Keynote Talk

The Minimum Informative Subset Problem

Sorin Istrail, Ph.D.
Senior Director, Informatics Research
Celera Genomics/Applied Biosystems
45 West Gude Drive
Rockville, MD 20850

A variety of challenging problems arising in computational biology when formulated as partition-distinguishing optimization problems can be cast into a common general framework that we call “the minimal informative subset:” Given a set of objects, find a minimal set of “attributes” of the objects that are “informative” with respect to the optimally distinguished partitions. The general framework can be formulated as a set-cover based feature-selection. This intertwining between discrete algorithms and statistics is interesting to explore in the search for effective heuristics. Although the general problem is hard, biology often provides additional constraints making the problems tractable for important special cases. We will discuss the framework and its application to several computational biology problems: SNP selection/haplotype tagging, assay design, disease associations, literature retrieval and tissue classification.