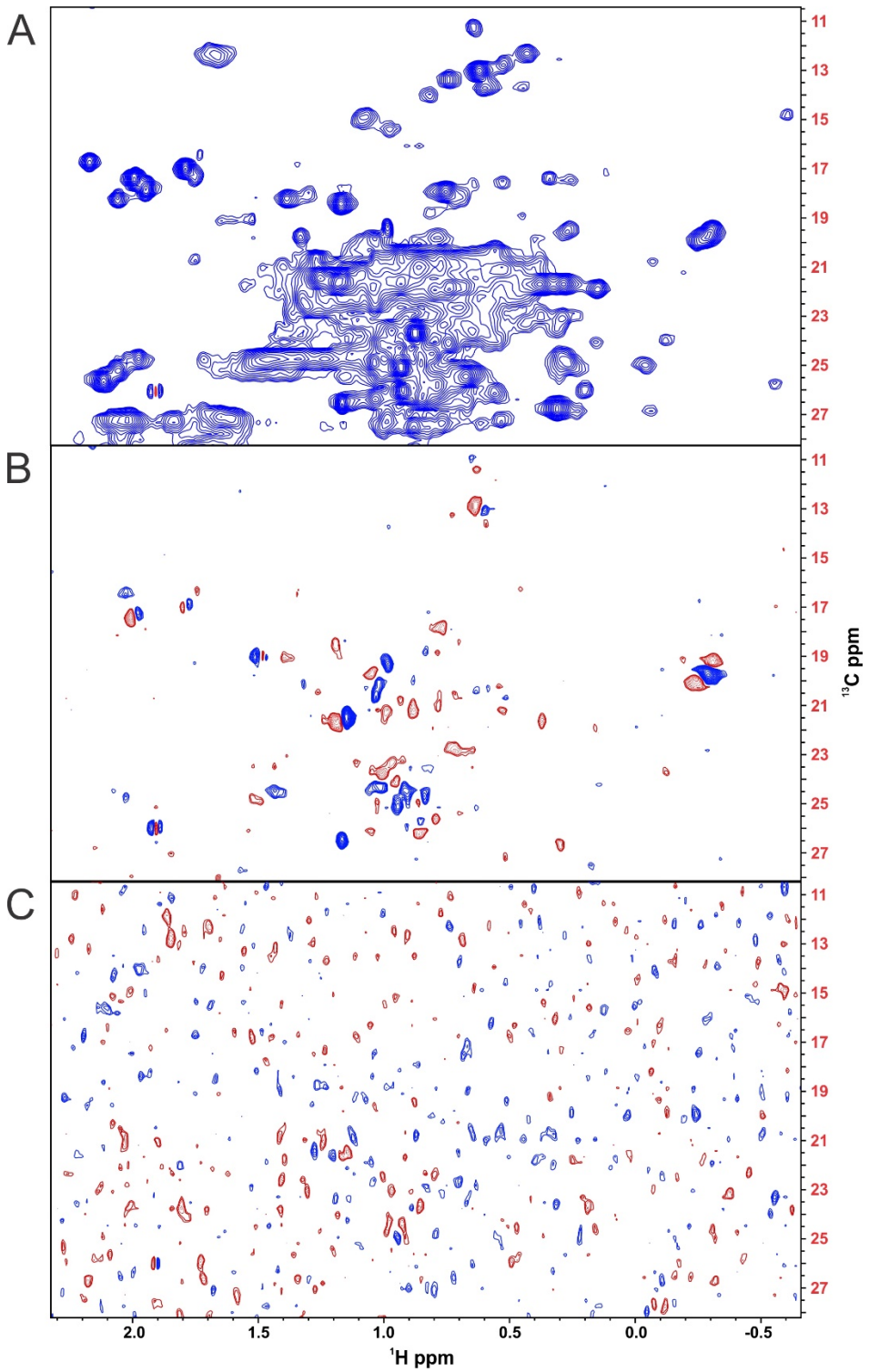

Figure 3. Principal component (PC) loading spectra from the L-alanine concentration series. A. 1St PC. The first PC represents the series average spectrum. B. 2nd PC. Signals that vary across the series appear in red/blue for frequencies of lesser/greater intensity relative to the series average. C. 3rd PC. In the L-alanine concentration series, relevant variation is confined to the 1st two PC and the 3rd PC contains only noise.

