BioFeatureFinder (BFF): Flexible, unbiased analysis of biological characteristics

Felipe Ciamponi[1], Michael Lovci[1], Katlin Massirer[1]

1 CBMEG - UNICAMP

Abstract

BFF interrogates interesting genomic landmarks (ex. alternatively spliced exons, DNA/RNA-binding protein binding sites, and gene/transcript functional elements) to identify distinguishing biological features (nucleotide content, conservation, k-mers, secondary structure, protein binding sites and others). BFF uses a flexible underlying model, combining classical statistical tests with big data machine learning strategies, that takes thousands of biological characteristics (features) and can interpret category labels in genomic ranges or numerical scales from genome graphs. The algorithm is python-based with scalable multi-thread capabilities, designed to be compatible with a wide array of servers ranging from notebooks to HPC clusters. Due to flexible nature of it's design, BFF can also be easily modified to include new functions and sources of data in it's analysis process. As proof-of-concept, we applied BFF to an eCLIP-seq (enhanced crosslinking-immunoprecipitation followed by RNA-seq) dataset for the mRNA targets of RNA-binding proteins (RPBs) RBFOX2. Our algorithm was capable of recovering several major features described previously in the literature, as the GCAUG binding motif for the RBFOX2 protein. To showcase the potential uses of BFF, we analyzed 112 eCLIP-seq datasets from RBPs available at ENCODE, identifying biological features associated with the binding sites for these proteins. From a total of 5498 input features, BFF predicts an average of 624 important features for each RBP. 98 RBP binding maps that were marked by co-location with other RBP binding maps, with known complexes (ex. IGF2BP1-3, U2AF1-2) being identified by BFF. 40 RBP binding maps were marked by their relative abundance of sequence motifs, with known examples from the literature (ex. TARDBP, SRSF1, PUM2, PTB and QKI) being successfully recovered by BFF. Secondary RNA structure was a distinguishing feature for 64 proteins, some which are known RNA-structure binding proteins (ex. TAF15, KHDRBS1 and ESWR1). Taken together, our results show that BFF provides a flexible and reliable analysis platform for large-scale datasets, while at the same time providing a way to control observer bias and uncover latent relationships in biological datasets.

Funding: FAPESP, CNPq