

CONNJUR Workflow Builder: Tutorial

Part I: Introduction

1.1 About CONNJUR and CONNJUR Workflow Builder (CWB)

The CONNJUR Project is to develop an open source integration environment for biomolecular NMR data analysis. CONNJUR's goal is not to build NMR data processing/analysis tools, rather to integrate them. As such, CONNJUR does not handle the licensing, installation and maintenance of the various third-party tools, such as NMRPipe and the Rowland NMR Toolkit – those tasks are left to the user. These various tools are required by CONNJUR for useful NMR data processing; however, CONNJUR runs independently of these tools.

CONNJUR is developed by a community of NMR spectroscopists and scientific programmers who aim to create and maintain NMR analysis tools bounded by the goals of excellent end-product, free of charge, open source software using the highest standards.

CONNJUR is being actively developed! That means that there may be some unexpected bugs. It also means that if you encounter an unexpected bug and report it to us, it is likely the bug will be fixed quickly. We also encourage feedback and communication. If there is a feature you especially like or hate, or if there is something you wish was implemented, let us know.

Where to download: <http://www.connjur.org/downloads>

1.2 Additional Information

Additional information on the CONNJUR project can be found at the following websites:

<http://www.connjur.org>

<http://sourceforge.net/apps/mediawiki/connjur>

There are also several publications on the CONNJUR project which can be accessed through Pubmed Central. A listing is found on the CONNJUR website.

Part II: Using CWB

2.1 Starting Workflow Builder

The CONNJUR-WB application consists of two files: the java-based 'jar' file which has the actual application and a startup script which runs the application. Both of these files should be in your "bin" directory, which is in your path. Therefore, all you need do is type `workflowbuilder` and the application should start.

2.2 Database

While CONNJUR-WB can be run in "Demo mode" without a database, the maximum utility of CWB requires data stored in a database. The database can be selected in the Preferences->Database Menu. Required information for establishing the database connection are hostname (i.e. computer name hosting the database), port (typically 3306), and the Username and Password authorized to connect to database. During this tutorial we will use a database pre-installed on your individual machines. Therefore, these parameters should read:

Hostname	localhost
Port	3306
User Name	Workshop
Password	xxx

When changing any preference information such as the database connection, CONNJUR-WB must be exited and restarted for changes to take effect.

2.3 Reconstruction layout

The reconstruction layout is the large left-hand window in the CONNJUR-WB application. It is within this window where you will create/load/run your spectral reconstruction tasks. The right-hand window contains an actor browser which will allow you to select reconstruction tasks to assemble for your reconstruction workflow. When you first open CONNJUR-WB, all options are greyed out as there is no dataset yet to be operated on.

2.4 Creating a New Reconstruction

To create a New Reconstruction select Reconstruction->New. At this point you will be prompted to select the format of the data you will be importing so the proper Importer can be invoked. Currently supported formats include "Agilent/Varian", "Bruker", "NMRPipe", and "Rowland Toolkit". The Rowland NMR Toolkit will be covered later in the workshop. Let's start by importing a 'Varian' file set.

2.5 Actors

The reconstruction workflow will consist of a series of “actors” (triangles, diamonds and circles) and datasets (squares). The simple rule is that an actor takes a dataset as input and produces one as output. However, at the start of the workflow no dataset is uploaded into the system yet, and so the first actor will stand alone. An actor can be thought of as analogous to a “function” in NMRPipe processing, in that it manipulates the data in some manner. However, there is not necessarily a direct correspondence between an individual actor and an individual NMRPipe function.

2.6 Importers

As mentioned before, to initiate a reconstruction workflow we must import the data from elsewhere. For this reason, the Import actors play a very important and unique role. Each reconstruction begins with an Import actor, and at the time being, the format of the imported data must be defined by selecting the proper Import actor. Don't fret, if you wish to apply a workflow to a different dataset of a different format, you will learn later how to save and reload workflows so they may be applied to different datasets.

2.7 Configuring the Varian Importer

Double click the Varian Import Actor (Purple triangle). A pop-up window should appear which will allow you to configure the Import Actor. The first step in configuration is to select with the file browser the dataset you wish to import. This must be done first as CONNJUR-WB will examine the file to assist with the available configuration choices for import. Varian data is stored as a directory of files, not a single file. Therefore, for the Varian importer you wish to select the directory in which the data is stored – not the actual “fid” file. This behavior is different for other formats of spectral data.

Parameter	Value	Units	Range
sensitivity Enhancement On Dimension	2	(Rance Kay)	
negate Imaginaries	<input checked="" type="checkbox"/>		
dimension One Channel	1		
dimension Two Channel	3		
dimension Three Channel	not present		
dimension Four Channel	not present		
directory*	ome/workshop/CONNJUR_data/nhsqc.fid		
ignore Extra Data	<input type="checkbox"/>		

Tool Arguments:

Info Error Detail Online help Online wiki Bypass actor

Fid input not available OK Cancel

Use the file browser to select our example HSQC dataset in the CONNJUR_data directory (filename: nhsqc.fid). The directory window should now show the directory path to that data. You will now see that the window has reconfigured such that we have options to select a dimension for sensitivity enhancement correction, as well as establishing dimension/channel associations for dimensions 1 and 2. As CWB has read the procpa data from the fid directory, CWB knows this is a 2D dataset and therefore dimension associations for dimension 3 and 4 are listed as “not present”.

