MDDGUI

Graphical User Interface to

Multidimensional Decomposition

Step by step guide from raw data to processed spectrum

AG & MK, UHN, Toronto, Canada

Introduction

Multidimensional decomposition (MDD) is a mathematical tool for processing 3D or higher dimensionality NMR spectra. For theoretical background, examples and applications, please see the references. For the purposes of this text, it is important to understand that MDD is able to reconstruct a complete spectrum from approximately 30% of FIDs that are actually recorded, the so called sparse or non-uniformly sampled (NUS) spectrum. The program MDDGUI combines a few scripts to prepare the input and process the output of the actual MDD calculation. So far it works only for 3D spectra. The whole process currently has 8 steps as follows

1. Insert zeros into the Bruker SER or Varian FID file. As we currently use NMRPIPE in some steps of the processing, it is required that the recorded FIDs are placed in their proper positions as if all 100% of the spectrum were recorded, not just the 30%.

2. Convert the newly produced SER-/FID-style file into NMRPIPE format (BRUK2PIPE or VAR2PIPE).

3. Find the proper phase corrections in the acquisition (directly detected) dimension by processing the first plane of the spectrum (it is complete for Bruker, but not Varian data sets). Uses NMRPIPE.

4. Fourier transform the acquisition dimension for the whole spectrum and extract the region to be calculated, *e.g.* ^{15}N, aromatic or aliphatic (NMRPIPE).

5. Convert the newly produced NMRPIPE file into the format understood by MDD software.

6. Calculate and save the result.

7. Use the result from step 6 to reconstruct missing data.

8. Fourier transform the indirect dimensions.

1   Recording the data

1.1   On Bruker TopSpin

When recording the non-uniformly sampled spectrum, one needs to indicate which FIDs are to be recorded. In our lab MK and CF rewrote the standard pulse sequences to use an additional file, VDLIST, which does exacly that. These pulseprograms usually have a suffix .NUS and before you launch the data acquisition, you should run a script from within TopSpin, called SPARSE on our 800MHz and 500MHz instruments or SPARSE600 on the 600MHz. These scripts generate the VDLIST file. If the entry in VDLIST is '1u', an FID pair (*Re* and *Im* or *Echo* and *AntiEcho*) is skipped, if the entry is 3u, it is recorded and stored in the SER file. To be able to process the spectrum you must have the correct VDLIST file.

The first step of the processing is thus rearranging the recorded FIDs and putting them into a new file, similar in structure to SER. Fill out the required fields on the first work page of the MDD program. Most often, you would need to use the prefilled values. You need two input files: SER and VDLIST and one output file SER.SP. Press the INSERT ZEROS button and check the console window for any error messages. Then move to the next page. For the purpose of clarity, this text will assume that the FIDs are now rearranged into SER.SP file.

1.2   On Varian VnmrJ with BioPack

Latest versions of Varian BioPack have a few options to include the non-uniform sampling. You need to read the help available from within VnmrJ and follow instructions for Orekhov's procedure. Any BioPack pulse sequence can be modified to include non-uniform sampling. In the end, before processing the data, you should have the modified pulse sequence file (PULSE_SEQUENCE_S.C) as well as a number of additional files, most importantly (PULSE_SEQUENCE_S.HDR_3), where PULSE_SEQUENCE signifies any standard BioPack sequence. Use macro BPSVF to save the data, as regular SVF will omit these necessary files.

From here, the processing follows same logic as that for Bruker data. Home-built software reads the recorded FID, PROCPAR and other relevant files and inserts zeros in the right places, so that the final FID.SP file has the actually recorded FIDs in the order they would be if the 100% of the final spectrum were recorded.

2   Convert SER.SP/FID.SP into NMRPipe format

The second step is to convert the Bruker-/Varian-style data into NMRPipe format. On the second page of the MDD wizard edit a FID_SP.COM script for BRUK2PIPE/VAR2PIPE conversion just as you would do for a regular spectrum. Most parameters should be read correctly already, but double check the acquisition mode for the indirect dimensions as these are often difficult to get right in an NUS experiment. Make sure that

- Input file is **./SER.SP** or **./FID.SP** (or the name you gave in Step 1).

- Double check the number of points in the indirect dimensions. You should use the values for complete spectrum. E.g., if you are recoding 30% of 300 x 50 data points, you should still use 300 and 50 for the corresponding dimensions in -yT and -zT fields. Change the -yN and -zN fields to corresponding number of real points (in this case 600 and 100).

