Genome-wide DNA-binding specificity of PIL5, an *Arabidopsis* basic Helix-Loop-Helix (bHLH) transcription factor

Hyojin Kang

Department of Bio and Brain Engineering,
KAIST, Daejeon 305-701, Korea
E-mail: hjkang@biosoft.kaist.ac.kr

Eunkyoo Oh and Giltsu Choi

Department of Biological Sciences,
KAIST, Daejeon 305-701, Korea
E-mail: roped@hanmail.net
E-mail: gchoi@kaist.ac.kr

Doheon Lee*

Department of Bio and Brain Engineering,
KAIST, Daejeon 305-701, Korea
E-mail: dhlee@biosoft.kaist.ac.kr
*Corresponding author

Abstract: PIL5 is a member of the basic Helix-Loop-Helix (bHLH) transcription factor superfamily. We previously showed that PIL5 binds to the G-box (CACGTG) motif with high affinity. However, since there are many randomly matched G-box motifs throughout the genome, other factors must account for the in-vivo PIL5 binding specificity. In this study, we investigated if in-vivo PIL5 binding sites can be explained by any other attributes extracted from various sources. Our results showed that PIL5 binding sites can be explained by attributes such as neighbouring motif composition, nucleosome density, DNA methylation and distance from transcription start site in addition to G-box.

Keywords: bHLH; basic helix-loop-helix; phytochrome interacting factor 3-LIKE5 (PIL5); g-box; ChIP-chip; DNA-binding specificity; *Arabidopsis*.

Reference to this paper should be made as follows: Kang, H., Oh, E., Choi, G. and Lee, D. (2010) 'Genome-wide DNA-binding specificity of PIL5, an *Arabidopsis* basic Helix-Loop-Helix (bHLH) transcription factor', *Int. J. Data Mining and Bioinformatics*, Vol. 4, No. 5, pp.588–599.

Biographical notes: Hyojin Kang is a PhD candidate in the Department of Bio and Brain Engineering at the KAIST, Korea. His research focuses on the integration of heterogeneous biological data to reconstruct gene regulatory network.

Eunkyoo Oh received the PhD Degree in Biology from the KAIST, in 2008. His research interests focus on light signaling in plant development.

Giltsu Choi received the PhD Degree in Biology from the University of Brandeis, Waltham, in 1997. Currently, he is an Associate Professor in the Department of Biological Sciences at the KAIST. His research interests include light, phytochrome, and plant development.

Doheon Lee received the PhD Degree in Computer Science from the KAIST, in 1995. He is a Professor in the Department of Bio and Brain Engineering at the KAIST. His research interests include systems bioinformatics and neural engineering.

1 Introduction

The bHLH proteins form a large superfamily of transcriptional regulators and play crucial roles in diverse biological processes. The bHLH family contains a highly conserved domain consisting of two functionally distinct regions: a basic region and an HLH region. The basic region binds to DNA, whereas the HLH region serves as a dimerisation motif (Jones, 2004). The interaction between the HLH regions of two bHLH proteins leads to the formation of a homodimeric or a heterodimeric complex, which recognises and binds to a core hexanucleotide DNA sequence (Shimizu et al., 1997).

Many bHLH proteins bind to the E-box element (CANNTG) or the G-box element (GACGTG). The first bHLH motif was identified in two murine transcription factors known as E12 and E47 (Murre et al., 1989). Since then, tremendous bHLH proteins have been identified in animals, plants and fungi and classified into six main groups based on their evolutionary sequence composition and their DNA-binding specificities (Atchley and Fitch, 1997; Ledent and Vervoort, 2001). In *Arabidopsis*, 147 bHLH protein-encoding genes were identified from the analysis of multiple sequence alignments. Sequence analysis suggested that most of these proteins would recognise E-box, and more specifically, 89 proteins (60% of the total number) would bind to the G-box (Toledo-Ortiz et al., 2003).