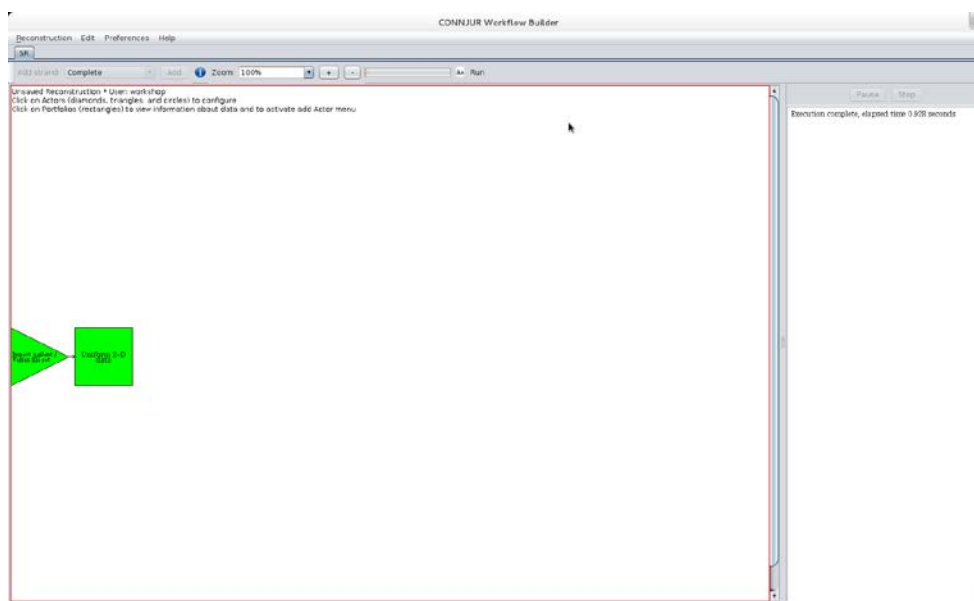
The individual dimensions to a spectrum are associated with individual nuclei, which on the spectrometer are determined by the amplifier channels. Dimension 1 is typically ^1H which is also typically set as channel 1. Channels 2 and 3 are set somewhat arbitrarily; in our labs, channel 2 is typically ^{13}C and channel 3 is ^{15}N . So for this N-HSQC, dimension 2 corresponds to channel 3 (Nitrogen). You should set “dimension Two Channel” to “3”. You should also set “sensitivity Enhancement on Dimension” to 2.

Finally, click OK to finish configuring the Importer.

2.8 Running Actors

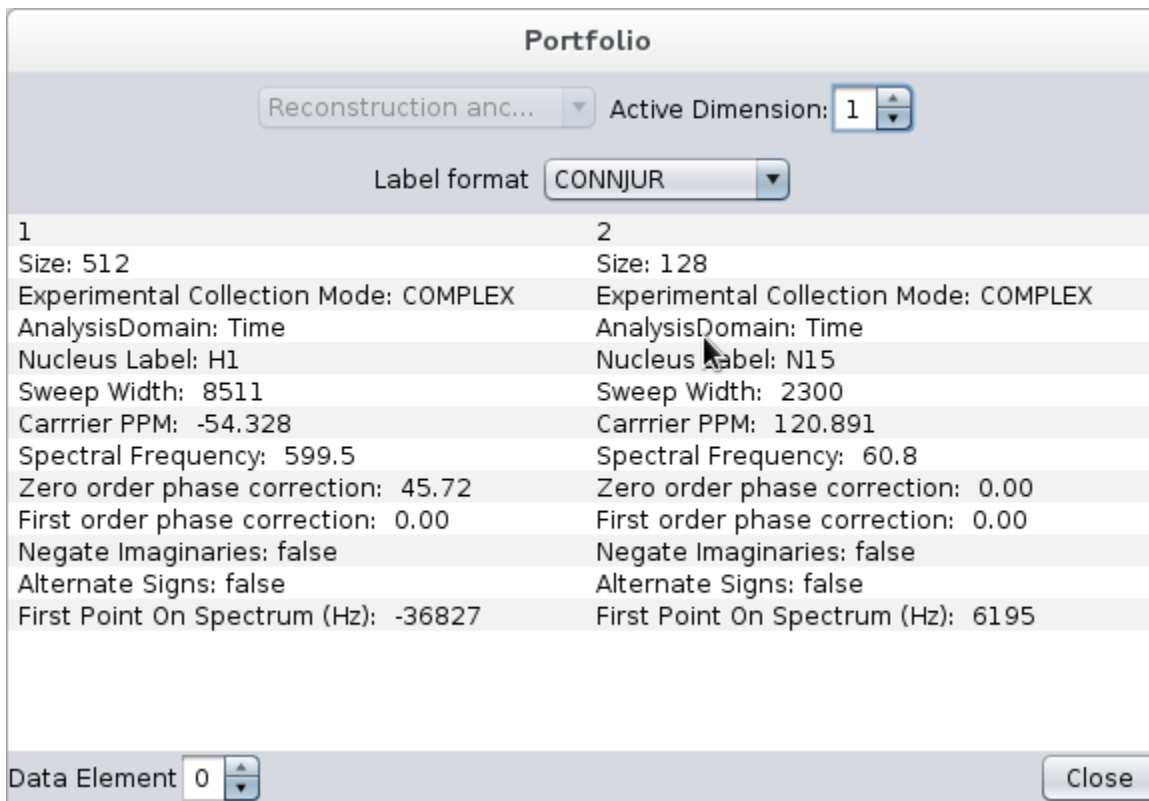
An actor which is not yet configured is colored purple. (By default. All coloring options can be customized in the Preferences menu). Once we click OK to finish configuring the actor, its color will change to pink to inform us that it is ready to be run. To run the importer, click the Run button in the top menu bar.

You may or not have noticed the Import actor flash to blue (signifying a running actor in progress) to green (signifying a completed actor). In the right hand status panel, CWB should report that the execution has completed and how long it took.



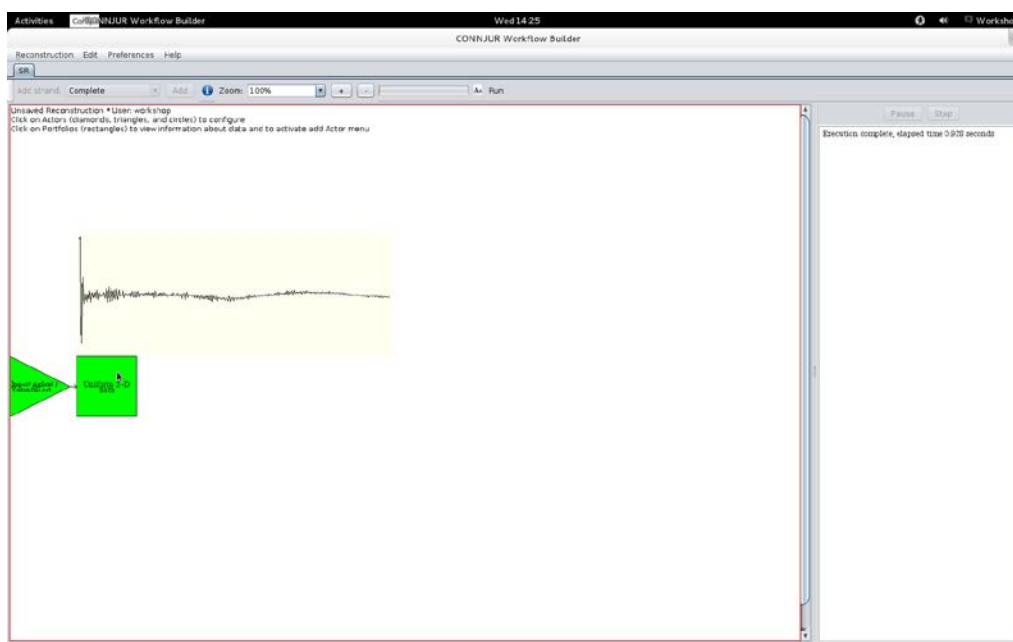
The actors are represented by diamonds, triangles, and circles. The datasets are represented by squares. So if we double-click the triangle actor, we see how it was configured. If we double-

