Detecting cis-regulatory motifs for cooperatively binding proteins

Liesbeth MM van Oeffelen, Bart De Moor and Yves Moreau

Unraveling regulatory pathways is a key step towards understanding of biological processes. A major problem that biologists are often confronted with, is that they want to retrieve new binding sites for a known regulatory protein, while reducing as much as possible the number of costly and time-consuming experiments. Therefore, they construct a PWM usually based on microarray data or a set of known binding sequences, and use this to score the putative promoter region of each gene. Subsequently, the highest scoring genes are validated in the wet lab (e.g., with RT-PCR).

Several methods have been developed to score genes based on a PWM, depending on the interaction between transcription factor and DNA. The first interaction mode studied involved only one protein copy. Afterwards, adaptations have been proposed to account for multiple binding sites within a promoter region, and cooperatively binding proteins. However, these adaptations require additional parameters that cannot be estimated from sequence.

In this work, we describe a method that takes multiple binding sites and cooperative binding into account with a minimum of additional parameters that cannot be determined from sequence. Therefore, we theoretically derive the binding probability within a putative promoter sequence. First we consider the binding probability at a single binding site, then the influence of multiple binding sites and cooperative binding is studied. Finally, as an illustration, we apply our method to the homotypic cooperative binding (i.e., cooperative binding with the same protein) of the Fur protein in *Pseudomonas aeruginosa*, for which binding sequences (to derive a PWM) and microarray data (for validation) are available in the literature.

As only the high scoring differentially expressed genes will be directly Fur-regulated, only those may be used for validation. Therefore, we do not use ROC curves to evaluate our method, but we determine the number of true positives (TP) versus the number false positives (FP) for a limited number of false positives, and evaluate the area under this curve. The higher this area, the better the method.

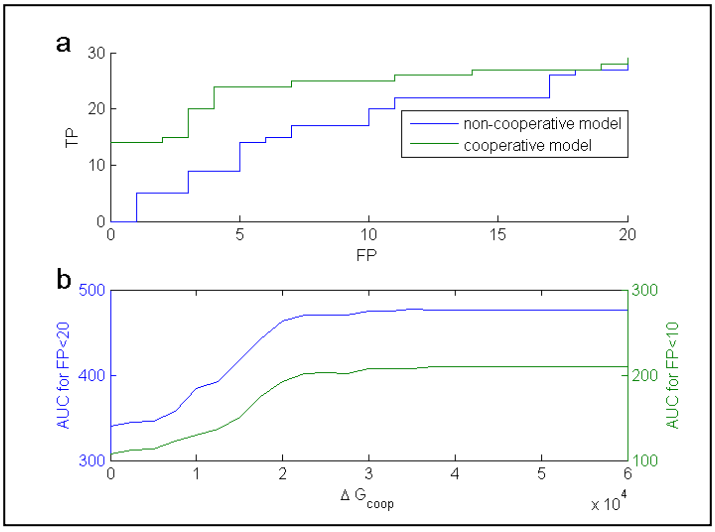

Figure 1: (a) TP versus FP curves for the non-cooperative and the cooperative model with ΔG_{coop}=40kJ. (b) The area under this curve as a function of ΔG_{coop} for FP<10 and FP<20, to make sure that the trend of this curve is not too dependent on the number of FP considered.

In conclusion, we propose a new scoring method that does not require additional parameters to account for multiple binding sites, and only requires the distance(s) between the proteins and the cooperative binding constant(s) to model cooperative interactions. For heterotypic cooperativity, these binding constants only have to be known up to a constant factor.