PIL5 is a bHLH transcription factor that inhibits seed germination. The in-vitro binding assay indicated that PIL5 binds to the G-box. However, the ChIP analysis indicated that PIL5 binds not to all G-boxes but only to some G-boxes. This raised a question how a bHLH transcription factor such as PIL5 binds to specific G-boxes. Since different G-boxes are present in the context of different DNA sequences, it has been postulated that other attributes such as flanking sequences may contribute to the binding specificity. Genome-wide analysis of PIL5 binding sites further indicated that PIL5 binds to only a fraction of G-boxes in *Arabidopsis* genome. We previously conducted chromatin immunoprecipitation coupled with microarray (ChIP-chip) assay to identify genome-wide binding sites of PIL5 (Oh et al., 2009). Through the analysis, we identified total 748 PIL5 binding sites. Although G-box was the most significant target binding motif, only 58.5% (438/748) of PIL5 binding sites had at least one G-box in their extended 500bp sequences, indicating that other factors must account for the in-vivo PIL5 binding specificity.

There have been many studies to identify other attributes involved in DNA-binding specificity. The flanking sequences outside of the hexanucleotide core motif have been shown to play a role in binding specificity (Atchley et al., 1999), and the loop residues in bHLH proteins known to be involved in DNA binding through flanking sequences

(Nair and Burley, 2000). The DNA-binding affinity of transcription factors can also be affected by chromatin structure. Recent studies showed that DNA methylation and histone modification play important roles in regulating chromatin structure (Zhang et al., 2006, 2007). However, genome-wide binding specificity analysis for bHLH proteins has not been reported yet.

In this study, we investigated which attributes contribute to the in-vivo PIL5 DNA-binding sites. First, we defined PIL5 binding sites and non-binding sites from our previous ChIP-chip assay result and compared them to see what aspects were different. Then, we extracted various attributes from public databases as well as the result of our previous ChIP-chip assay. Finally, we verified the significance of these attributes by using two popular machine-learning methods, Random Forest and SVM. The Random Forest and SVM classifier could classify PIL5 binding sites from non-binding sites with accuracy of 93.05% and 92.31%, respectively. We found that other attributes such as neighbouring motif composition, nucleosome density, G-box presence, DNA methylation and distance from Transcription Start Site (TSS) were involved in PIL5 DNA-binding specificity.

2 Results and discussion

2.1 Characterisation of the PIL5 binding sites and non-binding sites

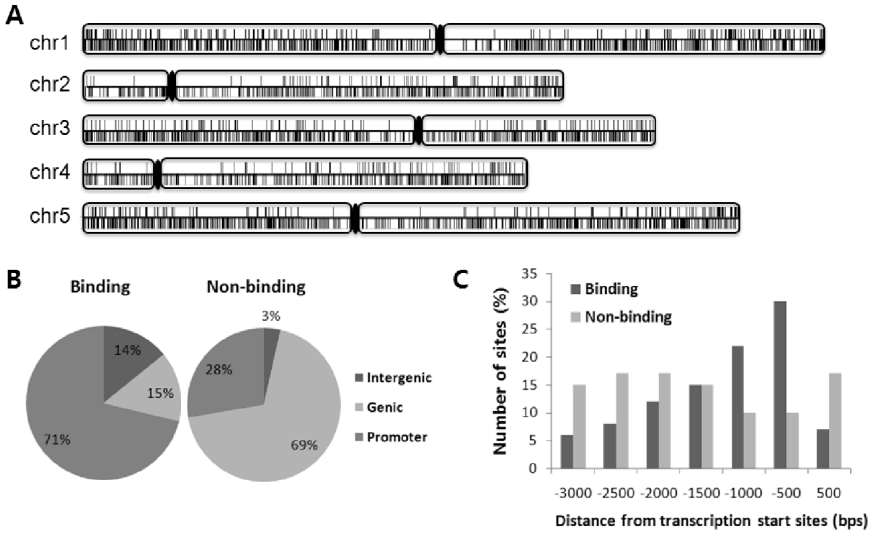
In our previous ChIP-chip assay (Oh et al., 2009), we identified 748 PIL5 binding sites using Tamalpais peak-calling algorithm (Bieda et al., 2006) with slight modification. In brief, a minimum of six consecutive probes in the top 1% of all probes on the array were used as threshold for identifying PIL5 binding sites. We adopted these 748 PIL5 binding sites as true binding data set. In similar way, PIL5 non-binding sites were identified from the result of ChIP-chip array with a minimum of six consecutive probes in low 10% of all probes. As a result, total 3678 non-binding sites were identified and assigned as non-binding data set.