click the square dataset, we see information (metadata) concerning this dataset and its current status. Standard metadata include the number of points (Size), Nucleus Label, Sweep Width, Carrier PPM for referencing, Phase Correction parameters (in this case as set on the spectrometer), etc. If we had assigned channel 2 to dimension 2 in the import actor, we would see information for Carbon in this window.



The nomenclature for these values can be customized using the menu option “Label Format”. Finally, the “Active Dimension” is set to 1. This signifies that the next actor added to the workflow will process dimension 1 (in this case ^1H) by default.

Mousing over a dataset will bring up a pop-illustration showing the first trace of the dataset. This can be quite instructional in getting an overview of what is happening in each processing step.



2.9 Actors

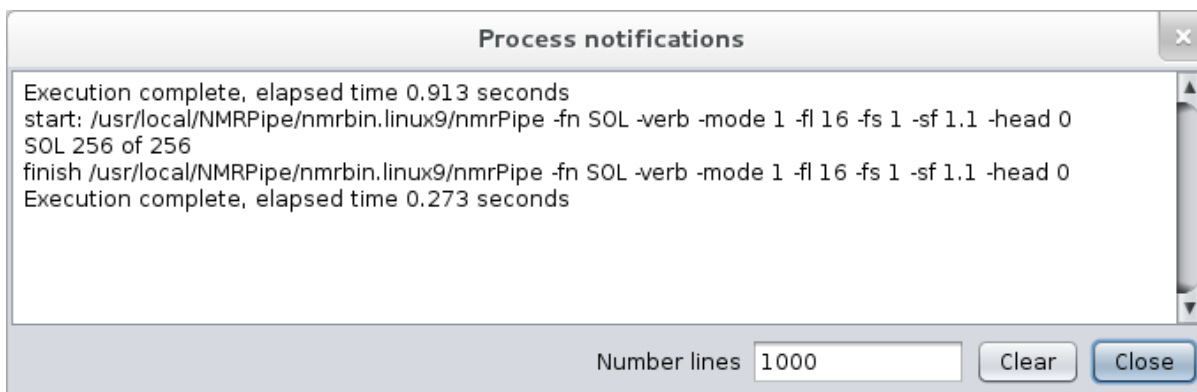
To add actors to the workflow, we must first select a dataset (square) in the workflow. We currently have only one dataset – the one imported by the import actor. Select the dataset by clicking with the mouse. In the right hand panel, an actor selection browser appears. Actors can be ordered by those most recently used, Domain groups (i.e. actors which apply to time domain vs. frequency domain), Functional groups of the processing (such as apodization), or finally the third-party tool utilized by the actor. For this session, we will concentrate on the Tool based actors, particularly, actors utilizing nmrPipe.

If we mouse-over the dataset again, we see a low-frequency roll to the data. This is emblematic of incomplete solvent suppression, as the NMR was collected in H₂O. The SOL function in nmrPipe is used to remove this artifact. We can add this process to our workflow by clicking the “Solvent Suppressed NMRPipe actor” while the dataset is selected. Do this now.

A pink diamond and square should now appear. Once again, pink signifies that the actor is configured but has not yet been run. If we double-click the “solvent suppress” actor, available configuration actors are given in the window. The defaults are typically a good choice for this actor – the important point to note is that the action is applied to dimension 1 (¹H). Click Run again to see the result of applying SOL to the first FID of the dataset (by mousing-over green squares). You should be able to notice that the low-frequency baseline roll has been removed by SOL.

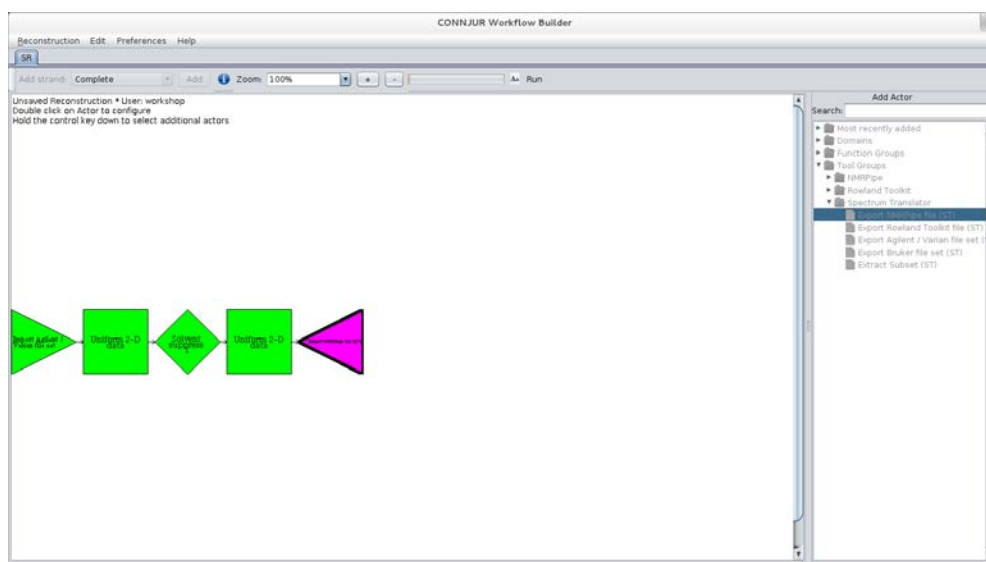
2.10 Info Button

The button on the menu bar will give a history of all the actual process run during this particular reconstruction. This is useful to examine in case an unseen error happens during processing. It also gives the user a log of the actual commands sent to nmrPipe.