- Double check the acquision modes. The automatic parameter readout sometimes places Echo-AntiEcho detection mode into a wrong dimension.

Apparently, with a newer version of TOPSPIN (2.0) there might be some trouble with the DECIM filter when converting. You might need to install the most recent version of NMRPIPE program.

Save the conversion script and perform the conversion by pressing the corresponding buttons. Then move to the next step.

3   Process the First Plane of the Spectrum

The third step is required to determine the processing parameters for the directly detected dimension. To allow to do so, the first plane is currently recorded completely for Bruker data

sets, i.e. all FIDs are present. For Varian data sets it is not the case, but one can still use the incomplete first plane to determin the phasing parameters, or even record the first plane separately and determine the phasing that way. Edit the prefilled NMRPipe script to suit your case. E.g., change input file, region of interest, phase corrections, etc. You might need to create a new directory for the output file. This step is interactive, in the sense, that you can modify the necessary parameters a few times and check the result with nmrDraw, just as you would normally do with a regular spectrum. Script lines between the capitalised comment lines are copied for the next step's script, which is why DO NOT DELETE OR MODIFY these comment lines. When you are satisfied with the phasing, move to the next step.

4   Process the Directly Detected Dimension

In the fourth step only the directly detected dimension is Fourier transformed for the whole spectrum. The indirect dimensions are incomplete and should not yet be transformed. The script is copied from the previous step, but you should still double check the relevant parameters, phase corrections being the most important. In this step you also make the decision on the size of the region you wish to process, thus change the corresponding line accordingly. Save the script and process the spectrum. The output should be in the form of a single file rather than set of planes, so don't change that. When done, move to the next step.

5   Convert NMRPipe File into MDD Format

In this step you need to produce a file directly used as input for the actual calculation program. You need to specify the input file (i.e. output from Step 4) and the output file (i.e. input for Step 6). There are two versions of the calculation software: one uses MPI for parallelization of the task on multiple processors, while the other can only use one processor. MPI version has the advantage of treating the whole spectrum in one go, but it is usually much slower. It is recommended that you use the single processor version. Since majority of NMR spectra are too large for only one processor, it is necessary to subdivide the spectrum into smaller overlapping regions and treat each region separately. Each region contains complete indirect dimensions and a portion of the acquisition dimension. E.g., if you process ^{15}N-NOESY-HSQC (say, between 11 and 5.5 ppm in direct ^{1}H) and would like to subdivide the spectrum into 16

4

regions, each region would be just under 0.5 ppm wide. Some overlap is usually desirable, so that edges of each region could be discarded later, when the result is put back into one spectrum. It is recommended to use region size of around 0.3-0.4ppm and the overlap to be on the order of 0.1ppm. These values will be adjusted to make the regions with integer number of points.

You can optionally provide a schedule file, which is produced together with the VDLIST. There is no schedule file for Varian data. Press the CONVERT TO MDD button and check the console window for progress output and any error messages. Then move forward.

6   MDD Calculation

The actual MDD calculation is taken out of the GUI wizard because it can take a long time to perform and it may be annoying to keep the graphical program idling. In addition, you may wish to perform the calculation on a more powerful machine or a cluster. You should thus transfer the output file from Step 5 to the desired location and start the calculation. The detailed instructions are repeated in the information window of the user interface and are given here for reference. You can fill out the relevant fields below the information window, and press the CREATE COMMAND FOR CALCULATION button. This will create a file with the correct command, which you can run or copy into command line.

1. If your dataset is large you may wish to login to the cluster. Please, check with your system administrators about the details of initiating a calculation on multiple processors

2. For single processor, the command is 'MDDNMR1.2C <MDD FILE> <NO. OF COMPONENTS> 0 <MAX ITERATIONS> <EPS> <LAMBDA> <RANDOM SEED OR INITIAL SOLUTION> <RESERVED MEMORY>'. Parameters are explained below. The created script takes a file with the list of regions as an argument and automatically starts a calculation process for each region. Unless you know what you are doing, do not modify the file containing the list of regions as it is required for Step 7. If any process needs to be restarted, copy the list into another file and modify accordingly. You can, of course, use the general command and start each calculation in each subfolder.