We compared genomic characteristics between PIL5 binding sites and non-binding sites to get insight what make differences. First, we compared chromosomal distribution. PIL5 binding sites and non-binding sites are uniformly distributed throughout the five chromosomes. In contrast to the binding sites that are largely absent in the centromeric regions, the non-binding sites are evenly distributed throughout the chromosome (Figure 1(A)). These results imply that chromosomal structures influence the in-vivo PIL5 binding sites.

Second, we compared genomic positions with regard to promoter regions. The promoter regions are defined as upstream 3000 bps and downstream 500 bps from the transcription start sites. While most of the binding sites (71%) are located in promoter region, majority of non-binding sites (69%) are located in the genic regions (Figure 1(B)). Considering the fact that 49.7% of the *Arabidopsis* genome consists of genic region, it seems that non-binding sites are uniformly distributed along the chromosome. In addition, we compared the distances of the binding sites or non-binding sites that are present in the promoter region from transcription start sites. While the locations of the PIL5 binding sites are further skewed towards immediate upstream from the transcription start sites, non-binding sites were uniformly distributed (Figure 1(C)).

These results are coherent with the fact that PIL5 is a key transcription regulator in seed germination process (Oh et al., 2006).

Figure 1 Comparison between PIL5 binding sites and non-binding sites: (A) comparison between PIL5 binding sites and non-binding sites in the five *Arabidopsis* chromosomes. Upper and lower bars represent the positions of the PIL5 binding sites and non binding sites respectively on each chromosome. Black circles indicate the location of the centromere; (B) comparison between PIL5 binding sites and non-binding sites in the *Arabidopsis* genome and (C) comparison between PIL5 binding sites and non-binding sites in the promoter regions (–3000 to 500 bps)



Third, we compared the presence of G-box because the G-box was the only motif identified by motif-finding programmes (AlignACE (Roth et al., 1998) and MDscan (Liu et al., 2002)) in PIL5 binding sites (Oh et al., 2009). Since the G-box motifs are highly enriched within +250 to −250 bps of PIL5 binding sites (Oh et al., 2009), 500bp extended sequences were searched in both binding sites and non-binding sites. While 58.5% (438/748) of binding sites have at least one G-box, only 3.3% (120/3678) of non-binding sites contain G-box. This is consistent with fact that G-box is the core binding motif for PIL5. However, at the same time, the 120 G-boxes in the non-binding sites indicate that G-box is not a sole determinant of PIL5 binding sites. Overall, these results imply that many attributes including other binding motifs may influence the PIL5 DNA-binding specificity.

2.2 Investigation of possible attributes involved in PIL5 DNA-binding

We extracted possible attributes from public databases as well as the result of our ChIP-chip assay. First, since the G-box was the only motif identified in PIL5 binding sites by motif-finding programmes, the presence of G-box within 500 bp extended window was considered as the first attribute.

Second, since the 41.5% (310/748) of PIL5 binding sites does not contain G-boxes, other motifs may serve as PIL5 DNA-binding sites. We searched over-represented motifs by Oligo-analysis (van Helden et al., 1998) and found that they more likely to be

clustered together around binding sites. In average, binding sites had top 20 ranked over-represented motifs by three times more than non-binding sites. The G-box is one of the well-known ABA-Responsive Elements (ABREs) and ABRE requires a second element called Coupling Element (CE) to form a functional ABA response complex (ABRC) (Shen et al., 1996). The clustered top-ranked motifs may represent the relation between ABREs and CEs. To formulate the neighbouring motif composition within 500 bp extended window as a single attribute, we devised neighbouring motif score as the second attribute (see Materials and Methods).

Third, the distance from transcription start site of nearest gene was considered as attribute. As we could see in Figure 1(C), there was strong correlation between the distance from transcription start site and the number of binding sites. The shorter the distance from transcription start site, the more the binding sites were observed. So, we believe that the binding specificity of PIL5 would rely on whether a binding occurs within a promoter region or not.

Fourth, the presence of CpG islands was considered. In *Arabidopsis*, DNA methylation is found primarily in the CG sequence and a large fraction of methylation is also found in the CNG and CHH (an asymmetric site, where H is A, C or T) sequence (Chan et al., 2005). Putative CpG islands were predicted by the CpG Island Searcher (Takai and Jones, 2002).