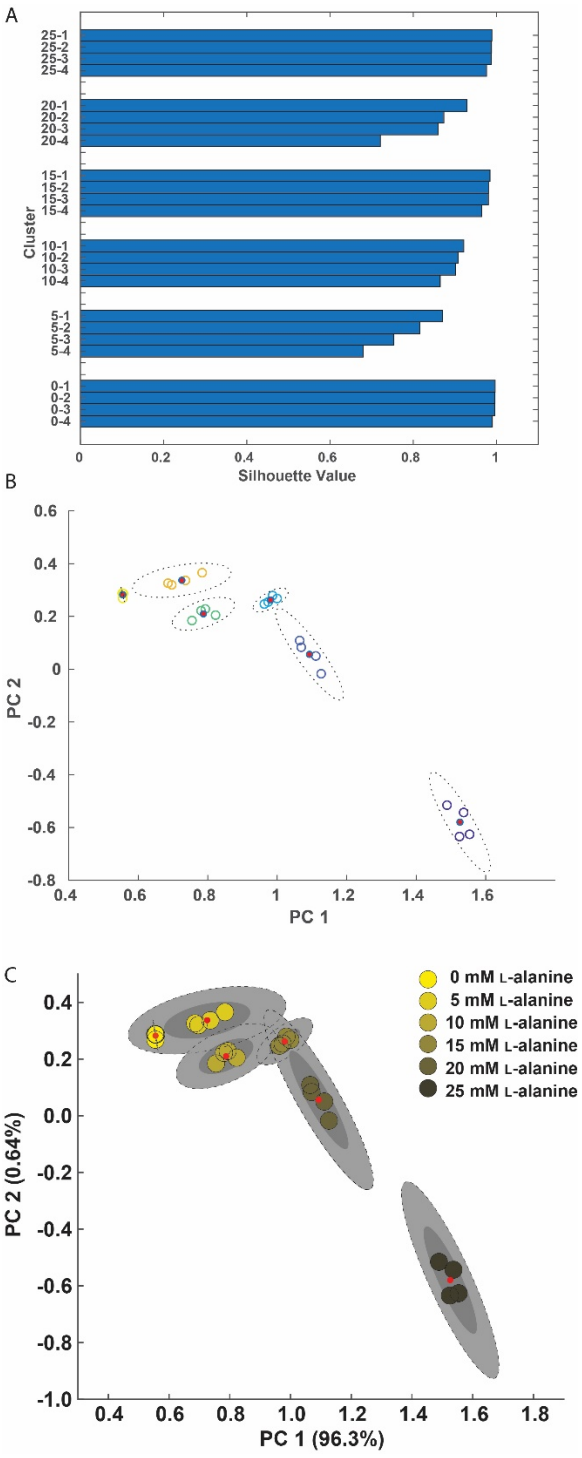

Figure 4 Results of the cluster analysis using Matlab. A. Silhouette plots for the six clusters identified by complete-linkage in NMRPipe. Values closer to 1 indicate better clustering. B. Raw output from Matlab plotting of the first two principal components from PCA performed in NMRPipe, with 95% confidence intervals indicated by dashed, empty ellipses. C. Edited plot of the first two principal components. 95% and 99% confidence intervals are shaded in dark and light grey respectively. Clusters are identified by color of their members as indicated in the legend. The image was manipulated using Adobe Illustrator. For panels B and C, red dots indicate the centroids of the clusters.

