

Prediction of Co-Regulated Genes in Eubacterial Genomes by Phylogenetic Footprinting

Yuko Makita^{1,3} Goro Terai^{2,3} Shigeki Mitaku¹
Makita@ims.u-tokyo.ac.jp terai@ims.u-tokyo.ac.jp mitaku@cc.tuat.ac.jp

Toshihisa Takagi³ Kenta Nakai³
takagi@ims.u-tokyo.ac.jp knakai@ims.u-tokyo.ac.jp

¹ Tokyo University of Agriculture and Technology, 2-24-16 Nakacho, Koganei, Tokyo 184-8588, Japan

² INTEC Web and Genome Informatics Corp., 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan

³ Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: regulon, phylogenetic footprinting, comparative genomics, *cis*-element.

1 Introduction

Since co-regulated genes are recognized by the same transcription factor, they must have the same motif in their upstream sequences. Thus, in principle, we can expect to understand the co-regulation relationship of genes by finding their common motif. However, it is generally difficult to find unknown motifs because they are usually short and are not strictly conserved. One way to reduce unavoidable noises is to use the evolutionary information. Namely, we can postulate that most binding sites of transcription factors are phylogenetically conserved (the method is often called the phylogenetic footprinting). In a previous work, we used such an approach to three closely-related species of *Bacillus* genus and could successfully identify many plausible *cis*-elements and co-regulated genes (regulons) [1]. A preliminary work to apply the same approach into the analysis of another genera was reported last year by us. In this work, we report more comprehensive results on *Mycoplasma*, *Chlamydia* and *Mycobacterium* genera.

2 Method

First, we defined orthologous genes with the bi-direction best-hit method and aligned their upstream 300bp sequences by Smith-Waterman's method. Next we extracted locally conserved regions as candidates of binding sites. We call these regions PCEs (Phylogenetically Conserved Elements). After that, we predicted the sets of potentially co-regulated genes by clustering these PCEs based on their local similarities.

To assess the signal/noise ratio of our result, we applied the same method into randomly-selected upstream sequences and decided the threshold value for the clustering based on that result. Furthermore, we compared the detected PCEs with known transcription factor binding sites.

3 Results and Discussion

With this research, we could obtain some biologically plausible clusters. For instance, a cluster consisting of inverted repeat sequences is tightly coupled with heat-shock related proteins in *Mycoplasma*. In

Table 1: Comparison of detected PCEs with known transcription binding sites.

Binding factor	Promoter	Location	Binding sequence	Orthologs*	Result**	Description
LexA	recA	-121	GAACaggtGTTC	BL	PCE	SOS box
LexA	lexA	-103	GAACacatGTTt	BL	PCE	SOS box
LexA	lexA	-279	GctCgctGTTC	BL	PCE	SOS box
LexA	ruvA	-362	GAACgggtGTTC	BL	CD	SOS box
LexA	ruvC	-35	GAACgattGTTC	BL	PCE	SOS box
LexA	dnaB	-42	GAAtatgcGTTC	BL	CD	SOS box
OxyR	oxyR	-	cTTATCggcnnngccGATAAg	B	-	Oxidative stress
OxyR	ahpC	-96	cTTATCggcnnngccGATAAg	B	-	Oxidative stress
IdeR	Rv1519	-	CTAGGTTAGGCTAGCCTTA	B	-	Iron Dependent
IdeR	Rv3403c	-	TTATGTTAGGCTTCCCTTA	B	-	Iron Dependent
IdeR	Rv3839	-	TTAACTTAGGCTTACCTAA	B	-	Iron Dependent
IdeR	Rv3403c	-	TTATGTTAGGCTTCCCTTA	B	-	Iron Dependent
IdeR	Rv1343c	-	TTAGGCTAGGCTAGCCTTG	BL	CD	Iron Dependent
IdeR	acpP	-	CTAGGCTAGGCTTGCCTAA	B	-	Iron Dependent
IdeR	Rv1347c	-	TTTGCTTGGCTAACCTAA	B	-	Iron Dependent
IdeR	Rv1348	-	TTCGGTTAGGCTACCCTTA	B	-	Iron Dependent
IdeR	hisE	-	ATAGGTTAGGCTACCCTAG	BL	CD	Iron Dependent
IdeR	irg2	-	CTAGGGTAGGCTAACCTAT	B	-	Iron Dependent
IdeR	Rv3402c	-	AGAGGTTAGGCTAACCTCA	B	-	Iron Dependent
IdeR	bfrA	-207	TTAGTGGAGTCTAACCTAA	BL	PCE	Iron Dependent
IdeR	mbtI	-	GTAGGTTAGGCTACATTTA	B	-	Iron Dependent
IdeR	bfrB	-	CTAGGATAGGCTATCCTGA	B	-	Iron Dependent
IdeR	Rv282	-	TTAGCTTATGCAATGCTAA	BL	CD	Iron Dependent
IdeR	mmpS4	-93	TTAGGCTAGGCTAAGTTGC	BL	PCE	Iron Dependent
IdeR	bfd	-	TTAGGCTAGACTCCACTAA	B	-	Iron Dependent
IdeR	bfrA	-230	TTAGTGGAGTCTAGCCTAA	BL	PCE	Iron Dependent
IdeR	mbtA	-	TTAGCACAGGCTGCCCTAA	B	-	Iron Dependent
IdeR	mbtB	-	TTAGGGCAGGCTGTCCTAA	B	-	Iron Dependent
IdeR	Rv766c	-	TGATCTTGGGCGGAGCTAA	B	-	Iron Dependent

* Name of species having the orthologous gene with the *Mycobacterium tuberculosis* gene. [B]*M.bovis* [L]*M.leprae*

**Result: [PCE]the binding site was a part of PCE. [CD] the binding site was within the Coding Region.

Table 1, a partial comparison result between PCEs and known binding sites in *Mycobacterium* genus is presented. As can be seen from the table, our method could extract all known sites as PCEs in case that the known sites are in the non-coding region and that we could define the orthologous genes over three species. Although we could not detect all of the known regulons with our method, we concluded that we could present some plausible regulons that are worthy for future experimental verification. Because our current knowledge on gene regulatory network of these general is quite limited, such an approach should be further investigated.

References

- [1] Terai, G., Takagi, T., and Nakai, K., Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species, *Genome Biology*, 2(11):research0048.1–0048.12, 2001.