

Keith Noto

University of California at San Diego, 9500 Gilman Drive (EBU3B,MC0404), La Jolla, CA 92093
ph: (858) 534-9655 · fax: (858) 534-7029 · knoto@cs.ucsd.edu

Research Interests

My primary research goal is to advance the field of machine learning and to investigate machine learning approaches to open problems. I have particular interest in probabilistic models, such as Bayesian networks and hidden Markov models, and to biological problem domains (bioinformatics).

Machine learning is the area of artificial intelligence that involves designing computer programs that improve with experience. *Supervised learning* concerns algorithms that infer generalized models from training examples. One of my specific research interests is improving state-of-the-art computational models for various tasks by making them more *expressive*, and developing novel algorithms to learn the form and parameters of these models from data. The expressivity of a model involves the number, type, and kinds of relationships among the features of a system that the model is able to represent. Molecular biology provides excellent examples of the need for expressive models. Genes are regulated by complex systems that are not fully understood. These systems involve a variety of factors beyond the genomic structure, including proteins, RNAs and other small molecules in the cell. As scientists continue to discover more factors that are involved, the most accurate scientific theories will be increasingly complex. Computational models inferred automatically from data will need to be correspondingly complex, because they will be used to provide insight to biologists about which factors are relevant in an enormous variety of cellular conditions.

The shift toward more expressive models presents difficult computational challenges. New types of models will require the development of novel algorithms to learn them. More expressive model spaces generally require more data to learn, which makes it a challenge to develop tractable algorithms, and also creates the need for new regularization techniques. Finally, these approaches will require increasingly sophisticated statistical analyses to measure their accuracy, and the effectiveness of different aspects of the models.

I am also interested in collaborating with experts to develop tools that make an impact in the scientific community. I make it a policy to publish my research code along with papers, but I am interested in going further and releasing tools and online resources so that biologists and experts in other domains will find it easy to take advantage of state-of-the-art machine learning solutions.

Research Experience

My graduate research focused on algorithms which learn models that explain how certain classes of genes are differentially expressed (that is, the extent to which genes are transcribed into RNA, for the purpose of making specialized proteins). This includes learning models of gene regulatory relationships, *i.e.* discovering which genes are involved in the activation and repression of which other genes, and learning models of the mechanisms by which this regulation is achieved (Noto and Craven, 2004, 2006a, 2006b). A primary mechanism in which genes are regulated is by proteins that recognize certain subsequences of DNA and bind to those locations in a control region, called a promoter, near the gene's protein coding region. There are often multiple protein factors at work, so genes that are regulated together share a common *pattern* of binding sites, sometimes called a *cis*-regulatory module. Some instances of these modules have been discovered and verified in laboratories, but in general the relevant binding sites that regulate a particular set of genes in a particular set of cellular or environmental conditions are unknown, and biologists are very interested in discovering the details of these modules.

A major contribution of my research is to learn models that are able to represent not only the DNA letters that make up these binding sites, but also the conserved relationships between them, such as which strand of the DNA double-helix a given type of site tends to be found, relative order between sites, and the conserved distance between sites. My experiments demonstrate that modeling these aspects increases the accuracy of the inferred modules, which is an important step forward in this area.

Fully sequenced genomes and high-throughput data sources such as microarrays provide a wealth of genomic data, and it is clear that computational methods are going to continue to play a significant role in analyzing these data. However, existing methods for analyzing sequence and discovering *cis*-regulatory modules are unable to make use of the real-valued expression measurements directly. One typical approach for mapping variable-length sequences or other types of unstructured data to real numbers is to turn the task into a classification problem by attempting to cluster the expression measurements into classes, but this approach loses information about specific expression values and there are no guarantees that the cluster definitions are correct. Another approach is to first learn a model of the sequence data, and then to map features of that model to the expression values. The problem with this approach is that there are too many potential model features to consider them all, so these types of approaches need to use some kind of search bias which may not have the right relationship with the real-valued responses that the model attempts to explain.

Another significant contribution of my research is an approach that is able to use the expression measurements directly to learn the parameters of the sequence model (Noto and Craven, 2008). It uses hidden Markov models (HMM), and therefore can represent uncertainty about the presence of features (e.g. a binding site that generally conforms to a consensus, but allows for variation), as well as sophisticated sequence features (e.g. the relative order or distance between binding sites). By using the real-valued response data to assign a likelihood distribution over the number of occurrences of each of a set of binding sites, my approach uses a modified version of the HMM backward procedure to propagate the likelihood information back through the model to characterize those binding sites.