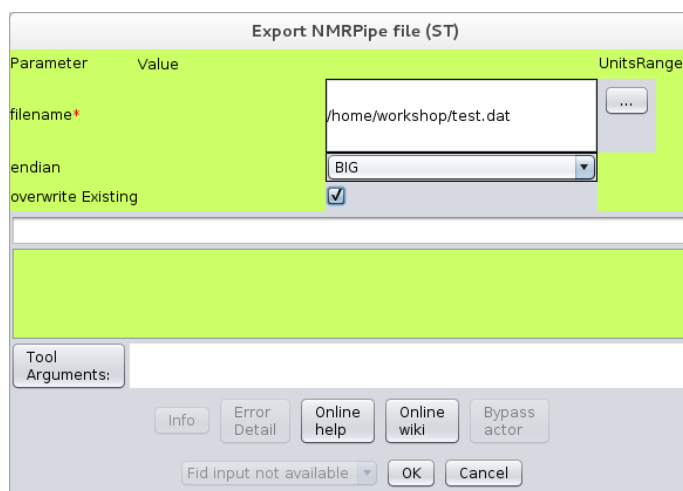


2.11 Exporters

While running workflow builder, all processed data exists either in CONNJUR memory or within the CONNJUR database (depending on configuration settings to be discussed later). However, at some point you will want to export your processed data from the CONNJUR system to be used with some other NMR analysis software. To export data from the system you need to add an Exporter. Select one of the two datasets in the workflow. Once again the actor panel should appear on the right hand side of the application. Exporters are components of the CONNJUR Spectrum Translator (to be discussed tomorrow). For the time being, suffice to say, simply chose a format for your output dataset (let's choose nmrPipe).



Once again the actor is purple, as it is waiting for a filepath and filename for our exported data. Choose something in your home path, like /home/workshop/test.dat.

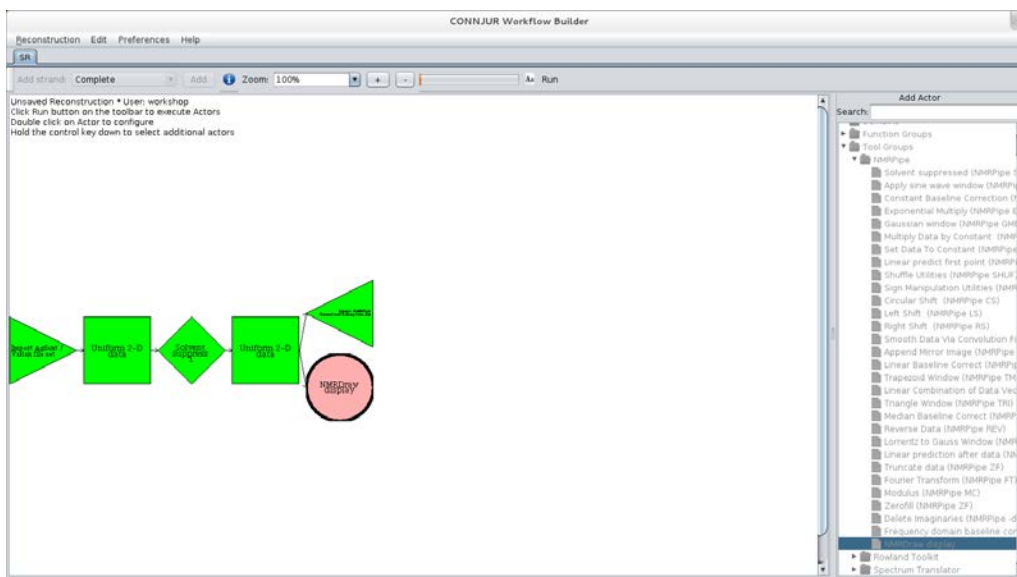


Now click run. The actor should turn green.

If we now move from CWB to a terminal, we can run NMRDraw and view our exported dataset. Simply type the command nmrDraw and select test.dat from the file menu.

2.12 NMRDraw Display Actor

We can also view our datasets using nmrDraw directly within CWB. Simply select a dataset (square), and select “NMRDraw display” actor from the right-hand panel. The display actor is visualized as a circle, pink because it is not configured. The major configuration choice is to either select it to display when run – useful when put at the end of a reconstruction workflow – or display now. Click display now.



2.13 Saving Reconstructions

Having used an exporter, our final dataset is saved to disc. However, the actual reconstruction workflow has not yet been saved. If we exit CWB, all record of how the data was processed will be lost. It is a good habit to save reconstructions to the CONNJUR database. This is done using file menu Reconstruction->Save. This brings up a pop-up window with the current username, a name for the reconstruction to be saved under, and you can arbitrary notes and keywords. The reconstruction window on the top left is now modified to reflect the new name for the reconstruction.

Reconstruction

ID:

User:

Entity:

Date created: Jun 14, 2012

Date last used: Jun 14, 2012

Notes

This is just a demonstration of how to save reconstructions

Keywords

Workshop, demo, SOL

2.14 Reconstructions vs Workflows

As you should have discovered by now, the major difference between reconstructions and workflows is that reconstructions are tied to a particular dataset. They describe and control the processing of a single dataset/portfolio. Workflows are generic – they refer to a series of processing steps which can be applied to any dataset – even one which is downstream in a reconstruction. In that sense, workflows can be considered modular building blocks for creating reconstructions.

2.15 Saving Workflows

Workflows can be saved much the same way as Reconstructions. Here we are given the opportunity to put verbose descriptions on each dimension. This is useful in case we wish to apply a workflow from one reconstruction on a dataset in which dimensions have been transposed. For example, perhaps our HNCO has dim1: ^1H , dim 2: ^{13}C and dim3: ^{15}N ; but our HN(CA)CO as dim2: ^{15}N and dim3: ^{13}C . We would like to apply the same workflow but wish to associate the ^{15}N processing steps to the correct dimension.

2.16 Retrieving Workflows

Retrieving Workflows is similar to retrieving reconstructions. Simply select a dataset and chose Reconstruction->Apply Workflow. A selector pop-up will allow you to choose the appropriate workflow, assign dimensions, and add to the reconstruction window.

