CMPUT 605 Syllabus Sept. 8, 2016

Title of Course: Applications of Machine Learning Towards Spectral Prediction

People Involved: David Wishart (Professor), Tanvir Sajed (Student)

Course Goals: This individual study course will involve exploring how machine learning techniques, especially probabilistic graphical models and hidden Markov models can be used to predict mass spectrometry spectra and NMR spectra. The main goals of the course are to: 1) gain an understanding of the underlying mechanisms and algorithms for MS and NMR spectral prediction; 2) learn how these PGM/HMM methods can be integrated with cheminformatics techniques to enhance the overall performance of MS and NMR spectral predictors and 3) to create a program or modify an existing program that performs accurate NMR or MS spectral prediction using chemical structures as the input.

Workload: The student will review papers on CFM-ID and relative spectral prediction techniques (see reading list below) and will discuss what they have learned at weekly meetings. They will also assembly a database of "training" NMR and/or MS spectral data from public databases. The student will prepare a brief (10-15 minute) powerpoint presentation on what they have read, learned or done and present that information at the weekly meetings. The student will write an original or extensively modify an existing program to perform NMR and/or MS spectral prediction (from input chemical structure data) and assess its performance against existing state-of-the-art programs. Finally, the student will write a 10-15 page paper describing what they have done and how the program was implemented/tested.

Reading List:

1) Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification.
Allen F, Pon A, Greiner R, Wishart D.
Anal Chem. 2016 Aug 2;88(15):7689-97

2) MetFrag relaunched: incorporating strategies beyond in silico fragmentation.
Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S.
J Cheminform. 2016 Jan 29;8:3. doi: 10.1186/s13321-016-0115-9

3) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra.
Allen F, Pon A, Wilson M, Greiner R, Wishart D.
Nucleic Acids Res. 2014 Jul;42(Web Server issue):W94-9.

4) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification.
Allen F, Greiner R, Wishart D.
Metabolomics (2015) 11: 98-107. doi:10.1007/s11306-014-0676-4

5) Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction.
BMC Bioinformatics. 2008 Sep 25;9:400. doi: 10.1186/1471-2105-9-400.
Kuhn S, Egert B, Neumann S, Steinbeck C.

6) Evaluation of a (1)h-(13)c NMR spectral library.
Smith SK, Cobleigh J, Svetnik V.
J Chem Inf Comput Sci. 2001 Nov-Dec;41(6):1463-9.

7) Fast metabolite identification with Input Output Kernel Regression.
Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J.
Bioinformatics. 2016 Jun 15;32(12):i28-i36.

8) Searching molecular structure databases with tandem mass spectra using CSI:FingerID.
Dührkop K, Shen H, Meusel M, Rousu J, Böcker S.
Proc Natl Acad Sci U S A. 2015 Oct 13;112(41):12580-5.

Timeline: Weekly meetings will be held on Wednesday afternoons (3:00 to 4:00 pm) in Athabasca 3-41 or failing that, on days/times that are mutually convenient. The first meeting will be held on Sept. 14. The deadline for completing the readings is Oct. 1. The deadline for completing the spectral databases is Oct. 15. The deadline for completing the program(s) is Dec. 1. The deadline for submitting the final paper is Dec. 10.

Evaluation Method: The project will be evaluated using the following criteria:
Participation and attendance – 10%
Quality of weekly presentations – 10%
Final program (performance, capability, robustness) – 40%
Final paper (quality of writing, completeness) – 40%