

The futility of Maximum Likelihood Estimation in Genomics
and an Alternative Ensemble Based Estimator

Thursday, October 19, 2006
Biolabs Lecture Hall, Harvard University

Professor Charles (Chip) E. Lawrence
Division of Applied Mathematics
Center for computational Molecular Biology
Brown University

Advances in data collection technologies have rendered increasingly large data sets available for analysis. While the emergence of such large data sets would seem to lead to increasingly more precise estimates of parameters, paradoxically just the opposite seems to be becoming increasingly common. This paradoxical circumstance has emerged because these technologies have simultaneously opened opportunities to draw inferences on previously unanswerable high dimensional questions. For decades maximum likelihood estimation (MLE) and related optimization procedures have been employed as the major tool of most inference procedures. Even though favorable properties of such optimization based inferences rest on an asymptotic foundation that requires the data to grow in comparison with the number of unknowns, such optimal estimators continue to be widely used even when these supporting conditions are not present. Genomics and computational molecular biology are among the more predominate fields experiencing the duality of the growth in data resources and inference expectations. In fact, prediction and inference of high dimensional objects are now arguably the most important activities in these allied new biological fields, and the inspiration for this talk. RNA secondary structure prediction offers a very special lens to examine the untoward consequences of the reliance on the MLEs in high dimensional inferences because polynomial time algorithms are available to comprehensively characterize the space of solutions, and a references set of structures is available for the comparison of alternative prediction methods. Through this lens we will examine these untoward effects, consider their boarder implications, and present an alternative “ensemble based” inference procedure.