Fifth, DNA methylation profile was also applied to reflect in-vivo methylation pattern. The interdependency between DNA methylation and gene transcription has long been the subject of many studies (Kass et al., 1997; Klose and Bird, 2006; Zilberman et al., 2007), so DNA methylation pattern would be involved in PIL5 binding specificity. Currently, genome-wide DNA methylation map was constructed using high-resolution methylcytosine immunoprecipitation (mCIP) method (Zhang et al., 2006).

Sixth, Low Nucleosome Density (LND) score was imported as an attribute. In eukaryotic organisms, promoter accessibility is influenced by chromatin structure and promoter regions are generally more accessible to transcription factors (Sekinger et al., 2005). Since PIL5 functions as a transcription factor, PIL5 would be more likely to bind LND regions.

Seventh, repeat element score was also considered as attribute. Large-scale DNA sequencing has revealed that most of the repetitive DNA is derived from the transposable elements. Recent studies showed that transposable elements were involved in the regulation and rearrangement of genes (Bennetzen, 2000).

Finally, average gene expression level was also considered as an attribute. The ChIP-chip array results showed that PIL5 rarely bound to centromeric regions in chromosomes. The centromeric regions are usually composed of heterochromatin, which cause tight compact of chromatin structure. It is known that the expression of genes in heterochromatin regions is down-regulated. So, if we calculate the average expression level within about 10 kbp window, it could represent whether it is heterochromatin or euchromatin.

To measure the significance of these attributes, we performed statistical comparison between PIL5 binding sites and non-binding sites. However, since the size of non-binding sites was five times larger than that of binding sites, the same sizes of 100 non-binding sets were generated to balance the size. Each non-binding set was composed of 748 non-binding sites, which were randomly chosen from 3678 non-binding sites.

As a statistical evaluation, the Welch two sample *t*-test was performed. Categorical attributes such as G-box presence, CpG islands and repeat regions were coded as 1 and 0

and applied for the *t*-test. The results showed that neighbouring motif score and G-box presence were most significantly different between two classes (Table 1). This is consistent with the fact that DNA-binding motif is a fundamental attribute involved in TF-DNA binding (Kadonaga, 2004). The *p*-values of other attributes such as distance from TSS, methylation score, gene expression level and repeat region were also significant between two classes, suggesting that they also influence the in-vivo PIL5 DNA-binding sites.

Table 1 The result of Welch two sample *t*-test

Attribute name	*p-value*[a]	*sd*[b]
Neighbouring motif score	1.11E-137	5.20E-137
G-box presence	1.47E-124	1.37E-123
Distance from TSS	6.50E-46	6.42E-45
Methylation score	4.14E-24	3.46E-23
Gene expression level	1.72E-22	9.50E-22
Repeat region	7.58E-10	4.88E-9
CpG island	3.83E-3	8.75E-3
LND score	3.89E-2	1.32E-1

[a]*p*-value: Averaged *p*-value of 100 two-sample *t*-test.
[b]sd: Standard deviation of *p*-values.

2.3 Classification of PIL5 binding sites

The purpose of this study is to verify whether other attributes except G-box were also involved in PIL5 DNA-binding specificity. We, therefore, built classifiers and validated accuracies using two popular machine-learning methods, Random Forest (Breiman, 2001) and SVM (Cortes and Vapnik, 1995) implemented in Weka (Witten and Frank, 2005) and BioWeka (Gewehr et al., 2007), respectively. Each classification task was performed 100 times with 1496 instances composed of 748 binding sites and the same number of non-binding sites from one of the 100 non-binding sets. Ten-fold cross-validation method was applied in test procedure.

