Supplementary Material for the *Bioinformatics* submission “KIRMES: Kernel-based Identification of Regulatory Modules in Euchromatic Sequences”

Sebastian J. Schultheiss,[1,2,*] Wolfgang Busch,[2,4] Jan U. Lohmann,[2,5]
Oliver Kohlbacher,[3] and Gunnar Rätsch[1]

[1]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
[2]Max Planck Institute for Developmental Biology, Tübingen, Germany
[3]Wilhelm Schickard Institute for Computer Science, University of Tübingen, Germany
[4]*Present Address:* Biology Department, Duke University, Durham NC, USA
[5]*Present Address:* Department of Stem Cell Biology, University of Heidelberg, Germany

ABSTRACT

In this supplement, we describe the sources of the data we used, how the multiple genome alignments for plants were computed and a detailed report on a comparison of KIRMES to PRIORITY.

S.1 Microarray Data

The expression datasets are publicly available at the EBI ArrayExpress repository for microarray experiments at `http://www.ebi.ac.uk/arrayexpress` as experiment number E-MEXP-98 for the heat shock dataset and E-MEXP-432 for the overexpression dataset (Busch *et al.*, 2005; Leibfried *et al.*, 2005). The datasets were combined according to their experimental logic (see dataset annotations below and original publications) and gene sets were obtained through a calculation of expression change using the software GeneSpring (Agilent Technologies). We labeled genes as co-expressed when they showed a four-fold change of expression in the experiment as compared to the control, and considered those genes not co-expressed if their levels remain the same, compared to the control, within a margin of 0.2 fold change. See also Section 3.3 of the publication.

The individual gene sets we derived are named after the gene whose expression was modified during the experiment. The gene name abbreviations are according to the standard TAIR7 notation of short gene names (Arabidopsis Information Resource (TAIR), 2007).

If a gene name is prefixed with AlcA, the promoter of that gene was modified so it could be over-expressed (gain of function, cf. Busch *et al.* (2005) for details). The FASTA files generated for the use within KIRMES are available as supplementary materials on the companion website to this publication. If the gene name is not prefixed, we used the loss of function experiment to obtain this list.

The abbreviation “4fold” stands for a four-fold change in expression, this is the threshold for significant change in expression we used, as mentioned in Section 3.3 of the main publication.

Subtractions are gene sets derived from subtracting the intersection of genes that have a significant change in expression in the two experiments from the first one. Intersections only contain genes with significant changes in expression levels in both experiments relative to the control. Joins contain genes of this kind from both experiments. The loop logic set contains genes that are downregulated when WUS is supressed and upregulated when WUS is upregulated.

Finally, the control experiment in the first column should not behave differently from a wild type plant and therefore there should be no discriminative information in this experiment *vs.* its control.

S.2 Conservation Data

We obtained the sequence conservation information on *Arabidopsis thaliana* aligning every gene according to the TAIR7 annotation Arabidopsis Information Resource (TAIR) (2007), including the surrounding intergenic regions, to other plant organisms with sequenced genomes. This differs from a whole genome alignment: it allows for multiple mappings of the same region of another plant’s genome onto the *A. thaliana* genome.

Available to us were the genome sequences of *Medicago truncatula* Cannon *et al.* (2006), *Oryza sativa* (Itoh *et al.*, 2007), *Carica papaya* (Ming *et al.*, 2008), *Arabidopsis lyrata*, and *Populus trichocarpa* (Tuskan *et al.*, 2006). To compute these alignments, we first did an initial BLASTn search of every *A. thaliana* gene against a database of each plant’s chromosome sequences (Altschul *et al.*, 1997). The best BLAST hit was used as a seed for an optimal local alignment by means of the Smith-Waterman implementation of the EMBOSS suite of tools (Rice *et al.*, 2000).

This data is entierely centered around *A. thaliana* and can therefore not be used with data from any of the other plants, but the method can be applied to any organism with annotations about the gene locations.

[*] to whom correspondence should be addressed

With this method, we align different numbers of genes to the total of 30 144 genes of *A. thaliana* from the TAIR7 annotation: 5 524 genes of *A. thaliana* with *Oryza sativa*, 7 732 with *Medicago truncatula*, 11 946 with *Populus trichocarpa*, 14 302 with *Carica papaya* and 24 088 genes with *thaliana's* closest relative, *Arabidopsis lyrata* (*cf.* Figure S.1. For the conservation information in the K_{WDSC} kernel, only matches contribute favorably, other conditions are ignored.

