Prediction of transcription factor binding sites in bacterial genomes: case study with *Pectobacterium atrosepticum*

Christelle Robert[1,2,3], Paul Birch[2], Ian Toth[2], Michael Ravensdale[2], Lucy Moleleki[2], Hui Liu[2], Sonia Humphris[2], Geoffrey J. Barton[3], Frank Wright[1], Leighton Pritchard[2].

(1) Biomathematics and Statistics Scotland (BioSS).
(2) Scottish Crop Research Institute (SCRI), Invergowrie, Dundee DD2 5DA.
(3) School of Life Sciences, University of Dundee.

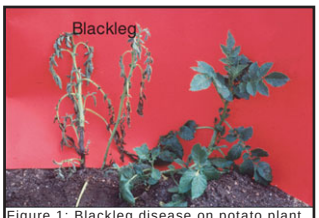


Figure 1: Blackleg disease on potato plant (left) vs. healthy plant (right).

Abstract

Understanding bacterial gene expression regulation poses a major challenge. We are interested in the phytopathogen *Pectobacterium atrosepticum* (*Pba*) which causes disease in potatoes (Fig.1). Using a training set of known transcription factor (TF) binding site sequences, we aim to predict the genome locations of previously unknown binding sites. Modelling the training set pattern is nontrivial, due to the heterogeneity of sequences to which a typical TF binds. Here we model *hrp* (hypersensitive response and pathogenicity) box sites which bind to the HrpL TF. In order to predict the locations of *hrp* boxes in the *Pba* genome, we use a simple modelling method based on regular expressions and a statistical method based on hidden Markov models (HMMs). These *hrp* box models are then used to search intergenic regions of *Pba*. Predicted binding sites exhibit a biased distribution towards the horizontally acquired islands (HAIs) of the genome and are shown to lie upstream of genes downregulated in a HrpL$^-$ mutant.

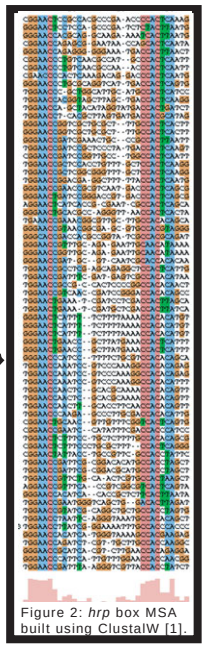


Figure 2: *hrp* box MSA built using ClustalW [1].

Modelling methods

- Building of multiple sequence alignment (MSA) using ClustalW (fig.2).
- Generation of models with regular expressions (right) and HMM (using HMMER package [2]). The models are derived from the MSA (fig.2).
- Model validation with 10-fold cross-validation method using test sets made up of 2000 random sequences, which were generated using either single nucleotide composition, or di-nucleotide composition of known promoter regions. A diagram of the architecture of the *hrp* box HMM profile is illustrated in figure 3.

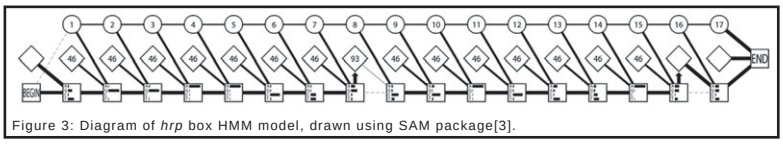


Figure 3: Diagram of *hrp* box HMM model, drawn using SAM package[3].

GGAAC[16-21]AC (RE1)
GGAAC[15-20]CAC (RE2)
GGAAC[16-21]AC..A (RE3)
GGAAC[15-20]CAC..A (RE4)

Predicted HrpL-regulated genes

GENE ID	HMM SCORE	DISTANCE	GENE NAME	PRODUCT
ECA2086	22.3	62	hrpJ	type III secretion protein
ECA2098	20.21	58	hrpF	type III secretion protein
ECA2113	19.23	33	dspE	putative avirulence protein
ECA2112	16.19	56	hrpW	type III effector protein
ECA2093	14.79	73	hrpA	type III secretion protein
ECA2103	14.67	78	hrpN	harpin
ECA1841	11.64	148	-	conserved hypothetical protein
ECA3987	11.3	136	metH	5-methyltetrahydrofolate-homocysteine methyltransferase
ECA3599	9.04	89	-	conserved hypothetical protein
ECA3043	9.03	195	-	putative phosphoesterase
ECA0804	8.07	73	rhiE	rhamnogalacturonate lyase
ECA2407	7.62	118	kdgM	oligogalacturonate-specific porin
ECA2713	6.95	602	-	LysR-family transcriptional regulator
ECA2062	**6.88**	**386**	**-**	**putative phosphatase**

Table 1: Selection of predicted HrpL-regulated genes (IDs, names, products) with HMM model. The selected putative *hrp* boxes exhibit the highest scores (column 'HMM score') amongst the predictions. The 'distance' refers to the distance between the binding site and the downstream gene.

Predicted HrpL-regulated genes are depicted in table 1 for HMM based results and table 2 for regular expression predictions. All predictions are found in intergenic regions in the same orientation as their downstream genes. The HMM predictions all have scores above zero. The hits with the highest scores correspond to the well-characterized *hrp* genes (red rows in table 1). These hits are also retrieved with regular expressions. *hrp* box predictions also highlight new candidate genes of interest, some of which have been experimentally validated to be HrpL-regulated genes.