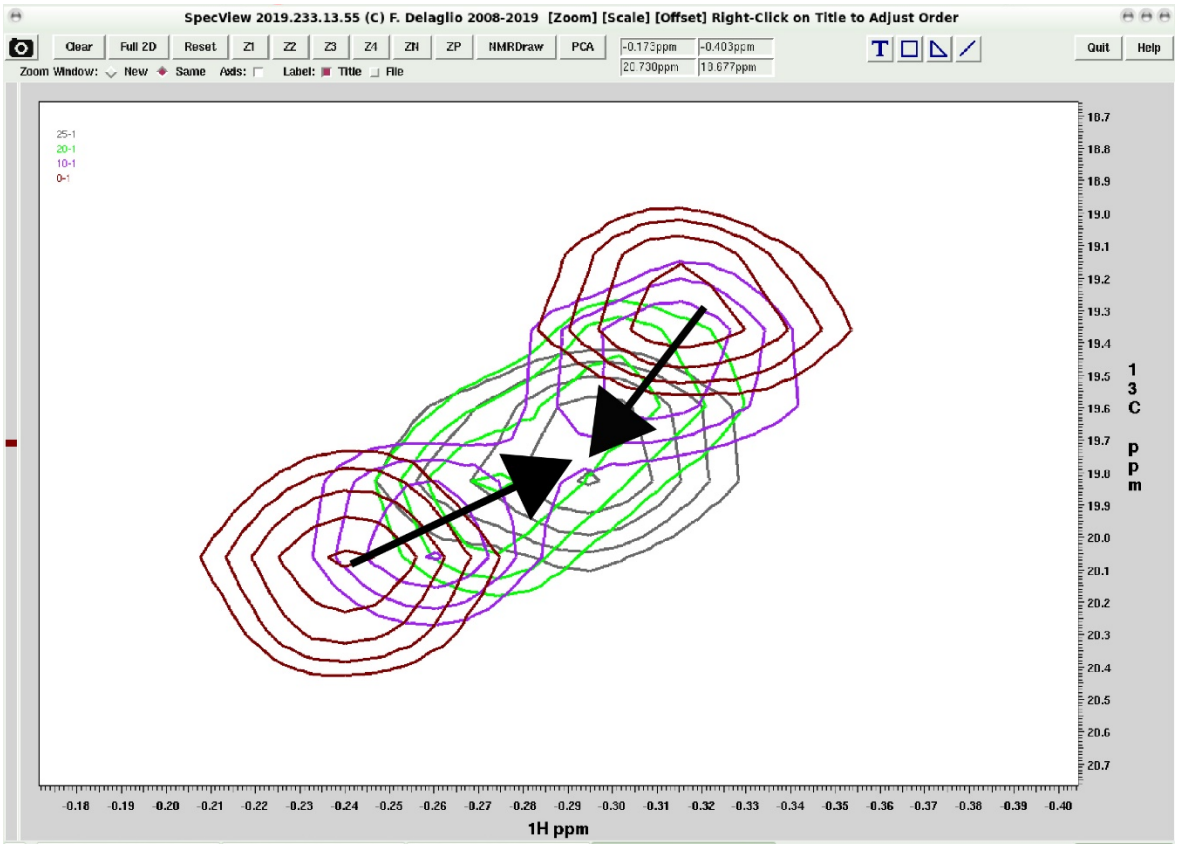

Figure 5 Extracted region of interest from the SIERRA-filtered ^{1}H-^{13}C gsHSQC methyl spectra from representative samples at 0 mM (brown) 10 mM (purple) 20 mM (green) and 25 mM (gray) L-alanine. Arrows denote the direction of change of the frequency positions of the two resonances going towards higher L-alanine concentration.

