Predictive Modelling of NMR Chemical Shifts of RNAs using Machine Learning approaches

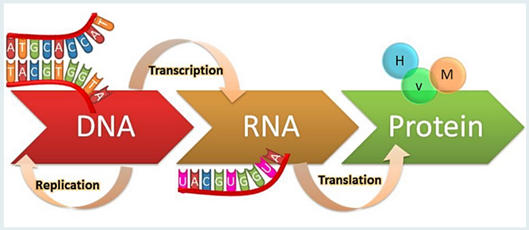
Sahil Chhabra[1,3] and Aaron T. Frank[2,3]

[1]Department of Chemistry, University of Michigan, Ann Arbor, Michigan, 48109

[2]Department of Biophysics, University of Michigan, Ann Arbor, Michigan, 48109

[3]Michigan Institute for Computational Discovery & Engineering, University of Michigan, Ann Arbor, Michigan, 48109

Introduction

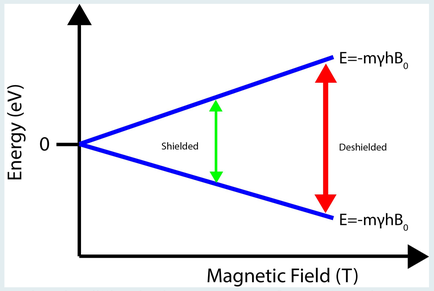


Central Dogma of Molecular Biology

In the present "**RNA World**", Ribonucleic acid (RNA) plays a vital role beyond translation of information from DNA to proteins as depicted in the central dogma of biology. In has been elucidated that various forms of non-coding RNA (ncRNA) are involved in different cellular processes from catalysis in protein synthesis to gene silencing. RNA is also associated with several diseases, so understanding the RNA structure - function is of vital importance.

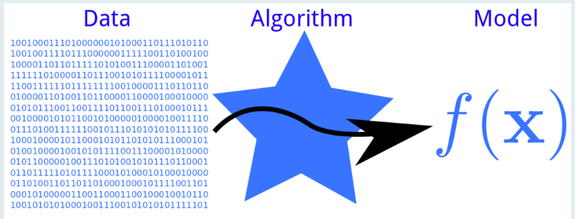
Chemical Shifts

"**Structural portraits**" of RNA functions will aid in understanding the relationship between RNA structure-dynamics-function. NMR chemical shifts (CS) which are the "**chemical fingerprints**" of any molecule are also capable of resolving the RNA 3D structure as they are readily accessible, precisely measured and are sensitive to the structure.



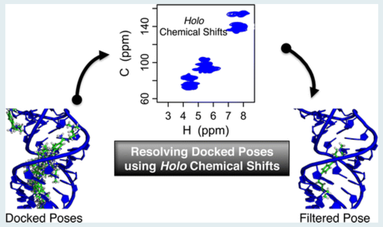
Energy level splitting in the presence of an external magnetic field

Methods



Schematic representation of Machine Learning

The electronic cloud around the nucleus is shielded from the external magnetic field, and chemical shift is the extent of representing this shielding, which is influenced by many "structural features", within the molecule. These structural features within the molecule, that influence the chemical shifts, can be hydrogen bonding, magnetic anisotropy, ring currents, stacking interactions and torsion angles. The extracted features from a given RNA coordinates are used to build predictors using the sci-kit learn python module.



Resolving Docked poses using Holo Chemical Sifts

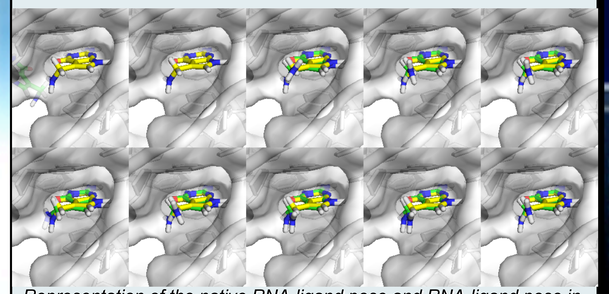
All the structural features are calculated (serves as features for the various models), and the model are trained on a training set of 19 RNAs The model is then tested on a set of 36 testing RNAs. Various machine-learning algorithms are tested using linear to Ensemble methods, and the accuracy of each model is assessed.

Results

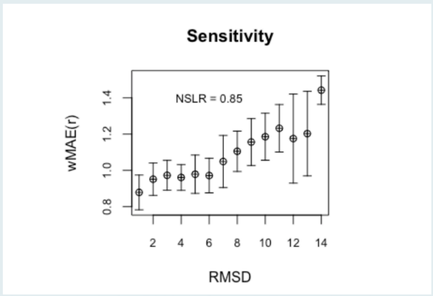
Linear Methods	MAE(^{13}C/^{1}H)
Linear Regression	0.893/0.159
Ridge Regression	0.878/0.172
RidgeCV	0.863/0.167
Theil-Sen	0.844/0.176
Bayesian Ridge	0.866/0.169
LassoLarsCV	0.819/0.153
Elastic Net	0.906/0.184
LassoLarsIC	0.856/0.156

Ensemble Methods	MAE(^{13}C/^{1}H)
Random Forest	0.757/0.145
Extra Randomized	0.748/0.140
Gradient Boosting	0.854/0.150
AdaBoost	0.844/0.155
Bagging	0.756/0.144

Mean Absolute Errors (MAE) for different Machine Learning Methods



Representation of the native RNA-ligand pose and RNA-ligand pose in the decoy pool that exhibited lowest MAE



Sensitivity analysis for RNA 1LC6 with ET machine learning method

Summary

Extra-Randomized trees (ET) and Lasso Lars CV (LLCV) methods were the most accurate ones in the ensemble and linear methods respectively. For Carbon "LLCV" method is the most sensitive whereas for protons "ET" is the most sensitive. The weighted MAE between predicted and measured chemical shifts increases as the structure becomes more non-native in a pool of decoy structures. (Sensitivity)

Discussion

Determining the structure of the RNA in the presence of drug-like molecules is a crucial step in any drug development campaign. As such, there is a need for the development of fast, easy, and "**precise prediction**" methods for determining the 3D structure of RNAs. Chemical shifts have become an important tool for the Structure-Activity Relationship (SAR). Therefore our research is focused on developing models, using this structural information, with better accuracy, that, in turn, could elucidate how the key structural features of RNA systems are associated with its function, how systems interacts, what they bind to and design drug agents that exploit these structural features to solve medical problems.

Reference

[1] A. T. Frank, "Can Holo NMR chemical shifts be directly used to resolve rna- ligand poses?" J Chem Inf Model, vol. 56, pp. 368–376, 2016.

[2] S. M. L. Aaron T. Frank and C. L. BrooksIII, "A simple and fast approach for predicting 1H and 13C chemical shifts: Toward chemical shift-guided simulations of rna," J. Phys. Chem. B, vol. 118 (42), pp. 12 168–12 175, 2014.