Constructing and Validating Sequence-Based Predictors

Mark R. Segal
University of California, San Francisco, USA

A wide variety of problem types surrounding the use of sequence data for prediction purposes exist. For instance, derived sequence features at (or near) a genomic location of interest have been used to predict associated phenomena: imprinting and X-linked gene inactivation are among recent examples. In tackling such problems not only do we need to contend with potentially high-dimensional predictors exhibiting complex associations but, even more fundamentally, differing strategies for devising the features themselves. Here we contrast two such approaches: scattershot generation and motif-finding based. Illustration, along with discussion of attendant validation issues makes recourse to recently published data on CpG island methylation.

[Mark Segal, Division of Biostatistics, 185 Berry Street, Lobby 4, Suite 5700, San Francisco, CA 94107; mark@biostat.ucsf.edu]