First, we measured the significance of each attribute by applying one attribute at a time. The result showed that neighbouring motif score was the most significant attribute with accuracy of 83.38% and 84.37% in Random Forest and SVM, respectively (Table 2). This can be explained by the fact that neighbouring motif score was extracted from in-vivo experimental result but other attributes were extracted from indirect sources. The fact that the classification performance of neighbouring motif score was better than the presence of G-box was consistent with our hypotheses that neighbouring other motifs were also involved in PIL5 DNA-binding specificity not only the presence of G-box. Previous studies showed that ABRE requires CEs to form a functional ABRC (Shen et al., 1996) and the ABRE–ABRE pairing is also possible in *Arabidopsis* (Gomez-Porras et al., 2007). Although the precise combinatorial binding mechanism among ABREs and CEs is remained to be determined, the actual binding occurrences seem to be well represented by neighbouring motif score. Among attributes, average expression level, repeat regions and CpG islands showed weak performance with accuracy of lower than

60% in both classifiers. Their actual effect for PIL5 DNA-binding may not be significant or it is possible that current resolutions of these data are not suitable for correct classification. Interestingly, although the *p*-value of LND score was not significant, it showed second best performance in classification. It seemed that computationally predicted CpG islands were not consistent with *in-vivo* methylation pattern.

Table 2 Classification accuracy for each attribute

	Random Forest		*SVM*	
	avg[a]	*sd*[b]	*avg*[a]	*sd*[b]
Neighbouring motif score	0.8338	8.75E-3	0.8437	9.07E-3
LND score	0.7874	1.01E-2	0.8122	9.98E-3
G-box presence	0.7760	3.37E-3	0.7760	3.37E-3
Methylation score	0.6783	1.24E-2	0.7205	9.81E-3
Distance from TSS	0.6520	1.12E-2	0.6703	1.23E-2
Gene expression level	0.5877	1.47E-2	0.5948	1.10E-2
Repeat region	0.5448	5.54E-3	0.5448	5.54E-3
CpG island	0.5368	6.97E-3	0.5368	6.97E-3

[a]avg: Averaged accuracy of 100 classification result.

[b]sd: Standard deviation of accuracies.

Second, when we applied all attributes in the classification, the average accuracies were 93.15% and 92.31% in Random Forest and SVM, respectively (Table 3). Surprisingly, the averaged value of accuracy, precision, recall and *F*-measure was very similar in both classifiers. This result indicated that Random Forest and SVM showed very robust performance. The accuracy of Random Forest was slightly higher than that of SVM. However, when we first applied SVM with default parameters, the accuracy was around 66%. Since it is known that the performance of SVM is comparable with Random Forest (Pang et al., 2006), we try to find optimal parameters for SVM with greedy search. With parameter adjustment, the accuracy of SVM went up to 92.31%.

Table 3 Performance of the classification

	Random Forest		*SVM*	
	avg[a]	*sd*[b]	*avg*[a]	*sd*[b]
Accuracy	0.9315	5.78E-3	0.9231	5.48E-3
Precision	0.9317	5.79E-3	0.9232	5.48E-3
Recall	0.9315	5.78E-3	0.9231	5.48E-3
F-measure	0.9315	5.78E-3	0.9231	5.48E-3
ROC area	0.9730	3.50E-3	0.9231	5.48E-1

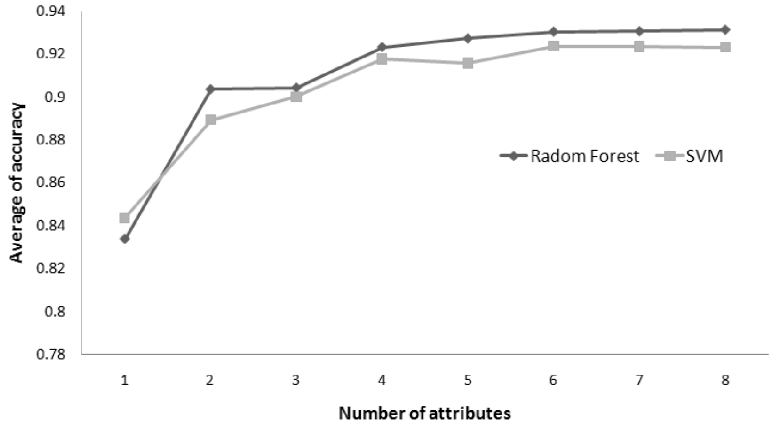
[a]avg: Averaged values of 100 classification result.

[b]sd: Standard deviation of values.