S.3 Comparison with PRIORITY

We performed some addional comparisons not mentioned in the main paper. The comparison was in that case closer to the original setup by Gordân *et al.* (2008): we compared the performance of PRIORITY using the literature consensus motifs, which were the benchmark in the original setup. The criterion for a successful prediction is somewhat changed: In the crossvalidation setup, we let PRIORITY run on 80% of the data and compare the top-scoring motif to the literature consensus. If in turn the occurrence is within the boundaries outlined by Gordân *et al.* (2008), *i.e.* the top-scoring motif is less than 0.25 of inter-motif distance from the literature consensus, we proceed to check for the performance on the remaining 20% as follows: we search for the bestmatching occurrence of the position-specific scoring matrix that PRIORITY reported as its first result. We count the performance of PRIORITY as a success, if the best match in each gene of the positive test set is on average below the inter-motif distance value of 0.25, and above that threshold in the negative set. In Figure S.2, we show the success rate of the two approaches on 158 gene sets from Harbison *et al.* (2004), preprocessed by Raluca Gordân. The number of successes of PRIORITY as reported by Gordân *et al.* (2008) is shown in the first row, the result of the experiments according to our crossvalidation setup is listed in the second row. The third row shows the performance of KIRMES on the split datasets. We count a success as KIRMES scoring an auROC curve of 0.75, with an average auROC of 0.71 overall and 0.79 on the successful datasets. The fourth column shows the successes of PRIORITY when using the positive sets obtained through a single run of KIRMES on the unsplit data rather than the initial sets. By applying KIRMES, these sets are filtered for significance and should contain less noise from false positives.

S.4 Determining the Length of the Motif Window

We tested different settings for the motif window length. A setting close to the actual length of the motif is not as good as a slightly larger setting. There might be another motif close by or some additional information about the sequence that helps the classifier with its discrimination in a window of 19–24 base pairs around the motif center. In the studied gene set, increasing length further seems to introduce more noise and therefore do more harm than good. Table S.1 shows the tested window lengths in a representative dataset, which was already used to evaluate the contributions of the different kernels and features (*cf.* Results section in the main publication). We give the area under the receiver operating characteristic curve (auROC), achieved with a constant set of splits for the fivefold crossvalidation.

Window length	auROC	Window length	auROC
6	0.76	22	0.85
8	0.79	24	0.83
10	0.79	26	0.79
12	0.80	28	0.79
14	0.80	30	0.79
16	0.80	35	0.76
18	0.80	40	0.74
19	0.83	50	0.71
20	**0.89**	60	0.71
21	0.85	100	0.68

Table S.1. Different window length tested and the resulting auROC.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Arabidopsis Information Resource (TAIR) (2007). Arabidopsis genome annotation tair7. `http://arabidopsis.org`.

Busch, W., Wunderlich, M., and Schoeffl, F. (2005). Identification of novel heat shock factor-dependent genes and biochemical pathways in a. thaliana. *Plant J*, **41**(1), 1–14.

Cannon, S., Sterck, L., Rombauts, S., Sato, S., and et al. (2006). Legume genome evolution viewed through the medicago truncatula and lotus japonicus genomes. *Proc Natl Acad Sci U S A*, **103**(40), 14959–14964.

Gordân, R., Narlikar, L., and Hartemink, A. (2008). A fast, alignment-free, conservation-based method for transcription factor binding site discovery. In M. Vingron and L. Wong, editors, *Lecture Notes in Computer Science: RECOMB 2008*, volume 4955, pages 98–111. Springer Verlag Heidelberg, Germany.

Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**(7004), 99–104.

Ipollito, F. (2007). Phylogenetic tree of flowering plants. *Natural History Magazine*, (0307), online.

Itoh, T., Tanaka, T., Barrero, R., Yamasaki, C., Fujii, Y., and et al. (2007). Curated genome annotation of oryza sativa ssp. japonica and comparative genome analysis with arabidopsis thaliana. *Genome Res*, **17**(2), 175–183.

