The Creation of a 2-D Web-Based NMR Prediction Program

James Andrew Surface – Hampden-Sydney College
Fall 2007

Introduction

The Internet has grown in its ability to host scientific applications in the past decade, and the web's potential for conducting calculations, storing data, and networking results for fast distribution has only just begun to be realized. One particular new technology, built off of an old framework of javascript and XML (a variant of html), called AJAX, is a new and emerging development framework that begins to realize the framework capability of the web. AJAX is a web2.0 programming methodology, and the usage of the paradigm has been popularized by Google in its star-studded GMail webmail application. Essentially what AJAX gives is a thinking paradigm to treat the web as a basis for computing applications by giving live-time results (or nearly live time results) by transforming the web-browser into a visual application on a client's computer that communicates with the main, highly-specialized server live-time. The idea that a program can be based on one highly-specialized computer (the server), and accessed from anywhere in the world via a regular web browser, is the framework that this project intends to explore. This proposal intends to layout the web-framework basis for designing a program that will enhance the accessibility of NMR spectra data and produce spectra predictions based on pure molecule structure. The program will furthermore predict 2-D spectral data as well, and will employ an intuitive interface that maximizes accessibility and usability by anyone anywhere.

It is generally known that scientific programs lack the intuitive, user-friendly features that characterize computer programs for the general populace. This problem is mostly due to the small client-base associated with most scientific computer literature and often the lack of resources devoted tightly-funded projects to so-called “eye-candy”. Nevertheless, because there is very little effort made to make programs more usable, such programs have difficulty in reaching the classroom and helping to educate new generations of students. Much benefit is therefore lost, benefit that would do nothing but help the future scientific community gain and grow and advance. A web-based framework, like the one being proposed, will not only meet those demands reached by so few scientific applications, but will also prove to be a valuable research and teaching tool in educational institutions.

Methods

It is important to note that this proposal does not incite an idea that is entirely novel. Many researchers have gone before to utilize the web's resources in delivering tools to the scientific community that have since proved helpful towards both education as well as research. Examples of this abound, from the famous Kegg (http://www.genome.jp/kegg/) Encyclopedia, an interactive, highly-designed framework that enables users to search and trace metabolic pathways at the genomic level, to the online PDB (http://www.rcsb.org/pdb/home/home.do). These tools have become successful because they are both highly accessible, but are also usable, and reasonably easy to learn.

One of the defining characterization methods in chemical research is Nuclear Magnetic Resonance Spectroscopy, especially amongst small organic compounds. This method of characterization is especially powerful because it measures the types of carbon based on their electromagnetic shift in a strong magnetic field. Thus, a compound can be characterized on the basis of its "types" of carbons. ^{1}HNMR is a magnetic resonance method that measures the chemical shifts of hydrogen atoms in a similar method to ^{13}CNMR, and it differentiates the types of Hydrogen atoms and provides information on the number of neighboring hydrogen atoms based on the splitting of the individual peaks. There is also another type of NMR that correlates the ^{1}HNMR and ^{13}CNMR spectra together, showing which peaks of the hydrogen are correlated to which carbons, that is which hydrogen atoms are actually connected to which carbons. This is a particularly powerful form of NMR spectroscopy because it enables larger molecule spectra to be evaluated. It also gives a nearly-perfect picture of the molecule being evaluated because it eliminates most of the peaks that 1D NMR sometimes cannot explain. There are many different types of 2-D NMR that correlates ^{1}HNMR to ^{1}HNMR and ^{1}HNMR to ^{13}CNMR. There are different pulse sequence methods for the NMR instrument to attain these correlations as well. The one that this proposal uses is HMQC 2-D NMR, which is a simple ^{13}C and ^{1}H correlation sequence. The spectrum attained from such an experiment on the NMR is a simple 2D correlation with carbon on one axis (typically the x-axis) and hydrogen on the other (typically the y-axis).

The difficulty with NMR, despite its powerful ability to attain "pictures" of molecules based on their carbon and hydrogen content, is that the spectra produced from these experiments must be "read" and understood. The general rule when running an NMR experiment on the NMR instrument is to make a prediction using NMR shift tables on the proton and carbon shifts that *should* appear on the experiment results. This forces the scientist to notate the shifts he should be expecting. The problem with this method is that there is a limitation on the size of the molecule that can be feasibly predicted. Not just that, but most of the carbon and proton shifts as predicted in shift tables are not altogether very accurate. If there is a small molecule with a lot of carbon shifts that will be in a similar range, it is virtually impossible to predict the spectrum accurately. Because of this some automated programs have arisen that actually predict complex carbon and proton spectra based on more complicated algorithms than humans can implement on the fly. These programs work well, but are limited in accessibility, are quite expensive, and usually have limits to the number of atoms allowed in the predictions.

