Dr. Shatkay, The Computational Biology and Machine Learning Lab.

Projects are available in biomedical data and text mining and machine learning.

Four specific projects are detailed below, concerned with genomics of autism, disease-gene association and SNP selection, protein subcellular localization, and biomedical text mining. If you are interested in working in the areas of biomedical data mining, and/or biomedical text mining, please email me at: *shatkay@cs.queensu.ca* to get more information about these and other possibilities.

As has happened before, results from outstanding projects may be published/presented in major conferences, be used by the international scientific community, and lead to an MSc thesis.

Specific projects:

1) Genomics of Autism. Narrowing the search for candidate genes.
Jointly with Dr. Jeanette Holden, Depts. of Physiology and Psychiatry and the Ongwanada resource centre.

"Autism" is not a single condition but a spectrum of disorders with a variety of symptoms, and multiple genetic and environmental causes. While autism spectrum disorders are devastatingly common (occurring at about 0.6% of the population), their genetic causes are not known or well-understood, and are thus the focus of much current research.

Dr. Holden and her collaborators have identified genomic regions that are highly mutated in individuals with Autism.
The goal of this project is to identify the genes located in those regions, and to try and explain the possible relationships among these genes. We will employ several bioinformatics tools, along with information from the Gene Ontology and methods in text analysis, to realize this goal.

The project can accommodate a team of two students working together.

Prereqs: Biomedical computing background and strong programming skills.

2) SherLoc. Improving protein subcellular location prediction.
Jointly with Dr. Kohlbacher's group, Division of Simulation of Biological Systems, University of Tübingen, Germany.

Knowing the location of proteins in the cell is an essential step toward understanding their function and their role in biological and physiological processes. As many proteins are known only by their postulated sequence, computational localization prediction is one of the fundamental tools used in current large-scale biology research.

Our group, in collaboration with Oliver Kohlbacher's group has developed SherLoc, which is a comprehensive and currently the most accurate system for computational prediction of protein location. (For some more details, see for instance:
http://bioinformatics.oxfordjournals.org/cgi/reprint/23/11/1410,
http://pubs.acs.org/subscribe/journals/jprobs/6/i07/html/0707toolbox.html,
http://www-bs.informatik.uni-tuebingen.de/Services/SherLoc/).

The current project is concerned with integrating several advanced features recently developed by Scott Brady (a former MSc student in the lab), improving the human interface to SherLoc, and studying several possible alternatives that may further improve its prediction accuracy for hard-to-predict locations and proteins.

The project can accommodate between 1 and 3 students.

Background in biomedical computing, strong programming skills, (experience in web programming is a plus), and a good understanding of probability and statistics.

3) **Integrating bioinformatics tools and databases for assessing the functional significance of SNPs**

Much effort in current human genomics, epidemiology, and pharmarcogenomics is focused on the identification of genetic variations that are responsible for common and complex human diseases. Specifically, single nucleotide polymorphisms (SNPs), i.e. substitutions of single nucleotides, are in the forefront of such studies, as they form the majority of genetic variations in the human population. Reliable identification of disease-causing SNPs is expected to enable early diagnosis, personalized treatment, and targeted drug design.

We have recently developed a new resource, F-SNP, integrating computationally predicted functional information about SNPs from 16 bioinformatics tools and databases, while particularly aiming to facilitate identification of disease-causing SNPs. F-SNPs thus helps researchers identify and focus on SNPs with potential pathological effects.

In this project, students will improve the contents and the usability of the F-SNP database by further integrating additional bioinformatics tools and databases that can help assess the deleterious functional effects of SNPs. Specifically, this project will involve:

A. Integrate data from human disease databases to identify candidate genes for specific target disease.
B. Survey additional bioinformatics tools that predict the deleterious effects of SNPs at the *transcriptional* and the *post-transcriptional* level, and integrate some of them into the F-SNP database;
C. Implement a web-service search page for user-defined queries.

Students working on this project will get familiar with the state-of-the-art in human disease databases and learn about SNP function prediction. In terms of programming, the project involves both development and integration of web services, and handling a large-scale MySQL databases.

This project can accommodate two students. Background in Perl CGI and/or Java programming are needed. Experience in database programming is a plus.

4) **BLIMP: Biomedical LIterature Mining Publications forum.**
Extending, and enhancing the widely-utilized web-based resource, BLIMP.

A former 499 student, Limin Zheng, has created a fabulous and widely used web site called BLIMP: Biomedical Literature and Text Mining Publications Forum. (Visit http://blimp.cs.queensu.ca).
Researchers are using and commenting on it, and there are several challenging extensions that will make it an even better tool.

This year's project is to create new functionalities in BLIMP, making it smarter, more accessible and more widely useable. Your contribution will be highly visible, and will be used by hundreds of prominent scientists who will also get to know your name ☺

Necessary skills: Programming in Java, javascript, Perl, and HTML. Experience with Unix. Experience with MySQL is a plus.