J Proteomics Bioinform 2018, Volume 11
DOI: 10.4172/0974-276X-C1-113

Genome Informatics: kmerHMM, SNPdryad and SignalSpider

Ka-Chun Wong
City University of Hong Kong, Hong Kong

There are three genome informatics methods are described here. The first kmerHMM is a pattern recognition method for discovering DNA motifs bound by proteins from Protein Binding Microarray (PBM) data. The novelty of kmerHMM lies in two aspects. First, it outperforms the existing methods in using Hidden Markov Models (HMMs) for modeling adjacent nucleotide dependency. Secondly, kmerHMM incorporates N-max-product algorithm and can derive multiple motifs. Comparisons of kmerHMM with other leading methods on several data sets demonstrated its effectiveness and uniqueness. Especially, it achieved the best performance on more than half of the data sets. In addition, the multiple binding modes derived by kmerHMM are biologically meaningful and will be useful in interpreting other genome-wide data. The second method named SNPdryad is a random forest method to predict the deleterious effect of non-synonymous SNPs on human proteins. It only includes protein orthologs in building a multiple sequence alignment. Among many other innovations, SNPdryad uses different conservation scoring schemes and uses Random Forest as a classifier. It has been demonstrated to have better performance than the existing methods (e.g. Harvard PolyPhen2 and JCVI SIFT) on well-studied datasets. It has been run on the complete human proteome, generating deleterious prediction scores for ALL possible non-synonymous SNPs in human. Lastly, the third method named SignalSpider will then be briefly introduced as a probabilistic graphical model for the integrative analysis of multiple ChIP-Seq (next generation sequencing) profiles from the ENCODE consortium. Comparing with similar existing methods, SignalSpider performs better in clustering promoter and enhancer regions. Notably, SignalSpider can learn higher-order combinatorial patterns from multiple ChIP-Seq profiles. The application of SignalSpider on the normalized ChIP-Seq profiles from the ENCODE consortium and learned model instances. We observed different higher-order enrichment and depletion patterns across sets of proteins. Those clustering patterns are supported by Gene Ontology (GO) enrichment, evolutionary conservation and chromatin interaction enrichment, offering biological insights for further focused studies. We also proposed a specific enrichment map visualization method to reveal the genome-wide transcription factor combinatorial patterns from the models built, which extend our existing fine-scale knowledge on gene regulation to a genome-wide level.

kc.w@cityu.edu.hk