IMA Program Application - Personal Statement

Large Data Sets in Medical Informatics, November 14-18, 2011

Nathanael Fillmore

In this document, I describe my scientific interests, research plans, and reasons for wishing to participate in the IMA's Large Data Sets in Medical Informatics program.

1 Scientific interests and research plans

I am interested in statistical machine learning and its applications to biology and medicine (and the humanities). My current particular focus is on using probabilistic generative models to study different aspects of gene expression. This includes "downstream" analysis, to study problems of direct medical importance, and "upstream" analysis, to evaluate and improve our future ability to do such studies.

1.1 Progression and gene expression in cervical cancer

With Michael A. Newton, Paul F. Ahlquist, and Paul Lambert, I am working on a model of how gene expression in cervical tissue changes through a progression from normal tissue to cancerous tissue. A key distinguishing feature of our model, compared to those typically used in gene-expression studies, is that we postulate the existence of several distinct (and unknown) types of cells which are present in all stages of the progression, but whose relative proportions (also unknown) change during the course of the progression. We then study differential expression across the postulated cell types. One immediate goal is to apply this model to an analysis of 128 cervical tissue samples, each measured by an Affymetrix whole genome microarray.

1.2 De novo transcriptome assembly from RNA-seq data

With Colin Dewey and Bo Li, I am working on de novo transcriptome assembly from RNA-seq data, i.e., assembly without knowledge of a reference genome. Our current goal is to develop a principled way to evaluate de novo asssemblies. Evaluation is difficult since, by assumption, no reference is available. To mitigate this issue, we are working on a probabilistic model of the process of generating RNA-seq reads from a transcriptome's isoforms; we use a statistic based on this model as our evaluation criterion.

Note that de novo transcriptome assembly is of importance for certain types of medical research, even though a high quality human reference genome is available, because some other organisms of interest to the research do not have reference genomes available. For example, James Thompson's lab at UW-Madison is interested in understanding how the Axolotl (Ambystoma mexicanum) is able to regenerate body parts, with the ultimate goal of applying this knowledge to regeneration in humans. The Axolotl does not have a reference genome available, and no reference genome is expected to be available in the foreseeable future, particularly since its genome is approximately 10 times larger than the human genome.

2 Reasons for wishing to participate in the program

The IMA's program is appealing to me for several reasons. First, the theme of the workshop is directly relevant to my own research. For example, a key issue in the cervical cancer study described above is that although our dataset is rather large - an Affymetrix whole genome microarray contains about 54,000 probe sets - only only around 30 tissue samples are available at each stage of the progression, so some information sharing is needed to make statistically valid conclusions. RNA-seq datasets are even larger - a collection of reads from just one sample can run into 10s of gigabytes - yet it is difficult to make sense of all this data, and often even more difficult to rigorously justify one's conclusions.

Second, although I have been developing an expertise in machine learning and related areas for several years, I am just getting started with applications to biology and medicine. There is much that I do not know about current research directions and challenges. The IMA's program would be an excellent opportunity to expose myself to a variety of different research problems in medical informatics.

Finally, the program's description mentions the difficulty in making productive use of results developed in mathematics, statistics, and elsewhere for applications in biomedical informatics. I have already experienced this difficulty myself. I think it would be helpful to me and quite interesting to see what other researchers have been doing along these lines.