

An Exceptional Motif of Infection-Induced *P. infestans* Promoters

W. Morgan; The College of Wooster, Wooster, Ohio, USA

Infection-Induced Promoters

Transcriptome analysis has revealed that dozens of *P. infestans* genes are specifically transcribed early during the infection process (Randall et al., 2005; Haas et al., 2009.). However, the transcription factors and DNA control elements required to regulate the expression of these genes is presently unclear.

As a first step to identify the DNA control elements of infection-induced genes, I have used newly developed computational tools to identify potential DNA control elements.

Input sequences

- Complete set of 22657 *P. infestans* upstream sequences (pi_utr_upstream_1000.fasta) was downloaded from the Broad Institute (www.broadinstitute.org)

- Subset of 436 putative promoter sequences of genes *induced* >2-fold on day 2 of infection (Haas et al., 2009)

- Subset of 113 putative promoter sequences of genes *repressed* >2-fold on day 2 of infection (Haas et al., 2009)

Identification of Exceptional 8-mers

Goal: Discover “words” that are exceptionally abundant in infection-induced promoters

Computational procedures

- Enumerated all 8-mers in putative promoter sets using WordSeeker (Lichtenberg et al., 2009) and MoSDi (Marschall and Rahmann, 2009)
- Computed p-values, using MoSDi and custom R scripts, based on occurrence counts in infection-induced promoters using a compound Poisson model, which can account for non-independence of word occurrences (Robin et al., 2005)

Main findings

- Found 16 exceptional words with p-value <0.2E-07 (Bonferroni correction for FWE rate of 0.01 with 50K+ words; Table 1)
- After visual inspection revealed potential overlaps, 15 of these 16 words were manually clustered into 2 related motifs (Table 1)
- Both motifs exceptionally abundant in infection-induced promoters (Table 1), but not infection-repressed promoters (Table 2)

Table 1. Manual clustering of exceptional words in infection-induced putative promoters

Word/Motif*	Observed#	Expected#	p-value
TGTACATG	60	25.5	6.51E-09
AGTACATG	50	20.9	4.69E-08
GTACATGT	101	42	2.06E-14
GTACCCGGT	57	20.1	1.57E-11
TACCCGTA	112	43.6	4.38E-18
ACCCGTAC	67	19.6	6.84E-17
ACATGTAC	96	45.1	4.93E-11
CATGTACA	61	27.6	4.05E-08
ATGTACAT	39	14.2	6.26E-08
-GTACMKGTACA-	70	19.1	1.78E-18
-GTACW-GTACA-	80	24.5	5.51E-18
GTACA-GTA	76	28.7	2.94E-13
GTAC-TGTA	66	30.2	1.40E-08
TAC-TGTAC	66	28.6	2.00E-09
TACA-GTAC	83	30.4	5.40E-15
ACA-GTACA	44	16.9	4.49E-08
TATTAATA	80	40.8	3.21E-08

* Consensus motifs in bold.

Thanks to the Ohio University Bioinformatics group for their support during my sabbatical.

Exceptionality of Degenerate Motifs

Goal: Examine exceptionality of consensus and degenerate motifs in infection-induced promoters

Computational procedures

- Computed p-values of both motifs allowing 1 mismatch using MoSDi
- Created logo at weblogo.berkeley.edu (Crooks et al., 2004)

Main findings

- Both degenerate motifs are exceptionally over-represented (Table 3)
- Altogether, over half (229/436) of infection-induced promoters contain at least one copy of degenerate motif (Figure 1)

Table 2. Non-exceptionality of both motifs in infection-repressed promoters

Motif (H=1)*	Observed #	Expected #	p-value
GTACMKGTAC	4	4.2	6.10E-01
GTACWGTACA	3	5.4	9.04E-01

Table 3. Exceptionality of degenerate motifs in infection-induced promoters

Motif (H=1)*	Observed #	Expected #	p-value
GTACMKGTAC	298	129.9	1.04E-23
GTACWGTACA	320	165.3	2.96E-16

* Degenerate motif allowing 1 mismatch (Hamming distance = 1)

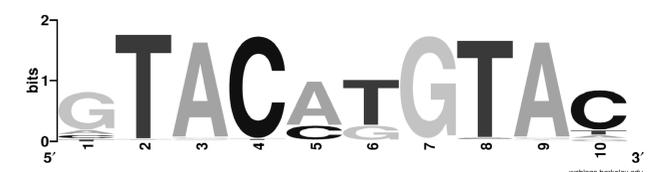


Figure 1. Logo of degenerate motif

Literature Cited

- Crooks et al. 2004. *Genome Research* 14:1188
- Haas et al., 2009. *Nature* 461: 393.
- Lichtenberg et al., 2009. *BMC Genomics* 10: 463.
- Marschall & Rahmann, 2009. *Bioinformatics* 25: i356.
- Randall et al., 2005. *MPMI* 18: 229.
- Robin et al., 2005. *DNA, Words and Models*.