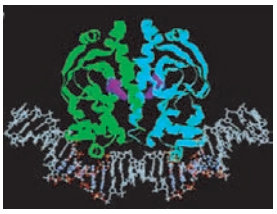


Development of a Computational Pipeline For Automated Prediction of Bacterial Transcription Factor Binding Sites

Christelle Robert(1,2,3), Paul Birch(2), Geoffrey J. Barton(3), Frank Wright(1), Leighton Pritchard(2).

(1) Biomathematics and Statistics Scotland (BioSS).
(2) Scottish Crop Research Institute (SCRI).
(3) School of Life Sciences, University of Dundee.

Abstract

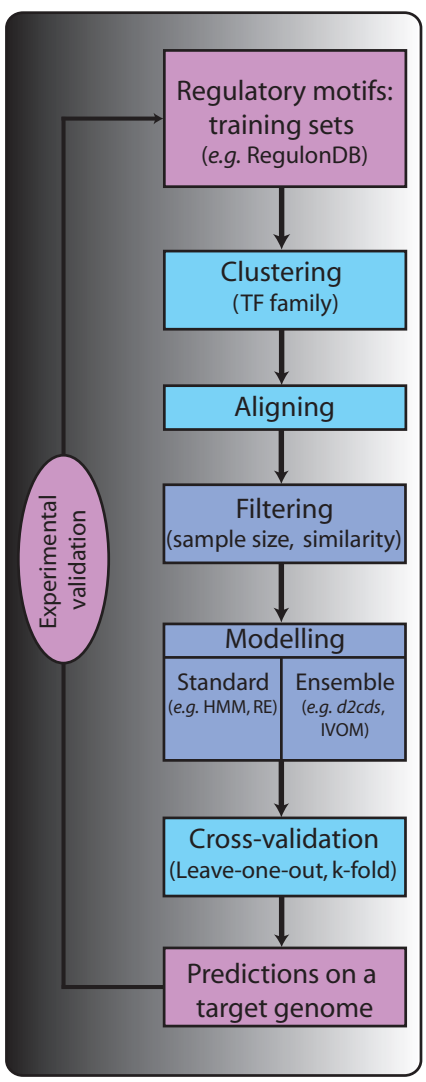
Understanding bacterial gene expression regulation is a major challenge. Using a training set of known transcription factor binding site (TFBS) sequences, we aim to predict the genome locations of previously unknown binding sites in bacterial plant pathogen genomes. Modelling the training set pattern is nontrivial, due to the heterogeneity of sequences to which a typical transcription factor (TF) binds. Here we present a supervised learning based pipeline to identify the locations of regulatory motifs in bacterial genomes. We use *Escherichia coli* K12 as a well-characterised model organism where the locations of most regulatory motifs are already known.

Introduction

A common problem found when predicting regulatory motifs is a high false discovery rate (FDR). We address this issue by evaluating methods to reduce the FDR using biologically-relevant information. Two main biological features are described here: the distance between a regulatory motif and its adjacent downstream gene (referred to as *d2cds*), and base composition of regulatory motifs. The former (*d2cds*) is used to weight alternative model output scores using the probability of observing the predicted *d2cds* distances for a validated set of TFBS [1]. The second method uses an Interpolated Variable Order Motifs (IVOM) approach [2]. We adapt this approach for use with shorter motifs, weighting base compositions of mono-, di-, and tri-nucleotides.

Analysis Pipeline

The pipeline shown below represents schematically the iterative process of model refinement. This pipeline is implemented in a set of Python modules, incorporating cross-validation and IVOM implementations.



Refining Training Sets Improves Model Performance

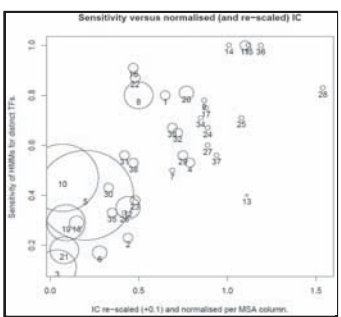


Figure 1: Sensitivity of HMM-only prediction for distinct TFBS (TF are indicated with numbers) versus the information content normalised per multiple sequence alignment (MSA) column. Bullet sizes reflect training set sizes.

