Statistical inference in high dimensional spaces of genomics: an RNA structural example

Charles E. Lawrence
Division of Applied Mathematics, Center for Computational Molecular Biology, Brown University, Providence (RI), United States

The emergence of genome scale data sets leads to increasingly more precise parameter estimates that are ideally suited for maximum likelihood methods and other highest scoring procedures, when the number of unknowns is modest. However, paradoxically just the opposite is becoming increasingly common in genomics. This paradox has emerged because these technologies have simultaneously opened opportunities to draw inferences on previously unanswerable high dimensional questions. In this regime the curse of dimensionality not only denies frequentist methods including maximum likelihood estimation of all their asymptotic advantages, but also often makes these estimates at best misleading if not downright wrong. However, ensemble based Bayesian inferences do not suffer from these afflictions, as they recognize that drawing inferences is an inherently uncertain process and employ the laws of probability to address this uncertainty. This talk will briefly introduce the ideas probabilistic statistical inference using the following example of RNA secondary structure prediction. RNA secondary structures play a crucial role in the function of many RNAs, and structural features are often esemtial to their interaction with other cellular components. But as we show the Boltzmann weighted space of RNA secondary structures can be very complex. Here we present a new algorithm, RGibbs, to identify RNA motifs in longer unaligned sequence, and predict consensus secondary structures for using the blocked Gibbs sampler, which has theoretical advantage in convergence time. This algorithm iteratively samples from the conditional probability distributions P(Structure | Alignment) and P(Alignment | Structure). We illustrate how these probabilistically drawn samples can characterize these potentially complex spaces using hierarchical clustering method to characterize the shape of the posterior space, γ-centroid estimator to generate a prediction from sampled structures, and credibility limits to characterize the uncertainty. An analysis of 17 RNA families shows substantially improved structural prediction based on PPV-SEN curves comparisons, compactness of sampled structures around their ensemble centroids, at least eleven families with well separated clusters. The fact that the distances between the references structures and the centroid structures were large compared to the variation among structures within an ensemble raises questions the aptness of the term maximum expected accuracy estimator.