Poster H-42

Discriminatory features of sequence profiles for transcription factor binding site discovery

Authors:
Gerald Quon (*Department of Biochemistry, University of Toronto*)
Shoshana Wodak (*Structural Biology and Biochemistry, The Hospital for Sick Children*)

Short Abstract: Transcription factor binding site (TFBS) prediction methods typically report a high rate of false-positive motifs. Using novel information content profile-based features, our method discriminates true- and false-positive motifs from yeast with an overall accuracy of 90%. Furthermore, we demonstrate significant improvement in predictions in a set of synthetic mammalian datasets.

Long Abstract:
Computational methods for the de novo discovery of transcription factor binding sites (TFBS) and their related profiles in DNA sequences play a very important role in the analysis of gene regulation. A major shortcoming of most available methods for TFBS discovery is that they produce a high rate of false-positive motifs (FP-TFBS): highly scoring sequence motifs that do not correspond to TFBS. Various ways of incorporating prior knowledge of specific characteristics of TFBS profiles have been explored as a means of guiding predictions towards profiles with such features. For example, groups of TFBS profiles from a particular family of TFs can be amalgamated into a ”familial profile” describing the entire family, and subsequently used as model priors during de novo TFBS identification. Another type of approach, most closely related to the work presented here, explores methods that search for motifs with specific patterns in the position specific information content of their profiles, using both supervised and unsupervised learning methods. Profile pattern-based methods are attractive because their prior knowledge is not family specific, but instead relies on general features of TFBS. For example, TFEM (Kechris et al., 2004) can search for profiles with one or two blocks of conserved positions, generically representing mono- and dimeric interfaces, respectively. Our work builds on and extends this approach by explicitly determining features of TFBS profiles that are useful for characterizing and discriminating them from false positives (FP-TFBS). To that end, using the yeast S. cerevisiae as a model system, we mapped out the entire landscape of FP-TFBS, defined as highly scoring sequence motifs discovered by the MEME prediction software, in groups of genes unlikely to be co-regulated. We show that this landscape can be described by a total of 274 FP-TFBS profiles, each representing a distinct cluster of discovered profiles. We find that some of these profiles have known properties favorable for transcription. Many also have clear positional preference with respect to the translation initiation site. We therefore suspect that several of these motifs may have intrinsic properties similar to the unary and binary tracts in aiding transcription. We computed a set of 2132 profile features, based on a reduced representation of the position specific information content, on the identified compendium of FP-TFBS profiles and a control set of known yeast TFBS. A combined sparse logistic regression and cross validation procedure was then used to test the power of our features to discriminate TFBS from FP-TFBS profiles. Our procedure was shown to yield an excellent performance, with an overall classification rate of 89.89%, sensitivity (fraction of correctly

assigned TFBS motifs) of 82.69%, and specificity (fraction of correctly assigned FP-TFBS) of 92.75%. Our procedure also identified the most salient features capable of discriminating between the TFBS and their FP-TFBS counterparts. We found that only 7 features were identified for TFBS profiles, and 66 for FP-TFBS profiles. The most striking characteristic feature of TFBS profiles was the consistently high variation in conservation of positions in each profile. Surprisingly, we found that some of the features we expected to positively score for TFBS, such as spaced dyads and stretches of highly conserved positions, were either not important, or scored positively for FP-TFBS. Our classifier, trained on yeast data, performed well when used to post-process MEME results on an unrelated, synthetic mammalian dataset. The dataset consisted of intergenic sequences from human of varying length, spiked with known TFBS. We show that while MEME returns more FP-TFBS as the intergenic sequence length increases, our classifier returns few, if at all, FP-TFBS. These results cannot be simply attributed to over-prediction of FP-TFBS, as the sensitivity of our method remains high despite the removal of many FP-TFBS. The application of our classifier to profiles identified by MEME is hence an effective means of reducing the rate of false positive predictions produced by this method and thereby correcting one of its most serious shortcomings. Also, the encouraging results obtained on mammalian sequences with our yeast-trained classifier indicate that to a crude first approximation, many TFBS and FP-TFBS profiles from yeast may share similar properties to their mammalian counterparts.