NMR Analysis Tools for RNA Structural Biology Integrated into NMRFX Analyst and RING NMR Dynamics

Ellen Koag and Bruce Johnson

Structural Biology Initiative, CUNY Advanced Science Research Center, New York, NY

Abstract

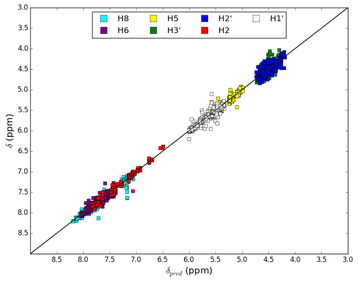
NMRFx Analyst[1] is an integrated software application for the processing and analysis of NMR data. RING NMR Dynamics[2] is an application for the analysis of NMR data that probes macromolecular dynamics. Both applications, written in the Java and Python programming languages, are cross-platform applications and run on Linux, Mac OS X and Windows operating system. NMRFx can incorporate RING as a plugin allowing facile transition from raw NMR data to parameters describing molecular motions. We describe here the current state of specific features added to NMRFx and RING to allow for the analysis of NMR data from **RNA** molecules.

These features include chemical shift prediction of RNA molecules based on sequence attributes and 3D structural information and NMR assignments of RNA molecules using our molecular network approach. Within RING, we provide capabilities for measuring chemical exchange between conformational states using CEST and R_{1rho} experiments. Structural calculations are based on torsion angle molecular dynamics. We've recently optimized these to allow subdividing larger RNAs into fragments that are connected with automatically inserted linkers. These allow rapid assembly of predefined fragments such as helices, turns and bulges.

We will also describe our recent work on incorporation of deep learning techniques that will allow direct access to state-of-the-art methods directly integrated with the NMR analysis tools. As one example, we'll show a deep learning language model that can directly translate RNA sequences into dot-bracket (Vienna) notation that summarizes secondary structures.

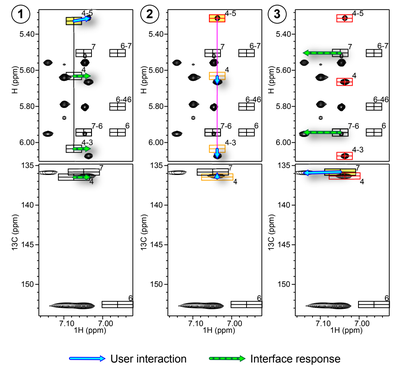
Chemical Shift Prediction

NMRFx Analyst includes chemical shift predictions based on secondary structure attributes and 3D structural information. Secondary structure attributes include a 5 residue window consisting of nucleotide type and base-pairing[3]. Predictions are available for H, C and N atoms and are of high quality as shown in the plot of predicted and experimental H chemical shifts. The predictions are integrated into NMRFx Analyst and are performed together with predictions for proteins and small molecules so a chemical shift predictions can be performed for protein-RNA complexes with bound small molecules.



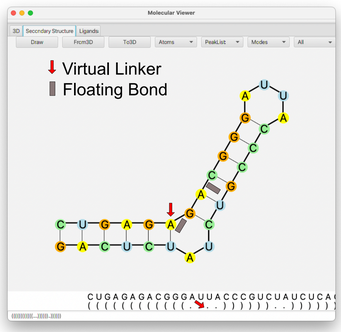
Chemical Shift Assignment

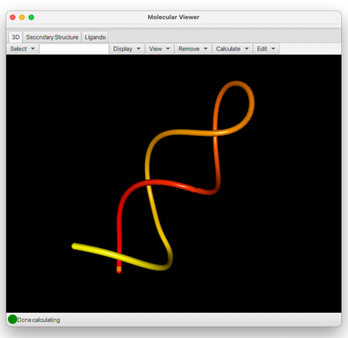
NMRFx Analyst includes tools for facilitated assignment of RNA chemical shift. We use an inverted approach that focuses on networks of coupled peaks that are predicted from the RNA secondary structure and type of NMR experiment[3]. Instead of picking peak positions and then attempting to assign them, we generate a linked network of assigned peak-boxes at these predicted positions that can then be interactively aligned with the observed spectra. This approach allows the spectroscopist to make simultaneous use of multiple spectral features that can minimize ambiguity in the assignments compared to the process of assigning individual peaks predictions. We are also working to provide tools that automate portions of this facilitated assignment process.



Structure Calculation with Automated Fragments

A standard method for NMR based structure calculation is to use torsion-angle dynamics programs such as CYANA and XPLOR-NIH. In torsion angle space the calculation uses a continuous tree of bonds from the 5' to 3' end. With larger RNAs the NMR analysis is often done on smaller fragments which are assembled to form the complete structure. NMRFx has torsion angle dynamics built-in and we now facilitate the structure calculation by automatically inserting virtual linkers and openings in the tree structure. This allows assembling fragments, which might have predefined structures. In the example below, linkers and tree openings are automatically inserted so the structure consists of an initial helix, a bulge, and a final stem loop. The 3D structure on the right is generated from pre-defined angles for these fragments.