As a postdoc, my research has focused on identifying relevant protein data for the Transport Classification Database (TCDB; www.tcdb.org), a specialized biological database. One primary task is to learn a classifier that distinguishes published literature that is relevant to TCDB from literature that is irrelevant, for the purpose of screening new publications, such as the thousands added to the Medline database each month. TCDB, like hundreds of other specialized databases, maintains a list of references to relevant published literature. It does not maintain a list of *irrelevant* literature. A significant contribution of my research is to show that under reasonable assumptions, one can learn a classification model from positive and *unlabeled* training examples that predicts the likelihood of new labels just as accurately as a model trained on both positive and negative examples (Elkan and Noto, 2008, Noto *et al.*, 2008).

Future Directions

Most of the significant results in discovering *cis*-regulatory modules have been achieved analyzing bacteria, yeast or even insects, but mammals have much larger and more complex control mechanisms by comparison. Modeling regulatory modules in mammals is a challenge because protein binding sites can be quite distant from the corresponding protein coding regions, and because other small molecules such as microRNAs may play a more significant role in regulation. Discovering these modules in mammals is a computational challenge as well, because of the amount of DNA sequence that must be analyzed. Even approaches that run in linear time may be too slow to provide useful tools, but I believe that many probabilistic methods can be made tractable by identifying and omitting calculations involving factors with negligible likelihood (Noto, 2007). I am interested in continuing this research to develop models for mammalian genomes that account for a variety of factors such as microRNA molecules, and to develop tractable algorithms and usable tools for inferring these models.

I believe that there is great potential in approaches like my HMM-based regression algorithm (Noto and Craven, 2008), but that this potential is under-investigated. For instance, this particular algorithm has applications for other sequences of events, like user activity on the web, or any kind of time-series data. I am interested in finding new applications for this approach, but beyond this, I believe that this research represents an important advance in machine learning algorithms because it takes what is fundamentally a classification algorithm and adapts it for a different type of problem setting. I am interested in generalizing other probabilistic approaches, especially probabilistic graphical models such as Bayesian networks and conditional random fields, so that they can be used with continuous variables and arbitrary probability density functions.

A further aspect of my postdoctoral research with TCDB is that, beyond classifying documents for relevance to TCDB, we need to extract the relevant information from these articles, such as protein names and their sequences (Saier *et al.*, 2009). Part of this information is not stored in the document, but it can be found in other documents or in online databases. Named entity recognition (NER) and information extraction (IE) are important open problems in natural language processing, and they are both very difficult tasks. I am interested in taking advantage of the fact that there are large, freely accessible online databases, Swiss-Prot for example, that can provide background information for doing NER and IE specifically on biological texts. I am currently researching methods to use online databases to help detect protein names and features in documents, both to automate the process of curating TCDB, and to provide key additional document features for learning classifiers. These techniques have applications not just to TCDB, but to hundreds of specialized databases, and I am currently in the process of preparing an online resource for the curators of those databases to benefit from pointers to new publications of likely relevance. This resource will be available by the time my postdoc ends, but I am interested in continuing to use it as a testbed for research on document classification and information extraction specifically for biological texts.

References

- Noto and Craven, 2004.
K. Noto and M. Craven. Learning Regulatory Network Models that Represent Regulator States and Roles. RECOMB 2004 Workshop on Regulatory Genomics. In *Lecture Notes in Bioinformatics* 3318, pages 52-64. Springer-Verlag, 2004.
- Noto and Craven, 2006a.
K. Noto and M. Craven. Learning Probabilistic Models of *cis*-Regulatory Modules that Represent Logical and Spatial Aspects. *Proceedings of the 2006 European Conference on Computational Biology*, In *Bioinformatics* 23(2):e156-162.
- Noto and Craven, 2006b.
K. Noto and M. Craven. A Specialized Learner for Inferring Structured *cis*-Regulatory Modules. *BMC Bioinformatics*, 7:528, 2006.
- Noto, 2007.
K. Noto. Learning Expressive Computational Models of Gene Regulatory Sequences and Responses. PhD thesis, Department of Computer Sciences, University of Wisconsin-Madison.
- Noto and Craven, 2008.
K. Noto and M. Craven. Learning Hidden Markov Models for Regression using Path Aggregation. *Proceedings of the 24th Uncertainty in Artificial Intelligence Conference (UAI 2008)*, 444-451.
- Elkan and Noto, 2008.
Learning Classifiers from Only Positive and Unlabeled Data. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, 213-220.
- Noto *et al.*, 2008.
K. Noto, M. H. Saier, Jr. and C. Elkan. Learning to Find Relevant Biological Articles Without Negative Training Examples. *Twenty-First Australasian Joint Conference on Artificial Intelligence (AI 2008)*, (To appear in *Lecture Notes in Bioinformatics*. Springer-Verlag.)
- Saier *et al.*, 2009.
M. H. Saier, Jr., M. R. Yen, K. Noto, D. G. Tamang, and C. Elkan. The Transporter Classification Database: Recent Advances. *Nucleic Acids Research* 2009 Database issue (to appear).