Shared challenges in data mining and computational biology

Charles Elkan
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093-0114
elkan@cs.ucsd.edu

ABSTRACT

This talk will explain several fundamental research issues that must be addressed to enable continued progress both in computational biology and in data mining. The issues include:

1. Supervised and unsupervised learning where the important class is very rare.
2. Estimating accurate, well-calibrated probabilities.
3. Measuring the statistical significance of rare predictive features.
4. Feature selection where only combinations of features are predictive.
5. Learning where the number of features is far higher than the number of examples.
6. Deciding which examples to find labels for experimentally, when labeling examples is expensive.
7. Learning when test examples and training examples are drawn from different populations.

For each challenge I will explain at least one real-world scenario where it arises, and I will show how a commonly used learning method fails.