Third, we tried to select minimum number of attributes, which give us satisfactory accuracy. We measured classification accuracies by adding one feature at a time with best

first search algorithm. Figure 2 showed that classification accuracy was saturated over 90% using four features (neighbouring motif score, LND score, G-box presence and methylation score) and remaining features barely contributed to the increase in accuracy. These results indicate that PIL5 DNA-binding can be mainly affected by consensus DNA-binding motifs including G-box, nucleosome density and DNA methylation pattern. Then, we selected optimal subsets of attributes with exhaustive search algorithm. The results showed that in addition to the above four attributes, distance from TSS was also included in optimal subsets and other attributes are not frequently included in optimal subset with 10-fold cross-validation procedure.

Figure 2 Classification accuracy according to the different number of attributes



3 Conclusion

Transcription factors regulate gene expression by recognising and binding to specific DNA motifs. So, their target DNA sequences have been conserved through evolution. To achieve precise specificity required for correct temporal and spatial transcription, the length of DNA-binding motifs should be long enough. However, in many cases, conserved binding motifs are not quite long, so there must be other attributes involved in DNA-binding specificity.

In this study, we analysed genome-wide DNA binding specificity of PIL5, a member of the bHLH transcription factor in *Arabidopsis*. Although our previous study shows that PIL5 binds to the G-box (CACGTG) motif with high affinity, sequence matching can be occurred randomly owing to the short length of G-box. In addition to this, large number of other bHLH proteins also shares the same or similar consensus-binding sequences. For these reasons, we investigated other attributes, which could affect PIL5 DNA-binding specificity.

First, we defined PIL5 binding sites and non-binding sites from ChIP-chip assay result. We adopted 748 PIL5 binding sites from our previous result and newly identified 3678 non-binding sites with a minimum of six consecutive probes in low 10% threshold. Then, we compared genomic characteristics between PIL5 binding sites and non-binding sites to get insight what make differences. We found that the location of binding sites

in promoter and chromosome were involved in PIL5 DNA binding as well as the presence of G-box motif.

Second, we investigated 8 attributes extracted from public databases as well as the result of our previous ChIP-chip assay. They were G-box presence, neighbouring motif score, distance from transcription start site, CpG islands, DNA methylation, nucleosome density, average gene expression and repeat region. The significance of these attributes was evaluated by comparing PIL5 binding sites and non-binding sites. The result of two sample t-test showed that *p*-values of these attributes, especially neighbouring motif score and G-box presence were very significant.

Finally, we further verified the significance of other attributes by using two popular machine-learning methods, Random Forest and SVM. The Random Forest and SVM classifier distinguished PIL5 binding sites from non-binding sites with accuracy of 93.15% and 92.31%, respectively. We found that top-ranked five attributes were optimal subsets for classification through exhaustive search algorithm. In conclusion, we believed that various other attributes such as neighbouring motif score, nucleosome density, G-box presence, DNA methylation and distance from TSS were involved in PIL5 DNA-binding specificity.

4 Materials and methods

4.1 Neighbouring motif score

Total 4096 hexanucleotides can be enumerated but only 2080 hexanucleotides are distinctive if reverse complement motifs are considered as identical motifs. For each hexanucleotide, significance score was calculated by Oligo-analysis programme (Li et al., 2006). Briefly, oligo-analysis programme calculates *p*-values from binomial test. Expected oligonucleotide frequencies were calculated from *Arabidopsis* whole genome sequences and observed oligonucleotide frequencies were obtained from 748 extended binding sites.

To define significant motif, only 99th percentile of hexanucleotides, top-20 ranked, were regarded as significant motifs.

$$\text{neighbouring motif score}_i = \sum_j \text{sig}_{ij}$$

$$\text{sig}_j = -\log_{10}(p \text{ value})_j$$

The neighbouring motif score of *i*th binding site is calculated by summing up the sig value of *j*th matching motifs within 500 bp window.

4.2 CpG island

The CpG islands were predicted using the CpG Island Searcher programme (Takai and Jones, 2002). The programme was downloaded from the website (http://cpgislands.usc.edu/) and run with parameter as GC% higher than 50, observed/expected (o/e) ratio higher than 0.6, and 200 bp window size. The CpG islands score for each binding site was calculated by checking whether any CpG islands were overlapping within extended 500 bp window.