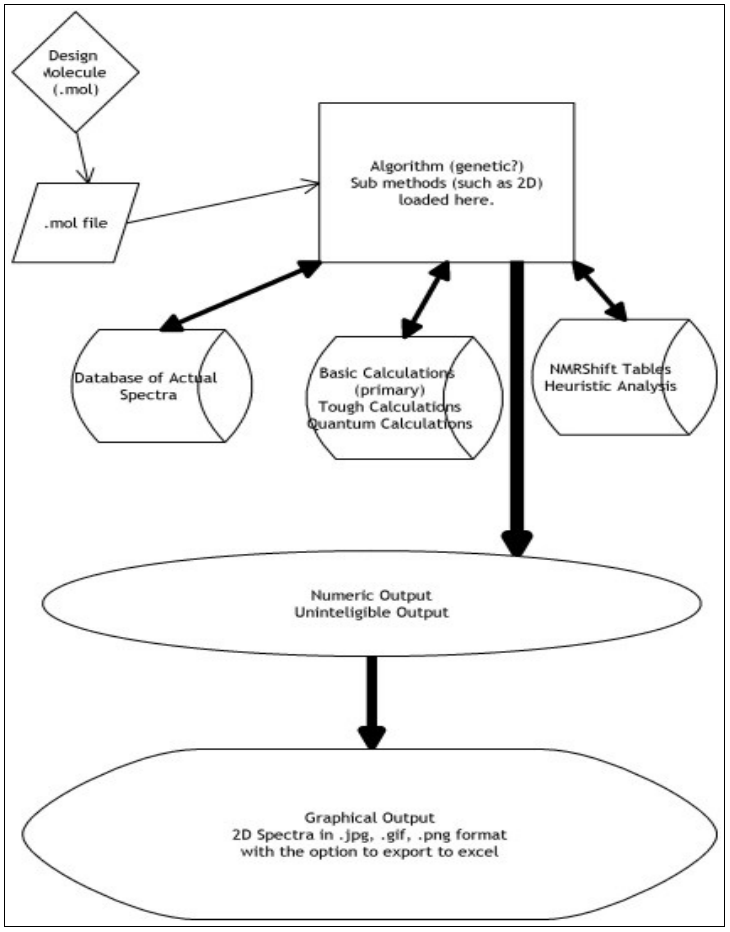
The problems with NMR prediction could easily be resolved by creating a web-based NMR prediction system, beginning with small molecules. To make the data better for prediction, a standard for molecule description should be created, a standard for the storing of this data, and a standard for developing an algorithm that can predict spectra based on the structure of the molecule.

The best method to develop such a program is to use a parser that processes the input information of the molecule who's spectrum is to be predicted. The parser then passes this molecular structure information from input to the algorithm which then performs a database search for the molecule. The database search searches for similar molecules, or ideally identical molecules, and then looks up the spectra already saved. The algorithm is also capable of completing basic and complex mathematical calculations on the molecule to predict the spectrum using a combinatorial method that

incorporates both nmr shift tables and quantum mechanical calculations, idealized for being utilized in the programming environment that is being run. This is a “divide and conquer” methodology that predicts spectra very accurately in practice, at least at this point, on small molecules. This methodology that is of this author's own invention description, is already in practice using along similar ideas by several programs on the web. Although they do not utilize the methodology as described exactly, it is similar enough to call the basic idea the same. The foremost of these programs is written in PHP/MySQL and is an opensource application that is community-maintained and community-trained. It is called NMRShiftDB, and works by uploading a .mol file onto their server which then parses and evaluates the molecule based on data contained within the file. The advantage to using this format is that the .mol file can already be edited and run using a molecular modeling program and energy-minimized such that the atoms are as accurately placed as possible. NMRShiftDB creates a list of different peaks for both ^{1}HNMR and ^{13}CNMR, tabulating them, then producing a predicted spectra in an image format projection. This is both a good visual aid and a way of predicting spectra that the normal human read cannot even do. It is very fast, and very accessible, although not the most user-friendly as many of its options and applications are difficult to find and to understand.

Diagram 1.1

The NMR Prediction Methodology



It is therefore possible to use the NMRShiftDB program as a foundation program upon which to build a more intuitive web-based NMR spectra prediction system. The project would not only encompass making the interface more user-friendly, but it would also include learning the "language" of .mol files (and other molecule structure files), making and storing NMR spectra (contributing to the already-existing NMRShiftDB database), but also creating an add-on application that will utilize the information stored in the database, and the spectra predictions from raw predictions on unknown molecules (molecules not stored in the database) to produce a 2-D HMQC NMR prediction.