2.17 Exporting/Importing Workflows

Saving and Applying workflows stores/retrieves them from your internal CONNJUR database. However, if you wish to move workflows between databases, for instance to share workflows with other collaborators, you wish to choose Export and Import. In these cases, the workflow is stored as an XML representation which can be transmitted between different machines and CONNJUR installations. Try Exporting our workflow. You will need to choose a filename. If you do a 'more' on your xml file, you will see the representation is in plain text and contains all of the configuration information for your workflow. These files should not be edited unless you really know what you are doing.

Part III: HSQC Example

In this example, we will process a standard, 2-dimensional ^1H - ^{15}N HSQC spectrum. Since this spectrum was collected in H_2O (not D_2O), we will wish to suppress the residual solvent signal in the direct dimension. Also, as we collected only a limited number of points in the indirect dimension, we will use LP to enhance our resolution in the ^{15}N dimension. Otherwise, the two processing dimensions will have similar spectral processing actors – apodization, zero-filling, fourier transform, phasing and deletion of the imaginary components.

We have already loaded our Varian/Agilent HSQC spectrum. Select any additional actors we've added and delete them either using the delete key on the keyboard or Edit->Delete. Select the dataset and import the HSQC workflow (HSQC_tutorial.xml). As you should be aware with this dataset, dimension 1 is ^1H and dimension 2 is ^{15}N .

You will note that there are three actors which appear white, signifying they are bypassed. These are the “Linear Prediction” and the two “Phase Corrections”. On first pass, we do not necessarily know the proper phase parameters. Linear prediction is typically time-consuming and is left out until the phase parameters have been calculated.

Let's begin by running the workflow with these three actors bypassed. All of the non-bypassed actors/datasets should turn green. As usual, the right-hand panel (or the information panel) report which commands were sent to nmrPipe and the status. Double click the NMRDraw Display actor and click “Display Now”. We can now use NMRDraw to determine the proper phase for the ^1H dimension. Record the value, go back to the workflow, double-click the “Phase Correction 1” actor and set these phases. Click Run again. Open the spectrum in the NMRDraw Display actor, and the spectrum should be properly phased (which means that no phasing in the ^{15}N dimension is necessary and that phase actor can remain bypassed).

Upon inspection of the spectrum, you will notice that while the ^{15}N dimension is referenced correctly, the ^1H dimension is not. This is a problem with the metadata. Let's take a look at the metadata by double-clicking the first dataset in the workflow (the one attached to the importer).

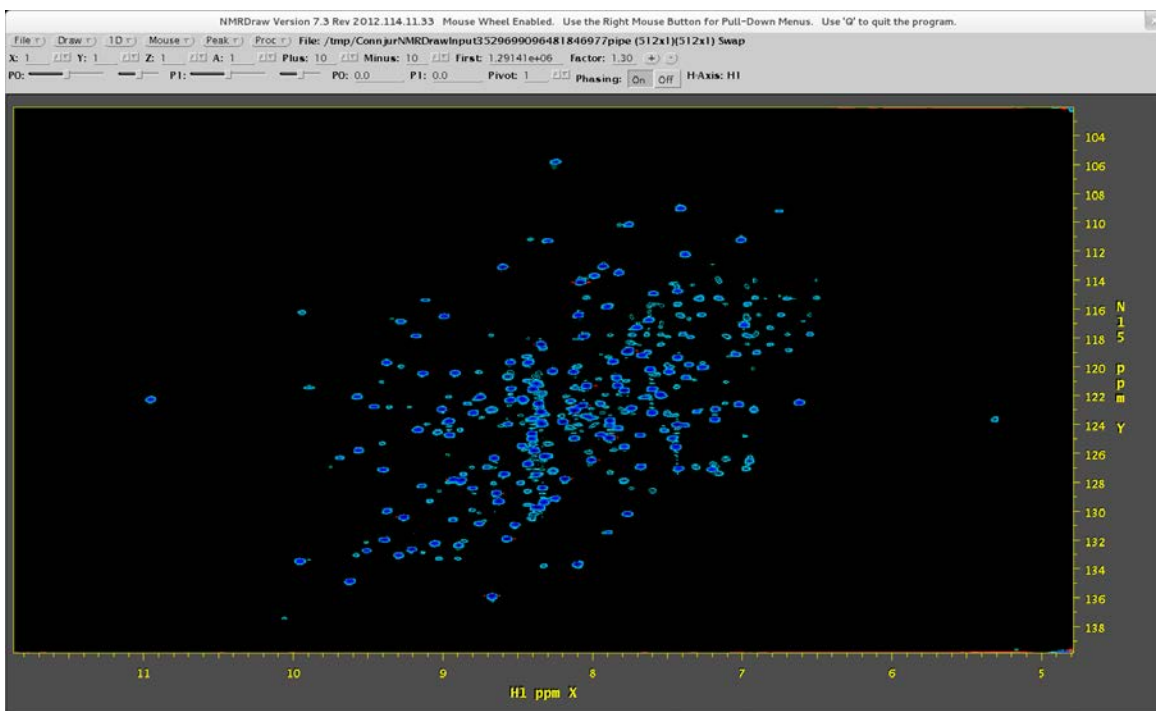
Portfolio	
Reconstruction anc... Active Dimension: 1	
Label format CONNJUR	
1	2
Size: 512	Size: 128
Experimental Collection Mode: COMPLEX	Experimental Collection Mode: COMPLEX
AnalysisDomain: Time	AnalysisDomain: Time
Nucleus Label: H1	Nucleus Label: N15
Sweep Width: 8511	Sweep Width: 2300
Carrier PPM: -54.328	Carrier PPM: 120.891
Spectral Frequency: 599.5	Spectral Frequency: 60.8
Zero order phase correction: 45.72	Zero order phase correction: 0.00
First order phase correction: 0.00	First order phase correction: 0.00
Negate Imaginaries: false	Negate Imaginaries: false
Alternate Signs: false	Alternate Signs: false
First Point On Spectrum (Hz): -36827	First Point On Spectrum (Hz): 6195
Data Element 0	Close

You will notice that the carrier ppm recorded for the 1H dimension is -54.328. This should be set to 4.772. To correct this, we need to add adjust the metadata. Delete the workflow by selecting the first SOL actor and hitting delete. Now add "Adjust Metadata" from the Spectrum Translator Tool Group. Double click the "Adjust Metadata" actor and set the carrier PPM to 4.772. Now reload the NHSQC.xml workflow, reset the phase parameters for the first phase correction, and hit run. If you open up the spectrum in NMRDraw Display – the spectrum should be referenced correctly.

