Predicting Molecular Structures using SMILES strings and Bayesian Networks

Student: Jake Rivett - Supervisor: Professor Simon Dobson

Background

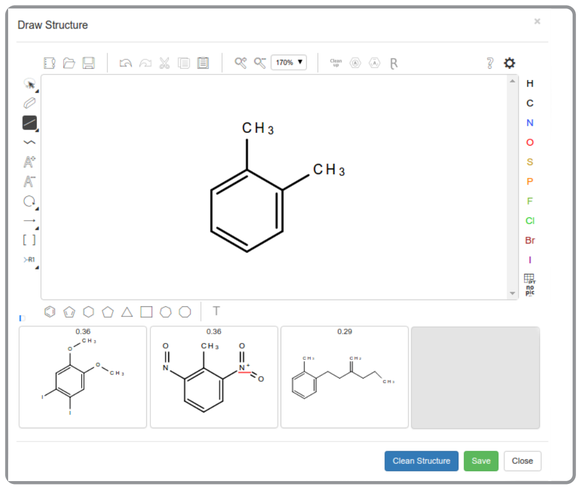
The School of Chemistry in St Andrews back in 2012 got funding to begin to develop an automatic data management system for storing their Nuclear Magnetic Resonance(NMR) spectroscopy data. The software solution was called online control system named NMR Online Management And Datastore (NOMAD) was created through a collaboration of the School of Computer Science and Chemistry in St Andrews.
Version 2.0 of NOMAD, that went into beta testing this year, introduced the ability to search for experimental data by molecular structure formulae. The problem was the process could be slow and error prone for many users and offers little assistance towards the process of drawing.

Project Aim

The project's aims were:

1. To explore the possibility of a tool that could wrap around a molecular structure drawing library, and give predictions about what the user might be trying to create.
2. To attempted to speed up the drawing process of a molecular structure for a given user, therefore directly increasing the speed and accuracy of searching for experimental data in NOMAD by using the structure drawn.
3. Create a tool that could learn from and cater for users using the system, improving predictions as the system is used.

Example of structure prediction tool



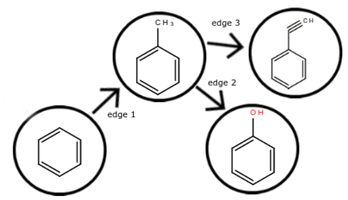
References

[1] Turner, H. Introduction to generalized linear models. http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf, 2017. Last accessed 5 April 2017.

Method

The system created uses a Naive Bayesian network, built from statistical data from previously drawn structures contained in a graph of users actions.

High level representation of database



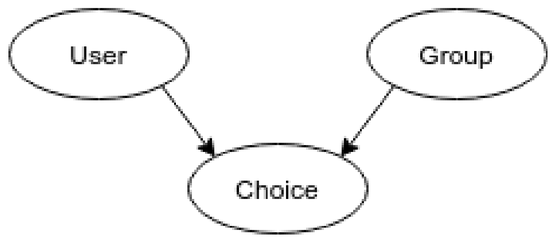
Each edge within the graph has statistical data stored about how many times each user within the system has drawn that edge in the past and the research group the user was within when he drew it.

Data associated with each edge

Edge 1			Edge 2			Edge 3		
User	Group	Times	User	Group	Times	User	Group	Times
User 1	Group 1	1	User 1	Group 1	1	User 2	Group 1	1
User 2	Group 1	1						

The application makes predictions by looking at a users currently drawn structure and quires the database for out going edges for that structure. It then uses that data to built the naive Bayesian network and return the possible edges in order of their calculated probability.

Example of Bayesian network

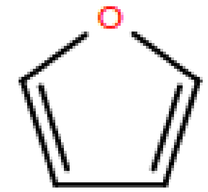


User Prediction View

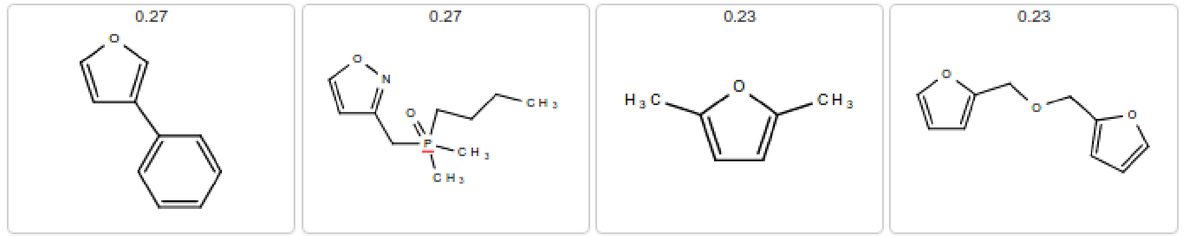
This section attempts to show the prediction from a user perspective.

- When a user begins to draw a molecular structure within the application, a list of predictions will appear at the bottom of the drawing application
- The structure on the right gives an example of something a user might draw
- The image bellow shows the predictions that application returns given the example structure on the right
- The user can click on any of the images to change the editor to that structure.

Example of input from a user when drawing a structure



Example of predictions shown to a user that has drawn the above structure



Evaluation

To evaluate the system designed, chemistry students were asked to draw structures with and without the prediction tool. 25 students completed the data acquisition phase; 14 of that group completed the evaluation part.

Summary of GLM variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.324e-02	2.699e-03	19.726	< 2e-16 ***
Predictions used	5.070e-03	1.218e-03	4.164	3.88e-05 ***
Smiles Length	-5.463e-04	5.116e-05	-10.678	< 2e-16 ***
Rubs	-3.031e-03	6.643e-04	-4.562	6.86e-06 ***
Undos	-1.381e-03	2.688e-04	-5.138	4.45e-07 ***

Time Taken To Draw Structure

The data gathered was then inputted in a Gamma Generalized Linear Model (GLM) to observe if there was any significant difference in the time taken for users to draw a structure when they used prediction compared to when they did not [1]. The GLM runs a t-test for each variable with and without a variable to give a significant value for each variable. The value for predictions used means the number of predictions a user clicks directly impacts the time taken.

Percentage of Structures Correctly Drawn

Another area explored was whether there was any significance difference in the number of errors made by each user when they used the prediction and when they did not. Comparing the groups, when users used prediction the structure they drew was $\approx 96\%$ of the time correct and when they did not use prediction users were correct $\approx 80\%$.

Future Work

Some of the future work that could be done to improve the application in the future:

- Improve data acquisition techniques by adding more validation of structures added by users
- Improve prediction by adding more nodes into the Bayesian network.
- Integrate the prediction tool into NOMAD and test the application using A-B testing.
- Explore the use of other drawing tools than the Ketcher library that was used within this application.