Biomedical and Health Informatics Lecture Series
Tuesday, January 29, 2008
Room RR-134, 12:00 - 12:50 p.m.

William Noble, PhD

Associate Professor of Genome Sciences and of Computer Science, Univ. of Washington

"Consistent probabilistic outputs for protein function prediction"

In predicting hierarchical protein function annotations, such as terms in the Gene Ontology, the simplest approach makes predictions for each term independently. However, this approach has the unfortunate consequence that the predictor may assign to a single protein a set of terms that are inconsistent with one another; e.g., the predictor may assign a specific GO term to a give protein ("purine nucleotide binding") but not assign the parent term ("nucleotide binding"). Such predictions are difficult to interpret. In this work, we focus on methods for calibrating and combining independent predictions to obtain a set of probabilistic predictions that are consistent with the topology of the ontology. We call this procedure "reconciliation."

We begin with a baseline method for predicting Gene Ontology terms from a collection of data types using an ensemble of discriminative classifiers. We apply the method to the data sources available in the MouseFunc assessment~\cite {pena-castillo:critical}, and we demonstrate that the resulting predictions are frequently inconsistent with the topology of the Gene Ontology. We then consider 11 distinct reconciliation methods: three heuristic methods, four variants of a Bayesian network, an extension of logistic regression to the structured case, and three novel projection methods: isotonic regression and two variants of a Kullback-Leibler projection method. We evaluate each method in three different modes -- per term, per protein and combined -- corresponding to three types of prediction tasks. For each of the biological process, molecular function and cellular component ontologies, we also consider several axes of evaluation corresponding to different recall values and different ranges of term sizes.

Our results show that the isotonic regression method generally performs well across evaluation modes, term sizes, ontologies and recall levels. Reassuringly, isotonic regression usually performs better than the underlying, unreconciled logistic regression method, and almost never performs worse. This implies that the structure of the Gene Ontology network can yield valuable information. On the other hand, successfully exploiting this information turns out to be difficult, and many apparently reasonable reconciliation methods end up yielding reconciled probabilities with significantly lower precision than the original, unreconciled estimates. Of particular interest are Gene Ontology terms to which few proteins have been assigned. For these terms, isotonic regression is not always convincingly better than logistic regression. If small terms are of particular interest, then we suggest using a Kullback-Leibler projection method instead.

William Stafford Noble (formerly William Noble Grundy) received the Ph.D. in computer science and cognitive science from UC San Diego in 1998. After a one-year postdoc with David Haussler at UC Santa Cruz, he became an Assistant Professor in the Department of Computer Science at Columbia University. In 2002, he joined the faculty of the Department of Genome Sciences at the University of Washington. His research group develops and applies statistical and machine learning techniques for modeling and understanding biological processes at the molecular level. Noble is the recipient of an NSF CAREER award and is a Sloan Research Fellow.

NEW: **Podcasts from MEBI 590 Lecture Series talks from earlier this quarter are available at http://courses.washington.edu/mebi590/schedule.htm**
Podcasts from Fall Quarter 2007 are available at
http://courses.washington.edu/mebi590/2007.Q4.Fall.htm