We can now restore the Linear Predict actor and configure it to linear predict an additional 128 points in the ¹⁵N dimension.

The final spectrum is shown below:

Connecticut NMR Workshop 2012



Part IV: HNCACB Example

In creating our NHSQC workflow, we have built one of the building blocks for processing any NHSQC-based experiment, for instance the HNCACB. Let's move on to this 3D experiment. Save any current work you have done and create a new reconstruction (Reconstruction->New), once again selecting Agilent/Varian format. Double-click the importer and select the hncacb.fid directory for import. The channel-dimension associations should all be as default (1=1, 2=2, 3=3) and sensitivity enhancement should be selected for dimension 3.

Parameter	Value	Units	Range
sensitivity Enhancement On Dimension	3	(Rance Kay)	
negate Imaginaries	<input checked="" type="checkbox"/>		
dimension One Channel	1		
dimension Two Channel	2		
dimension Three Channel	3		
dimension Four Channel	not present		
directory*	/home/workshop/CONNJUR_data/hncacb.fid		
ignore Extra Data	<input type="checkbox"/>		

Tool Arguments:

Info Error Detail Online help Online wiki Bypass actor

Fid input not available OK Cancel

Now let's import our NHSQC.xml workflow. The ^{15}N dimension in this experiment is dimension 3, so we want to assign dimension 3 to N15 and leave dimension 2 unassigned.

Assign Workflow dimensions by clicking on first row in table.

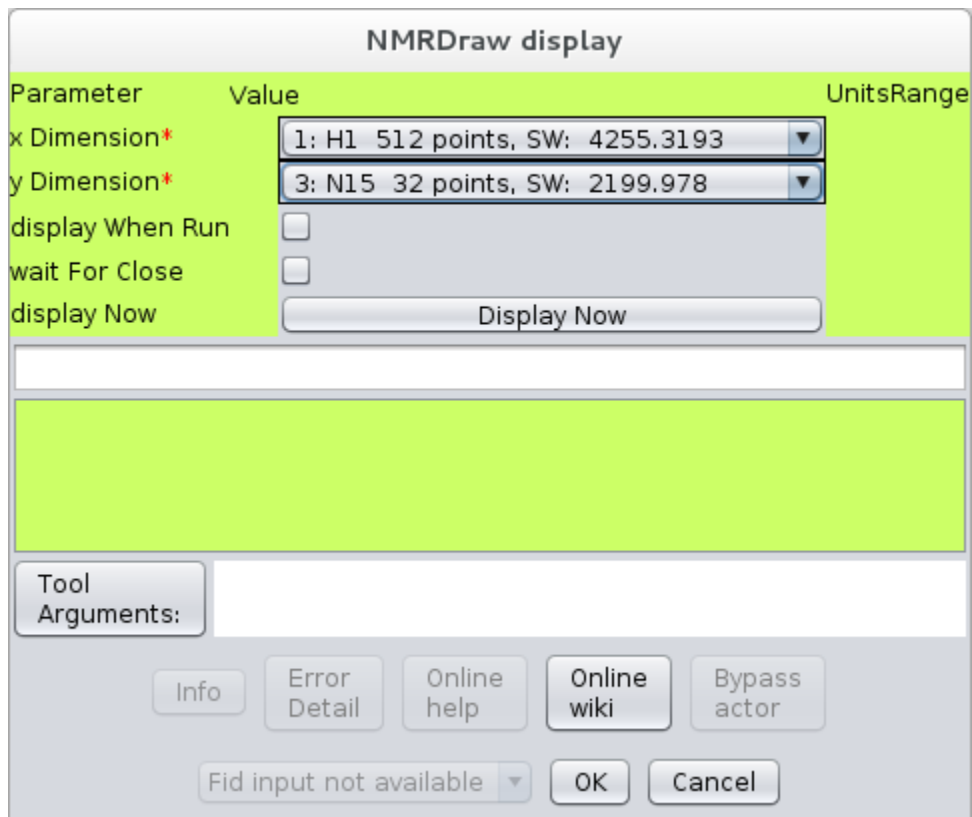
Label format: CONNJUR

H1 599.53973 Hz, with solvent	unassigned	N15 60.75768 Hz, linear predicted
1	2	3
Size: 1024	Size: 72	Size: 32
Experimental Collection Mode: COMPLEX	Experimental Collection Mode: COMPLEX	Experimental Collection Mode: COMPLEX
AnalysisDomain: Time	AnalysisDomain: Time	AnalysisDomain: Time
Nucleus Label: H1	Nucleus Label: C13	Nucleus Label: N15
Sweep Width: 8511	Sweep Width: 10600	Sweep Width: 2200
Carrier PPM: -27.163	Carrier PPM: 51.046	Carrier PPM: 120.065
Spectral Frequency: 599.6	Spectral Frequency: 150.8	Spectral Frequency: 60.8
Zero order phase correction: 62.09	Zero order phase correction: 0.00	Zero order phase correction: 0.00
First order phase correction: 0.00	First order phase correction: 0.00	First order phase correction: 0.00
Negate Imaginaries: false	Negate Imaginaries: false	Negate Imaginaries: false
Alternate Signs: false	Alternate Signs: false	Alternate Signs: false
First Point On Spectrum (Hz): -20541	First Point On Spectrum (Hz): 2396	First Point On Spectrum (Hz): 6195