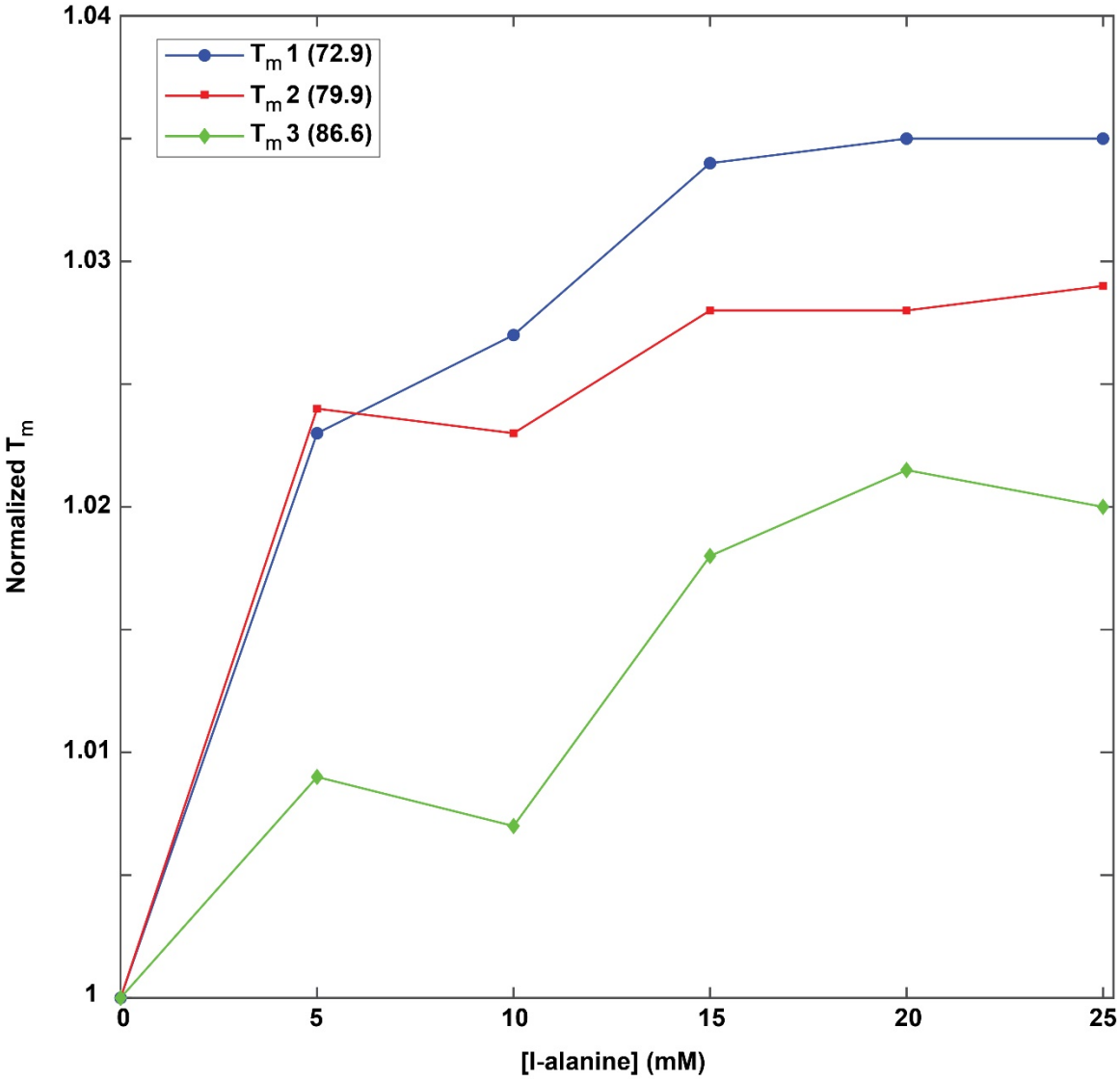

Figure 6 Thermal stability profiles as a function of L-alanine concentration for the three melting transitions (T_m1, T_m2, and T_m3) of NISTmAb. Melting transition temperatures at 0 mM L-alanine at shown in the boxed legend. An increase in all three transitions are observed up to 15 mM L-alanine, with diminishing gains at higher concentrations. Data are normalized to the T_m at 0 mM L-alanine.

TABLES:

Additional batch processing scripts run in all.com:

File Clean.com:

```
#!/bin/csh

set dirList = (`getTabCol.tcl -in $meta -var DIR_NAME`)
set proc    = (`getTabCol.tcl -in $meta -var PROC_FLAG`)
set origDir = `pwd`

set fList = (ext.dat ext.txt score.tab \
             ist.ft1 ist.ft2 scale.ft2 mask.fid mask.ft1 profY.dat ser_full clip.ft2    tmp.fid
             tmp.ft2 \
             test.fid test.ft1 test.tmp test.ft2 ps.ft1 mask.ft2 ext.dat ps.dat \
             orig.ft2 sierra.ft2 \
             phase1.out phase2.out conv.out proc.out)

@ i = 0


foreach d ($dirList)
   @ i++

   if $proc[$i] then

      cd $d

      echo "Cleaning $d ..."

      foreach f ($fList)
         if (-e $f) then
             echo " $f"
             /bin/rm -f $f
          endif
      end

      cd $origDir

   else

      echo $d/ keeping files

   endif
end

foreach d (`find ./ -type d -name pca\* -print`)

   if (-d $d/pca) then
      echo "Cleaning $d ..."
      /bin/rm -rf $d
   endif
end

foreach d (interp ext ascii_full ascii_ext)
   if (-d $d) then
      /bin/rm -rf $d
   endif
end
```

File title.com

```
#!/bin/csh

set dirList = (`getTabCol.tcl -in $meta -var DIR_NAME`)
set tList   = (`getTabCol.tcl -in $meta -var TITLE`)

@ i = 0

foreach d ($dirList)
   @ i++
   echo $d $tList[$i]
   echo $tList[$i] > $d/title
end
```

File fid.com:

```
#!/bin/csh

set dirList = (`getTabCol.tcl -in $meta -var DIR_NAME`)
set proc = (`getTabCol.tcl -in $meta -var PROC_FLAG`)
set origDir = `pwd`

@ i = 0

foreach d ($dirList)
   @ i++

   if $proc[$i] then

      cd $d

      if (!(-e title)) then
          echo None > title
      endif

      set tStr = (`cat title`)

      echo $d/test.fid $tStr

      bruker -auto -nosleep -notk -exit >& conv.out

      fid.com >>& conv.out

      sethdr test.fid -title $tStr

      $origDir/report2D.com test.fid

      echo ""

      cd $origDir

   else

      echo $d pre-processed

   endif
end
```