4.3 DNA methylation score

Genome-wide DNA methylation map was constructed using high-resolution mCIP method (Zhang et al., 2006). The raw tiling array data was downloaded from Gene Expression Omnibus (GEO) and accession number is GSE5094. Data normalisation and analysis was conducted using Tiling Analysis Software (TAS version 1.1.02, Affymetrix) as described in the user manual. DNA methylation information was mapped to chromosome position and the methylation score for each binding site was calculated by averaging methylation level within extended 500 bp window.

4.4 Low Nucleosome Density (LND) score

Genome-wide LND regions were identified using high-resolution ChIP-chip assay (Zhang et al., 2007). The raw data of ChIP-chip experiment was downloaded from GEO and accession numbers are GSE7062 and GSE7063. Data normalisation and analysis was performed using Tilemap (Ji and Wong, 2005) with Hidden Markov model option as previously described (Zhang et al., 2006). LND regions were defined as those giving higher signal when probed with the input DNA samples than with H3 ChIP-chip DNA samples. The LND score for each binding site was calculated by averaging LND level within extended 500 bp window.

4.5 Repeat element region

Arabidopsis repeat element information was downloaded from MIPS Repeat Element Database (http://mips.gsf.de/proj/plant/webapp/recat/). They provide all repeat elements mapped to chromosome position. The repeat score for each binding site was calculated by checking whether any repeat elements were overlapping within extended 500 bp window.

4.6 Averaged gene expression level

Since the samples for ChIP-chip assay were extracted from seed tissue, seed-specific microarray data were collected from GEO. Total 37 *Arabidopsis* ATH1 Genome array from 4 different experiments were collected and their accession numbers are GSE14374, GSE7227, GSE5687 and GSE5701. Data analysis was performed using Limma package (Wettenhall and Smyth, 2004) implemented in the Bioconductor R project (http://www.bioconductor.org/). The averaged gene expression score for each binding site was calculated by averaging all gene expression intensities within extended 10 kbp window.

4.7 Classification and attribute selection

As a classification tool, the Weka (Witten and Frank, 2005) and Bioweka (Gewehr et al., 2007) programme was used. WEKA is a comprehensive workbench for machine learning and data mining. The programmes were downloaded from their homepages (http://www.cs.waikato.ac.nz/ml/weka/ and http://bioweka.sourceforge.net/). The classification process was conducted in command line batch mode with ten-fold cross-validation option. Whereas Random Forest was applied with default parameters, SVM parameters

were adjusted to increase classification accuracy with greedy search. Attribute selection was performed with *CfsSubsetEval/ExhaustiveSearch* method.

Acknowledgements

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-C1090-1021-0012)**.** This work was also supported by WCU (World Class University) program (R32-2008-000-10218-0) and National Research Lab. Program (R0A-2005-000-10094-0 to D.L. and R0A-2007-000-20024-0 to G.C.). The authors would like to acknowledge the support from the Korea Institute of Science and Technology Information (KISTI) Supercomputing Center.

References

Atchley, W.R. and Fitch, W.M. (1997) 'A natural classification of the basic helix-loop-helix class of transcription factors', *Proc. Natl. Acad. Sci., USA*, Vol. 94,pp.5172–5176.

Atchley, W.R., Terhalle, W. and Dress, A. (1999) 'Positional dependence, cliques, and predictive motifs in the bHLH protein domain', *J. Mol. Evol.*, Vol. 48, pp.501–516.

Bennetzen, J.L. (2000) 'Transposable element contributions to plant gene and genome evolution', *Plant Mol. Biol.*, Vol. 42, pp.251–269.

Bieda, M., Xu, X., Singer, M.A., Green, R. and Farnham, P.J. (2006) 'Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome', *Genome Res.*, Vol. 16, pp.595–605.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, pp.5–32.

Chan, S.W., Henderson, I.R. and Jacobsen, S.E. (2005) 'Gardening the genome: DNA methylation in Arabidopsis thaliana', *Nat. Rev. Genet.*, Vol. 6, pp.351–360.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, Vol. 20, pp.273–297.

Gewehr, J.E., Szugat, M. and Zimmer, R. (2007) 'BioWeka–extending the Weka framework for bioinformatics', *Bioinformatics*, Vol. 23, pp.651–653.