Integrated Deep Learning Tools - Secondary Structure Prediction

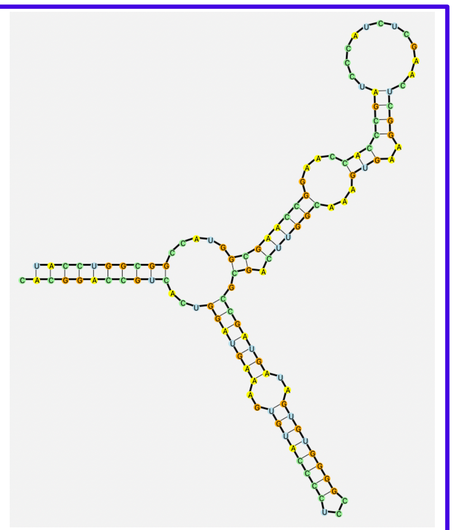
We are integrating deep learning technologies directly into NMRFx Analyst. One use is to derive structural information from the RNA sequence. In our first approach we have trained a model for secondary structure prediction.
We treat the prediction as a direct translation from the sequence to the dot-bracket (Vienna) code.
Initial training is with a masked language model as previously done for proteins[5] and RNA.[6] Training is in Keras/ Tensorflow with a combination of FNetEncoder (Fourier Transform Encoder) and Transformer layers.
Approximately 20 million unlabeled RNA sequences are used to train the masked language model.
Approximately 20 thousand labeled (with secondary structure) RNA sequences are used to fine-tune the model for secondary structure prediction. We augment the training set with new sequences with conservative changes.
Input to the fine-tuning state is the RNA sequence and the model is trained to translate the sequence characters into secondary structure characters, where:

- A: first base of a base paired segment
- B: middle bases of a base paired segment
- C: last base of a base paired segment
- a,b,c : base paired complements of above

We have not trained yet on pseudo-knots, but this will be done with additional characters (D,E,F,f,e,d etc.).
The current, preliminary model, has precision and recall scores on a test set of 92% and 91%, respectively

```
input    : UACCUGGCGGCCAUGGCGAACCGGAACCACCCGAUCCCAUCUCGAACUCGGAAGUGAAACGGUUCAGCGCCGAUGAUAGUGUGGGGCCUCCCCAUGUGAAAGUAGGUCACUGCCAGGCAC
actual   : ABBBBBBBBC.....ABBBBBBC....ABBBBBC............cbba..cba...cbbbba.ca.ABBBBBC..ABBBBBC...cbbbbbba..cbbbbba...cbbbbbbbba.
predicted: ABBBBBBBBC.....ABBBBBBC....ABBBBBC............cbba..cba...cbbbba.ca.ABBBBBC..ABBBBBC...cbbbbbba..cbbbbba...cbbbbbbbba.
actual   : ((((((((((.....(((((((((....(((((((.............)))))..)))...)))))))..)).(((((((..(((((((((...)))))))))..)))))))...)))))))))).
predicted: ((((((((((.....(((((((((....(((((((.............)))))..)))...)))))))..)).(((((((..(((((((((...)))))))))..)))))))...)))))))))).
```

- The above shows an example of an input sequence, our base-pair character representation, the dot-bracket representation and the model predictions. The secondary structure is then represented (with the secondary structure viewer that is built in to NMRFx Analyst.
- As noted above, the model has high precision and recall in predicting the secondary structure. However, the output is not always balanced (equal left and right parentheses) so we are working on a final post-processing step.

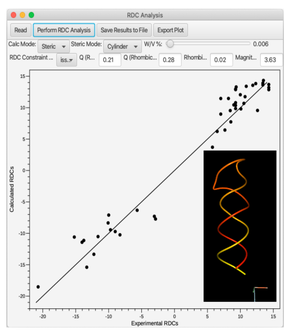


RNA Dynamics

Biological function is dependent on both structure and dynamics. NMR spectroscopy is one of the most powerful methods for probing structural dynamics and can be used to probe motions over multiple (12 - orders) timescales. Here we show the use of RING NMR Dynamics (embedded in NMRFx) to analyze CEST experiments on an RNA molecule to probe millisecond dynamics.

Residual Dipolar Couplings

RNA structure is well defined locally by NMR distance restraints but the orientation of helices can be under-constrained. Residual dipolar couplings can provide angular information that can define these orientations. NMRFx includes the ability to calculate the alignment of RNA molecules in oriented media and thereby predict the RDC values. Comparison with experimental values allows filtering trial structures for consistency between experimental and predicted values. Shown in the figure are predicted and experimental values of the HIV-1 intron splicing silencer (PDB ID 2N4L from the Tolbert lab).



Development of NMRFx Analyst and RING NMR Dynamics

Includes support from NIH grants:

- Center for Structural Biology of HIV RNA (CRNA) U54 AI170660-01
- Center on Macromolecular Dynamics by NMR Spectroscopy (CoMD/NMR) 1RM1 GM145397
- NMRFX: An integrated software suite for macromolecular NMR Analysis 5R01GM123012

References

1. Norris M, Fetler B, Marchant J, Johnson BA (2016) NMRFx Processor: a cross-platform NMR data processing program. J. Biomol. NMR 65:205-216.
2. Beckwith MA, Erazo-Colon, Teddy, Johnson BA (2020) Ring NMR Dynamics: software for the analysis of multiple NMR relaxation experiments. J. Biomolecular NMR. ttps://doi.org/10.1007/s10858-020-00350-w
3. Brown JD, Summers MF, Johnson BA. (2015) Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. J Biomol NMR. 63:39-52.
4. Marchant J, Summers, MF, Johnson BA (2019) Assigning NMR Spectra of RNA, Peptides and Small Organic Molecules using molecular network visualization software. J. Biomolecular NMR. https://doi.org/10.1007/s10858-019-00271-3
5. Rives, A et al 2021 PNAS 118 (15) e2016239118)
6. Chen, J et al., 2022, arxiv.org/abs/2204.00300).