There are generally four variables in the acquisition of 2-D HMQC NMR spectra. Those four variables are first, the ^{1}HNMR spectrum, second, the ^{13}CNMR spectrum, the structure of the molecule (the actual accurate interpretation of the 1D spectra), and finally the correlation data to tie the two 1D spectra together. Without knowing the interpretation ahead of time, the NMR instrument gathers the correlation data through a complicated pulse sequence that analyzes protons and carbons nearly simultaneously. What if we were to want to get the 2D NMR spectrum with JUST the two sets of 1D data? Well, according to the presupposition proposed above, we'd have to have the interpretation information from the 1D spectra, that is know which peaks from ^{1}H and ^{13}C spectra correspond to which hydrogen and carbond atoms. If we have three of the variables above, we can calculate the fourth. Thus, for regular 2D NMR experiments the instrument has to analyze the molecule because it doesn't know which ^{1}H and ^{13}C peaks correlate to which hydrogens and carbons, BUT, if we already have that information (contained in the molecule file analysis in the program) then we already know which peaks go to which atom. The 2D NMR spectrum can then be easily predicted, because we have the molecule structure already. HMQC theoretical predictions would be very powerful for interpretive analysis, and would be a fast way of knowing whether the results from the NMR could possible be the target molecule, the molecule for which we have the theoretical calculations from the 2D NMR prediction program.

Potential Results

The discussions about this project prior to this proposal have been full of mixed feeling. It is certainly an ambitious project, but the author is confident that the results desired can be attained, despite the heavy learning curve. Finding a 1D web-based prediction program that is already in place is an important discovery that has saved much time and made the 2D prediction portion actually possible. The first step in the project is to familiarize myself with taking 2D NMR spectra. I've already taken nearly 30 2D NMR spectra, experimenting by changing different parameters and seeing what affect that has on the final result. While editing the spectra virtually, I've learned about the structure of the spectra and have attained a good feel for using and reading the spectra. The next step is to learn the different molecule storage formats, such as SMILES and .mol. During this step I also need to learn the NMRShiftDB methodology and program so that I can figure out how to integrate the 2D methodology into the system. The primary goal is to first get it predicting x and y datapoints for the 2D spectra, and then to send those datapoints to a graphing application that will then display it in an image that can be saved.

IF this works (which it will undoubtedly, the only concern is getting it to work in the time we have) then there will be an amazing program that H-SC will be able to host on their webservers, or just contribute back to the NMRShiftDB project. Ideally, at the end, a user-friendly, intuitive interface will have been made for NMRShiftDB along with an add-on that enables 2D prediction capabilities in addition to the 1D already existing.

References

[1] W. Bremser. (1985). Expectation Ranges of ^{13}C NMR Chemical Shifts. *Mag. Res. in Chem.* **23**, 271-275.

[2] Murray-Rust, P.; Rzepa, H.S. (1999). Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **39,** 928-942.

[3] Thomas, S.; Strohl, D.; Kleinpeter, E. (1994). Computer Application of an Incremental System for Calculating the ^{13}C NMR Spectra of Aromatic Compounds. *J. Chem. Inf. Comput. Sci.* **34,** 725-729.

[4] Weininger, D. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28,** 31-36.

[5] PHS 398/2590 (Rev. 05/01) Revised Application to NIH for HINT Biomacromolecular Prediction Programs. Kellogg, Glen Eugene.

[6] Blinov, K.; Kvasha, M.; Lefebvre, B.; Sasaki, R.; Williams, A. NMR Prediction Accuracy Validation. ACD/CNMR Predictor. Version 10.05. [http://acdlabs.typepad.com/NMRShiftDB_Validation.pdf]

[7] *NMRPredict* - http://www.modgraph.co.uk/product_nmr.htm , http://www.mestrec.com/index.php?idtp=18&i18n=1

[8] *NMRShiftDB* - http://www.nmrshiftdb.org

[9] *CSEARCH* - http://nmrpredict.orc.univie.ac.at/csearchlite/comparison.html

[10] http://drx.ch.huji.ac.il/nmr/techniques/expts.html

[11] *XdrawChem* - http://xdrawchem.sourceforge.net/

[12] Steinbeck, C.; Kuhnm, S. (2004) NMRShiftDB – compound indentificationand structure elucidation support through a free community-built web database. *Phytochemistry*. **65,** 2711-2717.

[13] Munk, M.E. (1998). Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **38,** 997-1009.

[14] Meiler, J.; Maier, W.; Will, M.; Meusinger, R. (2002) Using Neural Networks for ^{13}C NMR Chemical Shift Prediction-Comparison with Traditional Methods. *J. Magn. Res.* **157,** 242-252.