Gomez-Porras, J.L., Riano-Pachon, D.M., Dreyer, I., Mayer, J.E. and Mueller-Roeber, B. (2007) 'Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice', *BMC Genomics*, Vol. 8, p.260.

Ji, H. and Wong, W.H. (2005) 'TileMap: create chromosomal map of tiling array hybridizations', *Bioinformatics*, Vol. 21, pp.3629–3636.

Jones, S. (2004) 'An overview of the basic helix-loop-helix proteins', *Genome Biol*., Vol. 5, p.226.

Kadonaga, J.T. (2004) 'Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors', *Cell*, Vol. 116, pp.247–257.

Kass, S.U., Pruss, D. and Wolffe, A.P. (1997) 'How does DNA methylation repress transcription?', *Trends Genet*, Vol. 13, pp.444–449.

Klose, R.J. and Bird, A.P. (2006) 'Genomic DNA methylation: the mark and its mediators', *Trends Biochem. Sci.*, Vol. 31, 89–97.

Ledent, V. and Vervoort, M. (2001) 'The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis', *Genome Res.*, Vol. 11, pp.754–770.

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., Guo, J., Liang, W., Chen, L., Yin, J., Ma, H., Wang, J. and Zhang, D. (2006) 'Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis', *Plant Physiol.*, Vol. 141, pp.1167–1184.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002) 'An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments', *Nat. Biotechnol.*, Vol. 20, pp.835–839.

Murre, C., McCaw, P.S. and Baltimore, D. (1989) 'A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins', *Cell*, Vol. 56, pp.777–783.

Nair, S.K. and Burley, S.K. (2000) 'Recognizing DNA in the library', *Nature*, Vol. 404, No. 715, pp.717–718.

Oh, E., Kang, H., Yamaguchi, S., Park, J., Lee, D., Kamiya, Y. and Choi, G. (2009) 'Genome-wide analysis of genes targeted by phytochrome interacting factor 3-like5 during seed germination in arabidopsis', *Plant Cell*, Vol. 21, pp.403–419.

Oh, E., Yamaguchi, S., Kamiya, Y., Bae, G., Chung, W.I. and Choi, G. (2006) 'Light activates the degradation of PIL5 protein to promote seed germination through gibberellin in arabidopsis', *Plant J.*, Vol. 47, pp.124–139.

Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E. and Zhao, H. (2006) 'Pathway analysis using random forests classification and regression', *Bioinformatics*, Vol. 22, pp.2028–2036.

Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998) 'Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation', *Nat. Biotechnol*, Vol. 16, pp.939–945.

Sekinger, E.A., Moqtaderi, Z. and Struhl, K. (2005) 'Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast', *Mol. Cell.*, Vol. 18, pp.735–748.

Shen, Q., Zhang, P. and Ho, T.H. (1996) 'Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient for ABA induction of gene expression in barley', *Plant Cell*, Vol. 8, pp.1107–1119.

Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y. and Hakoshima, T. (1997) 'Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition', *EMBO J.*, Vol. 16, pp.4689–4697.

Takai, D. and Jones, P.A. (2002) 'Comprehensive analysis of CpG islands in human chromosomes 21 and 22', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp.3740–3745.

Toledo-Ortiz, G., Huq, E. and Quail, P.H. (2003) 'The arabidopsis basic/helix-loop-helix transcription factor family', *Plant Cell*, Vol. 15, pp.1749–1770.

van Helden, J., Andre, B. and Collado-Vides, J. (1998) 'Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies', *J. Mol. Biol.*, Vol. 281, pp.827–842.

Wettenhall, J.M. and Smyth, G.K. (2004) 'Limmagui: a graphical user interface for linear modeling of microarray data', *Bioinformatics*, Vol. 20, pp.3705–3706.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, San Francisco.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J. and Jacobsen, S.E. (2007) 'Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis', *PLoS Biol.*, Vol. 5, p.e129.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. and Ecker, J.R. (2006) 'Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis', *Cell*, Vol. 126, pp.1189–1201.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T. and Henikoff, S. (2007) 'Genome-wide analysis of arabidopsis Thaliana DNA methylation uncovers an interdependence between methylation and transcription', *Nat. Genet.*, Vol. 39, pp.61–69.