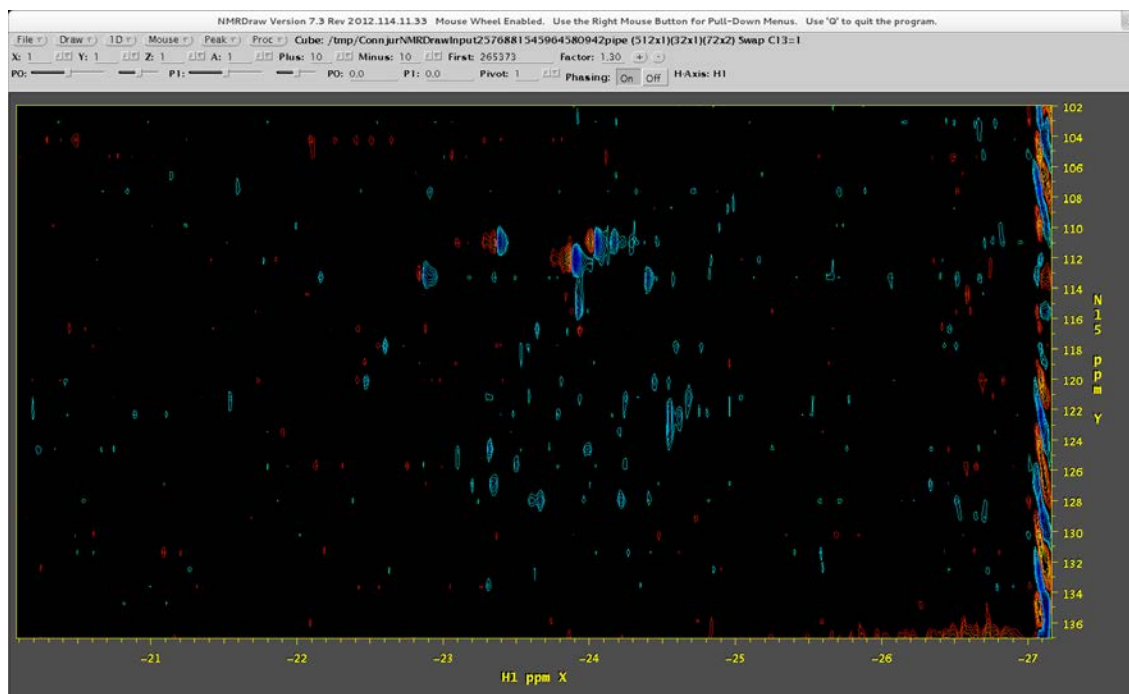
Cancel Apply Workflow

As before, our linear prediction and phase correction actors are bypassed. To speed things up a bit, let's also bypass the zerofill actors. Now run the workflow. This 3D dataset is much larger than the 2D, so the workflow takes longer to run. When finished, click the NMRDraw Display

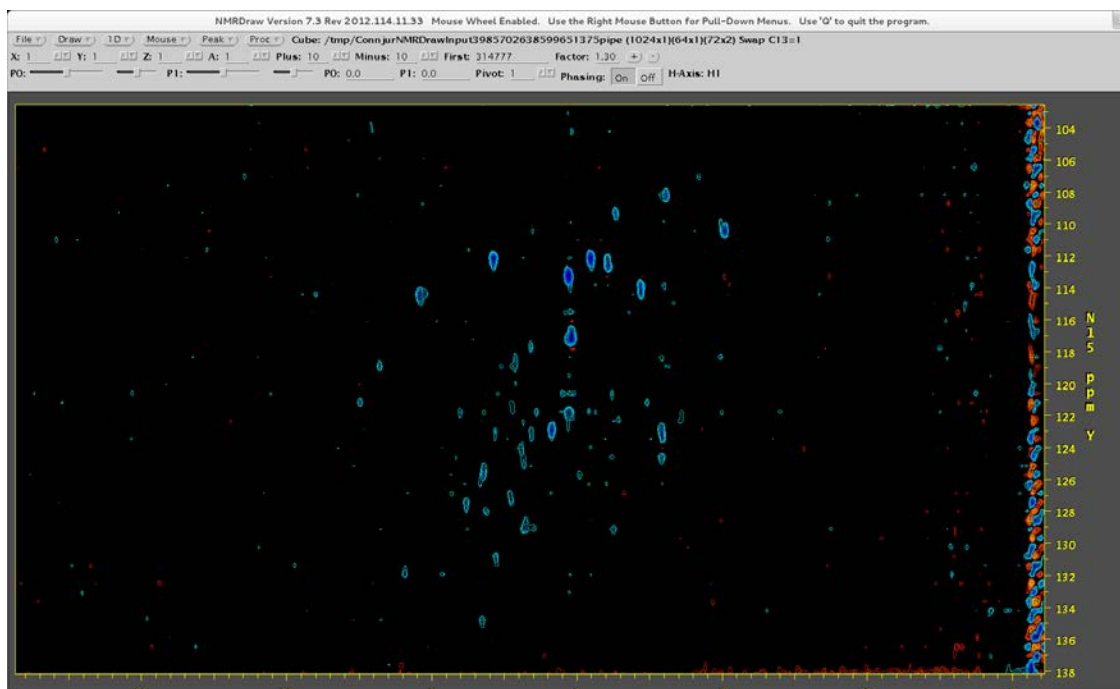
actor. Note that by default, the x and y dimensions of the NMRDraw Display are set to Dimensions 1 and 2 of the spectrum, in our case H1 and C13. Our workflow processed dimensions 1 and 3, so select dimension 3 for the y-axis of the display. Next click "Display Now".



You should see the first ^1H - ^{15}N plane of the 3D. You should note that the spectrum is improperly phased and referenced in the ^1H dimension. We know how to fix both. Determine the phase correction and close the NMRDraw Display.



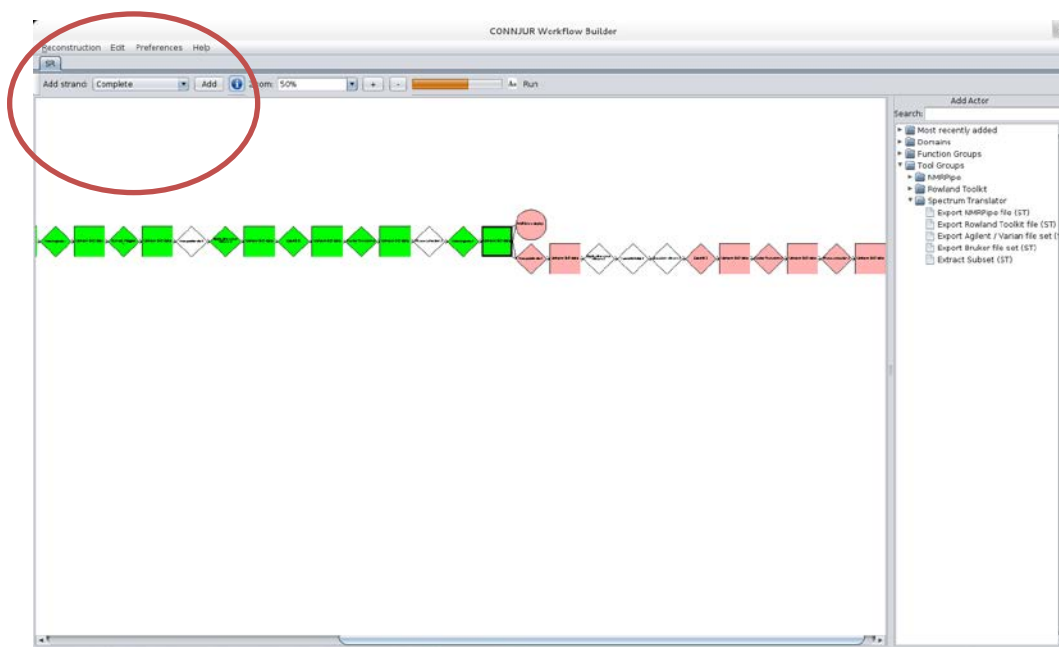
Now do the following, delete the current workflow starting with the SOL actor. Add an adjust metadata actor after the importer. Configure the metadata actor to set the ^1H carrier ppm to 4.772. Re-import our HSQC workflow. Restore the 1st phase correction actor and set the proper phase parameters (Hint: they are the same as in the HSQC example). Now re-run the workflow. (If you left the zerofill actors in place, it will take longer to run than last time.) The first ^1H - ^{15}N plane should look something like the figure below.