GENE ID	SYNONYM	PRODUCT	RE1	RE2	RE3	RE4	DISTANCE
ECA2058	-	probable short-chain dehydrogenase	X				272
ECA2062	-	putative phosphatase	X		X		391
ECA2086	hrpJ	type III secretion protein	X	X	X	X	65
ECA2093	hrpA	type III secretion protein	X		X		78
ECA2098	hrpF	type III secretion protein	X	X	X	X	63
ECA2103	hrpN	harpin	X	X	X	X	83
ECA2104	-	VgrG protein	X	X			255
ECA2112	hrpW	type III effector protein	X	X	X	X	61
ECA2113	dspE	putative avirulence protein	X	X	X	X	38
ECA2150	-	putative membrane protein	X		X		167

Table 2: Predicted HrpL-regulated genes with four regular expressions (RE1-4). A cross indicates that a *hrp* box has been identified by the regular expression in the corresponding column, *e.g.* RE1, RE2, RE3, RE4. The distance refers to the distance between the last nucleotide of the *hrp* box and the first of the downstream gene.

Visualisation of predictions on *Pba*

The *hrp* boxes predicted are represented on *Pba* genome (fig.4) using GenomeDiagram [3]. Each circle represents the *Pba* genome labeled with predictions from distinct methods (labels are in violet for RE2, in green for RE3, in red for HMM). The outer circle indicates the horizontally acquired islands (HAIs) known so far on *Pba*.

Comparisons between hits found in HAI regions versus non-HAI regions with regular expressions show a bias towards HAI regions (Table 3). This bias is statistically significant (P<0.05) with both regular expressions tested, but not with the HMM.

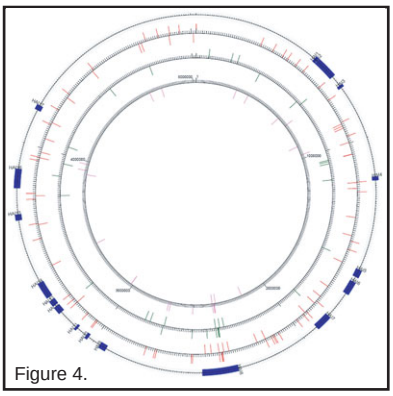


Figure 4.

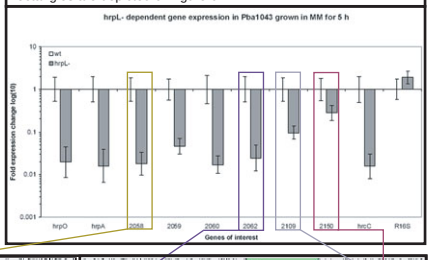
Method	Number of hits in HAI regions (versus expected)	Number of hits in non-HAI regions (versus expected)	Observed Chi-square	Expected Chi-square (for a given P-value)	Statistically Significant result
HMM	16 (13.6)	72 (74.4)	0.51	3.84 (0.05)	No
RE1	7 (3.4)	15 (18.6)	4.51	3.84 (0.05)	Yes
RE2	13 (4.9)	19 (27.06)	15.55	10.83 (0.001)	Yes

Table3: Chi-square results between number of hits in HAI versus non-HAI regions.

Experimental validation

A selection of the predicted hits is shown on figure 5 and 6. Figure 5 shows the QRT-PCR experiments in a HrpL$^-$ mutant. All specified genes are downregulated except the control (R16S). These genes of interest are located downstream of predicted *hrp* boxes with either or both of the methods (*cf.* figure 6).

Figure 5: QRT-PCR expression level of some selected genes in a HrpL$^-$ mutant. The genes which expression levels are highlighted in rectangles are depicted on figure 6.



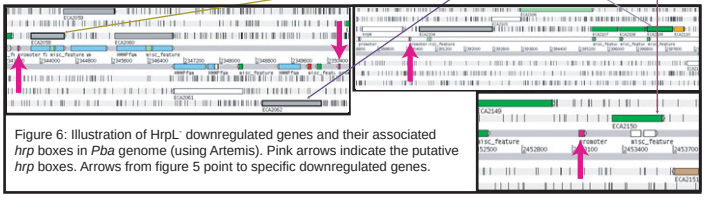


Figure 6: Illustration of HrpL$^-$ downregulated genes and their associated *hrp* boxes in *Pba* genome (using Artemis). Pink arrows indicate the putative *hrp* boxes. Arrows from figure 5 point to specific downregulated genes.

Conclusion

The predicted *hrp* boxes include all expected *hrp* boxes upstream of well-characterized *hrp* genes, which gives confidence in the existing models to represent members of the *hrp* box family. Some candidate genes downstream of the putative *hrp* boxes are found to be downregulated in a HrpL$^-$ mutant . Future work will investigate the combination of biological features into the models and generalize the predictor for all promoter binding sites on enterobacterial species.

References

[1] Thompson, J.D. *et al.* Nucleic Acids Res.; 22(22): 4673-4680; 1994. [2] Eddy, S.R. Bioinformatics; 14: 755-763; 1998. [3] Pritchard, L. *et al.* Bioinformatics; 22(5): 616-617; 2006.
[4] Hughey, R. *et al.* Compt. Appl. Biosci.; 12(2): 95-107; 1996.