File proc.com:

```
#!/bin/csh

set dirList  = (`getTabCol.tcl -in meta.tab -var DIR_NAME`)
set proc     = (`getTabCol.tcl -in meta.tab -var PROC_FLAG`)
set SIERRA   = (`getTabCol.tcl -in meta.tab -var SIERRA_FLAG`)
set xP0List  = (`getTabCol.tcl -in meta.tab -var XP0`)
set clipList = (`getTabCol.tcl -in meta.tab -var CLIP_THRESH`)
set title    = (`getTabCol.tcl -in meta.tab -var TITLE`)
set origDir  = `pwd`

@ i = 0

foreach d ($dirList)
   @ i++


   cd $d

   set tStr = (`cat title`)

   if $proc[$i] then

      echo $d/ auto-processing

      if $xP0List[$i] == 0.0 then
                set xP0 = 'auto'
      else
                set xP0 = $xP0List[$i]
      endif

      if $SIERRA[$i] then

         set region = (`getTabCol.tcl -in $origDir/sierra.tab -var REG`)
                set base   = (`getTabCol.tcl -in $origDir/sierra.tab -var TDN`)

         basicSierra.com -in test.fid -ft orig.ft2 -out sierra.ft2 \
                         -reg $region -tdN $base \
                         -procArgs -xP0 $xP0 -xEXTX1 -0.75ppm -xEXTXN 2.5ppm \
                         -xBASEARG POLY, auto,ord=0 -yZFARG zf=2,auto -yP0 90

      else

         basicFT2.com -in test.fid -out orig.ft2 \
                     -xP0 $xP0 -xEXTX1 -0.75ppm -xEXTXN 2.5ppm \
                      -xBASEARG POLY,auto ord=0 -yZFARG zf=2,auto -yP0 90

          cp orig.ft2 sierra.ft2

      endif

      sethdr sierra.ft2 -title $title[$i]

      set maxVal = (`specStat.com -in sierra.ft2 -x1 3% -xn 97% -brief -stat vMaxAbs`)
      set scale = (`MATH 100.0/$maxVal`)

      nmrPipe -in sierra.ft2 -out sierra.ft2 -ov -inPlace -fn MULT -c $scale

      #uncomment for threshold noise removal
      #nmrPipe -in sierra.ft2 -out sierra.ft2 -ov -fn CLIP -below -$clipList[$i] \
      #        -above $clipList[$i] -inside 0.0

      nmrPrintf "%s/%s %s xP0: %.1f Scale: %.3e\n" $d sierra.ft2 $xP0List[$i] $tStr $scale

      $origDir/report2D.com sierra.ft2

      echo ""
      cd $origDir

   else
```

```
      echo $d/ using pre-processed data


   endif

   cd $origDir

end
```

File interp.com:

```
#!/bin/csh

set dirList = (`getTabCol.tcl -in $meta -var DIR_NAME`)
set origDir = `pwd`

if (-d interp) then
     /bin/rm -rf interp
endif

mkdir interp

@ i = 0

set refName = $dirList[1]/test.ft2


foreach d ($dirList)
   @ i++

   set tStr = (`cat $d/title`)
   set outName = (`nmrPrintf interp/test%03d.ft2 $i`)
   set inName  = $d/test.ft2

         endif

  echo $inName $outName $refName $tStr


          interpNMR -in $inName -out $outName -ref $refName

end

series.com interp/*.ft2
```

File show.com:

```
#!/bin/csh

set colorList = (`getTabCol.tcl -in meta.tab -var COLOR`)

if (`flagLoc $argv -multiColor`) then
   set colorArg = "-multiColor"
else
   set colorArg = "-colors $colorList -pcaArgs -pcaColors $colorList --"
endif

set bgCol  = "#ffffff"
set winCol = "#d3d3d3"

specView.tcl $* -in interp/*.ft2 -noscale -noshowSize -zoom1 2.3ppm -0.75ppm 28.0ppm 10.0ppm \
                -colors $colorList -bgColor $bgCol -winColor $winCol \
                -pcaArgs -pcaColors $colorList -bgColor $bgCol -winColor $winCol -- \
                 -pcaFont "-adobe-helvetica-bold-r-*-*-*-140-*-*-*-*-*-*" \

exit 0

set dirList = (`getTabCol.tcl -in meta.tab -var DIR_NAME`)
set inList = ""

@ i = 0

foreach d ($dirList)
   @ i++

   set inName = testsierra.ft2

   if (-e $d/$inName) then
       set inList = ($inList $d/$inName)

   else
      echo Missing $d/$inName

   endif

end

specView.tcl $* -noscale -ref None -hi 2.5 -in $inList -colors $colorArg
```