   - Here <MDD File> is the input file (output from Step 5)

- <No of Components> is the maximum number of either ^{15}N or ^{13}C-HSQC peaks in each region of the spectrum + 10-20% in most cases (such as ^{15}N-NOESY-HSQC, CBCA(CO)NH or HCCH-TOCSY). It is a safe bet to have more components than necessary.

- 0 – A zero (ask no questions!)

- <Max Iterations> is the number of iterations you wish to perform. Somewhere between 250 and 500 is usually a good number. You can always continue with more iterations if necessary

- <Eps> is a convergence criterion. Use 0 (or 1e-8)

- <Lambda> is the Tikhonov regularization parameter. Use 0.01

- <random seed or initial solution> is either a file containing initial approximation (i.e. if you restart a calculation after a blackout or just want to add more iterations) or a random positive integer

- <reserved memory> is given in Mb. Necessary to store both the data and solution. Increase if your calculation crashes without starting the iterations.

3. For multiple processors, the command is 'MPIRUN -NP <NO OF PROCESORS> MDDN-MRP_UHN <MDD FILE> <NO. OF COMPONENTS> <MAX ITERATIONS> <LAMBDA> <RANDOM SEED OR INITIAL SOLUTION> <RESERVED MEMORY>'

- <No of Components> is the total number of corresponding HSQC peaks in the spectrum + 10-20%

- <No of Processors> - obvious, no? Anything above 1

- Note that there is no zero after the number of components and that parameter <Eps> is also absent

- Other parameters are the same as for a single processor run

4. The result of the calculation is by default stored in file RES.DAT. Copy it to a file of your liking. This unfortunately puts a limitation - you may start only one calculation from each directory, as otherwise RES.DAT would get overwritten by multiple runs. RES.DAT may be used to restart the calculation as <random seed or initial solution> parameter above. Note that somewhat newer MDDNMR1.2C program (for single processor) produces a slightly different file format, which includes some comments and size information in

the beginning. You do not need to be concerned with that unless you want to restart your calculation from a previous solution and switch from sinlge to multiple processors or *vice versa.*

7 Converting back into NMRPipe Format

In this step you have to transfer the results from the res.dat file (or its copy) back to NMRPipe format, filling in the missing FIDs by calculated values. The input files are the output from Step 4 (same as input for Step 5) and res.dat file for MPI version or the region list file for the multiregion version. The output is a new reconstructed NMRPipe file with the directly detected dimension Fourier transformed. Press the RECONSTRUCT button and check for progress output and any error messages in the console window. Move to the last step.

8 Fourier Transform of the Indirect Dimensions

This is the final step of the processing. A sample script is given, which you sould modify to suit the circumstances and then use to finalize the processing, just as you would with an ordinary spectrum.

Finally, the whole processing produces a number of temporary files that can be deleted once you are happy with the final spectrum. You can delete SER.SP/FID.SP, all *.MDD and RES.DAT files. A single 3D NOESY spectrum can easily waste 300Mb of disk space in these kind of files.

References

[1] Orekhov VY, Ibraghimov IV, Billeter M. *MUNIN: a new approach to multi-dimensional NMR spectra interpretation*. J Biomol NMR. (2001) **20** pp.49-60.

[2] Gutmanas A, Jarvoll P, Orekhov VY, Billeter M. *Three-way decomposition of a complete 3D 15N-NOESY-HSQC. J Biomol NMR.* (2002) **24** pp.191-201.

[3] Orekhov VY, Ibraghimov I, Billeter M. *Optimizing resolution in multidimensional NMR by three-way decomposition*. J Biomol NMR. (2003) **27** pp.165-73.

[4] Luan T, Jaravine V, Yee A, Arrowsmith CH, Orekhov VY. *Optimization of resolution and sensitivity of 4D NOESY using multi-dimensional decomposition*. J Biomol NMR. (2005) **33** pp.1-14.

[5] Jaravine V, Ibraghimov I, Orekhov VY. *Removal of a time barrier for high-resolution multidimensional NMR spectroscopy*. Nat Methods. (2006) **3** pp.605-607.

[6] Jaravine VA, Orekhov VY. *Targeted acquisition for real-time NMR spectroscopy*. J Am Chem Soc. (2006) **128** pp.13421-13426.

[7] Jaravine VA, Zhuravleva AV, Permi P, Ibraghimov I, Orekhov VY. *Hyperdimensional NMR spectroscopy with nonlinear sampling*. J Am Chem Soc. (2008) **130** pp.3927-3936. Erratum in: J Am Chem Soc. (2008) **130** p.13182.