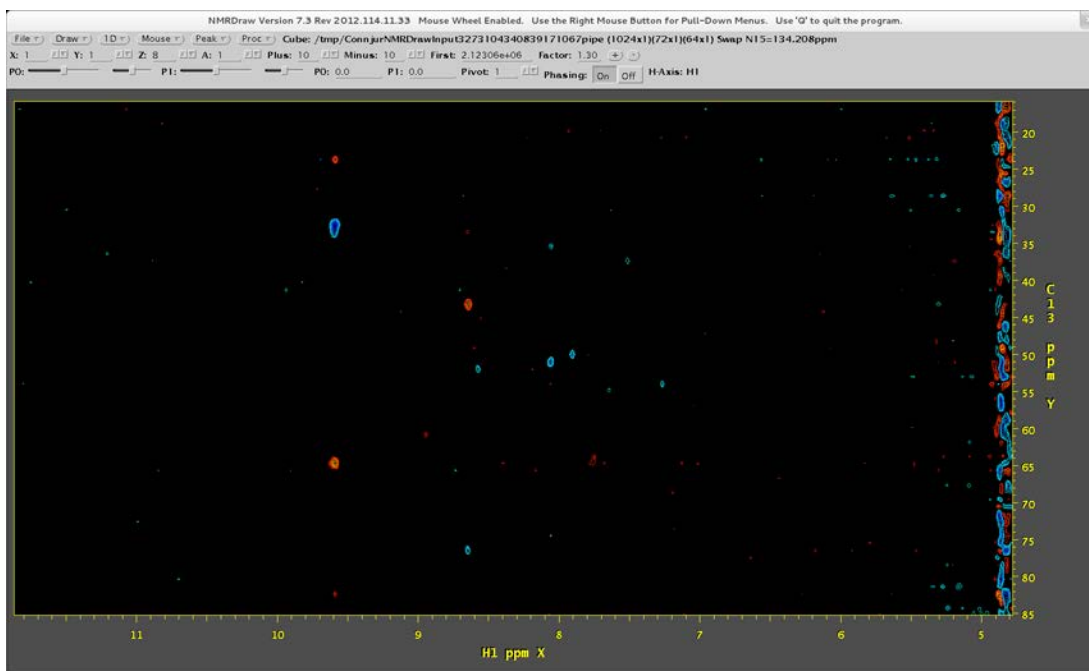
4.1 Strands

As you may have noticed, the ability to bypass actors gives you the ability to preconfigure multi-functional workflows and switch on/off the functionality at a whim. This concept can be extended to smaller processing modules, which perform a series of processing steps on a single dimension. We call these sub-workflows strands.

We currently have processed dimensions 1 and 3 of our 3D HNCACB. We need to process dimension 2, C13. Select the last dataset (the one prior to the NMRDraw Display actor). In the upper left corner of the application you will see a pull down menu for “Add Strand”. Select “Linear Predict” and click the “Add” button. The workflow will have forked in two and added a new series of actors for “linear prediction”, a series of apodization actors, zerofill, ft and phase correction.



Let's bypass the Linear Predict and Zerofill actors. We also need to add “delete imaginaries” and “NMRDraw Display” actors at the end. Click Run. When finished, click the NMRDraw Display actor and view the ^1H - ^{13}C planes. Scroll through the z-planes using NMRDraw. The ^{15}N plane at 134 PPM should seem strange. In the HNCACB experiment, alpha carbons and beta carbons appear with opposite signs. The residue at 9.6 ppm shows the alpha carbons to be at around 33 ppm which indicates that the ^{13}C dimension needs to be reversed. This is accomplished by negating the imaginaries of that dimension. This can be done in either of two ways – we can configure the FT actor to negate the imaginaries, or we can adjust the metadata again.



Configure the Adjust Metadata actor to “negate Imaginaries” and “correct Negate Imaginaries” along the C13 dimension.

Adjust MetaData (NMRPipe syntax)

Parameter	Value	Units	Range
correct Domain	<input type="checkbox"/>		
domain	Time		
correct Sweep Width	<input type="checkbox"/>		
sweep Width*	10600.0	Hz	<input type="checkbox"/>
correct Spectral Frequency	<input type="checkbox"/>		
spectral Frequency*	150.76414		<input type="checkbox"/>
correct Name	<input type="checkbox"/>		
name	C13		<input type="checkbox"/>
correct PPM	<input type="checkbox"/>		
carrier PPM*	51.045967	carrier	<input type="checkbox"/>
correct Zero Phase	<input type="checkbox"/>		
zero Order Phase*	0.0		<input type="checkbox"/>
correct First Phase	<input type="checkbox"/>		
first Order Phase*	0.0		<input type="checkbox"/>
correct First Scale	<input type="checkbox"/>		
first Scale Point*	2395.9019	Hz	<input type="checkbox"/>
correct Sign Alteration	<input type="checkbox"/>		
sign Alteration			
correct Negate Imaginaries	<input checked="" type="checkbox"/>		
negate Imaginaries	<input checked="" type="checkbox"/>		
dimension	2: C13 72 points, SW: 10600.0		

Tool Arguments:

Info Error Detail Online help Online wiki Bypass actor

Fid input not available OK Cancel

When we re-run the workflow, we will see that our ¹³C axis is now correct.