Prediction sensitivity (*Sn*) increases with information content (IC) of the alignment used to generate the models (figure 1). Sequence heterogeneity leads to models with low IC and resulting poor predictors (*Sn* <0.5). The performance of models may be improved by restricting the TFBS training set on the basis of sequence identity. An example for CRP binding sites is shown in figure 2. By restricting the training set to CRP TFBS sharing a minimum of 90% identity, the *Sn* increases to 0.77, from *Sn*=0.4, when no restriction is in place.

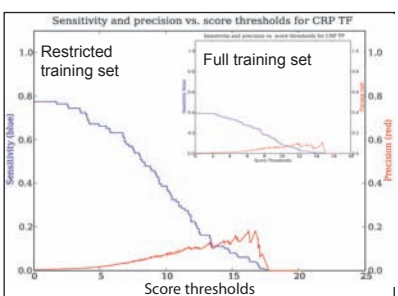


Figure 2: Sensitivity and precision for HMMs based on TFBS of CRP TF, with and without restricting the training set based on sequence identity.

TFBS-CDS Distance (*d2cds*)

The distribution of distances between known binding sites and their downstream genes is used to weight the scores from alternative models (figure 3). The model score and the probability of the associated *d2cds*, are combined using a joint probability as shown in equation 1.

$$P[s \in \mathcal{S} \text{ and } d_s \in (x-a, x+a)] = P(s \in \mathcal{S}) P[d_s \in (x-a, x+a) | s \in \mathcal{S}].$$

Equation 1: Joint probability that sequence *s* belongs to the set of promoters ($\mathcal{S}$) and that *s* lies at a distance, d_S, from its adjacent downstream gene, where d_S is in the interval $(x-a, x+a)$.

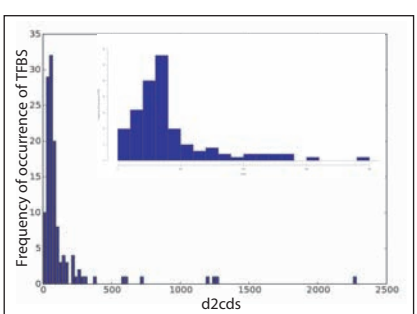


Figure 3: Observed frequency of occurrence of TFBS by distance to the adjacent downstream gene (*d2cds*).

Interpolated Variable Order Motifs (IVOM)

The IVOM approach [2] is used to distinguish between true and false positive predictions on the basis of the divergence of weighted mono- di- and tri-nucleotide compositions from compositions observed in a reference sequence set. Figure 4 shows the distribution of IVOM-based entropy distance measures for a selection of distinct TFBS in *E.coli*, compared to the reference set of all TFBS in *E.coli*. Clearly, we expect the scores for real TFBS to be close to zero. Figure 5 shows the modulus entropy score separation between TFBS (*i.e.* promoter) and both coding (CDS) / intergenic (IG) regions. ROC plots (figure 6) show IVOM-based model performance much better than random when comparing scores observed for TFBS to CDS and IG regions of same lengths (60-65 nt).

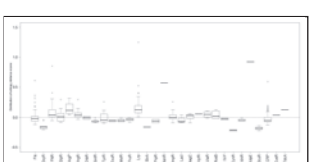


Figure 4: Boxplots of the distribution of IVOM-based entropy scores for a selection of distinct *E.coli* TFBS.

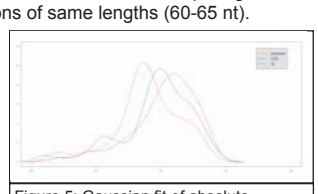


Figure 5: Gaussian fit of absolute frequency of promoter, CDS and IG regions vs. |scores| (sequence length is restricted to [60-65] nucleotides).

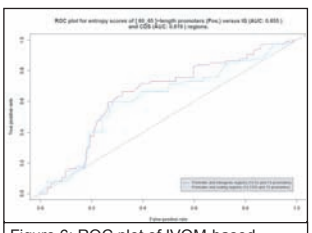


Figure 6: ROC plot of IVOM-based entropy scores for promoter and coding regions, and promoter and intergenic regions (bootstrap 1000 times).

References

[1] Burden, S. *et al.* Bioinformatics; 21(5); 601-607; 2005.
[2] Vernikos, GS. *et al.* Bioinformatics; 22(18); 2196-2203; 2006.