Poster H-13

Combined ab initio and comparative analysis of putative cis regulatory modules (CRMs)

Authors:
Nora (*Pierstorff*)
Rodrigo (*Nunes da Fonseca*)
Casey (*Bergman*)
Thomas (*Wiehe*)

Short Abstract: We have developed a method to predict cis-regulatory modules (CRMs) which identifies regions containing perfect local ungapped sequences based on multiple species comparison and overrepresentation of transcription factor binding sites without using prior information. Applied to a dataset of CRMs of Drosophila, our method outperforms all other methods tested.

Long Abstract:
Predicting the cis-regulatory modules (CRMs) in higher eukaryotes is a challenging computational task, since aside from the basic template of multiple binding sites for multiple transcription factors, the architecture of CRMs can vary dramatically. Commonly used methods to
predict CRMs based on the signal of transcription factor binding site (TFBS) clustering are limited by prior information about transcription factor specificity. Thus more general methods that bypass the reliance on TFBS models are needed for comprehensive CRM prediction.
We have developed a method to predict CRMs called CisPlusFinder which identifies high density regions of perfect local ungapped sequences (PLUSs) based on multiple species conservation between closely related species. By assuming that PLUSs contain core TFBS motifs that are locally overrepresented even in flanking nonconserved sequences, our method attempts to capture the expected features of CRM structure and evolution using no prior information. Applied to the HexDiff benchmark dataset of CRMs involved in anterior-posterior patterning in Drosophila, CisPlusFinder achieves 92% sensitivity and 61% specificity at the CRM level and outperforms all other methods tested. Using the REDfly database, we find that some "false positive" predictions in the HexDiff dataset correspond to recently annotated CRMs, and that CisPlusFinder can achieve 68% sensitivity and 71% specificity for a set of CRMs involved in general aspects of Drosophila development.
Our results show that methods that use only comparative genomic data to extract the intrinsic information encoded in CRMs can achieve reasonable performance in higher eukaryotes, and may allow reasonably accurate
whole-genome CRM prediction in absence of TFBS motif models.