Leibfried, A., To, J., Busch, W., Stehling, S., Kehle, A., Demar, M., Kieber, J., and Lohmann, J. (2005). Wuschel controls meristem function by direct regulation of cytokinin-inducible response regulators. *Nature*, **438**, 1172–1175.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Wang, M.-L., Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J.-K., Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Perez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Luo, M.-C., Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Mitchell-Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (carica papaya linnaeus). *Nature*, **452**(7190), 991–996.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends Genet*, **16**(6), 276–277.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., and et al. (2006). The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, **313**(5793), 1596–1604.

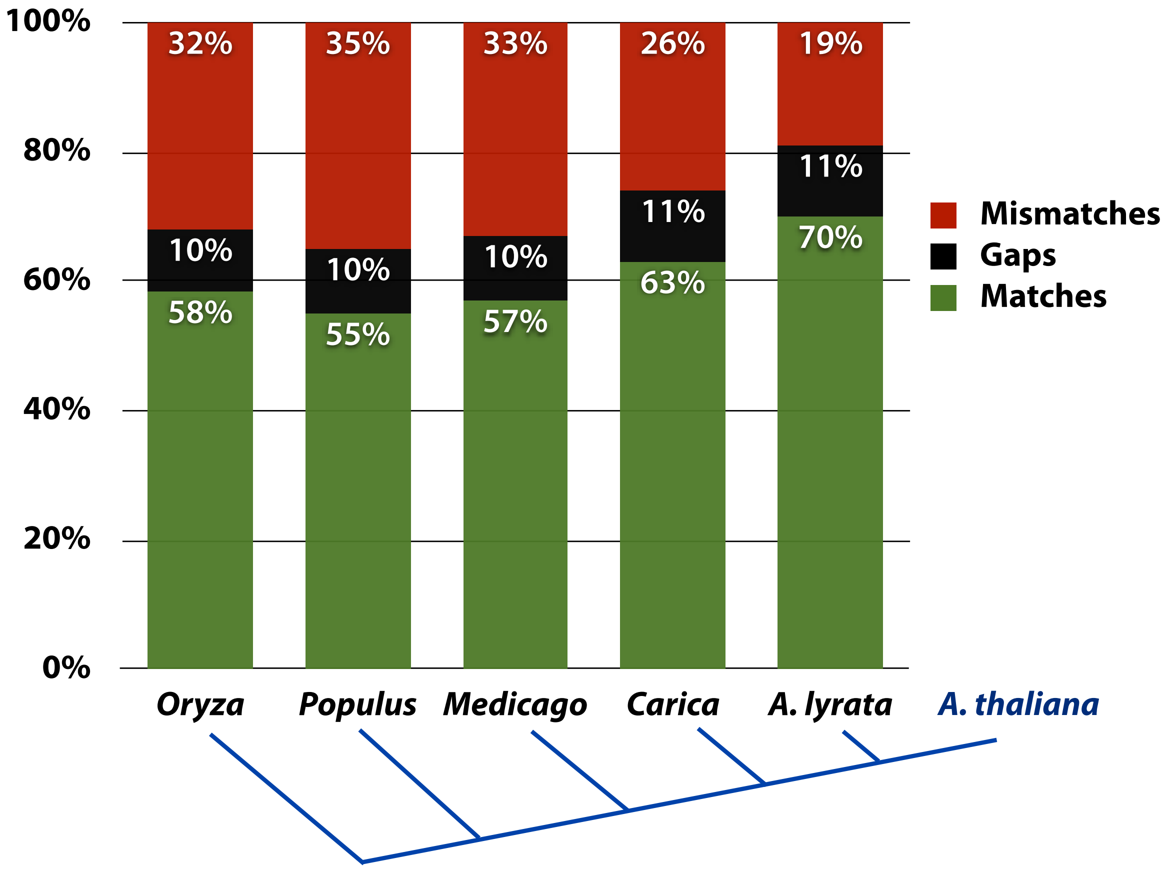

Fig. S.1. Results of the alignment of three other plant genomes against *Arabidopsis thaliana*. Every bar sums up to 100% and is divided between matches (green), mismatches(red) and gaps(black). The similarity decreases almost exactly as the evolutionary distance increases, as shown through the phylogenetic tree (blue lines) (Ipollito, 2007).

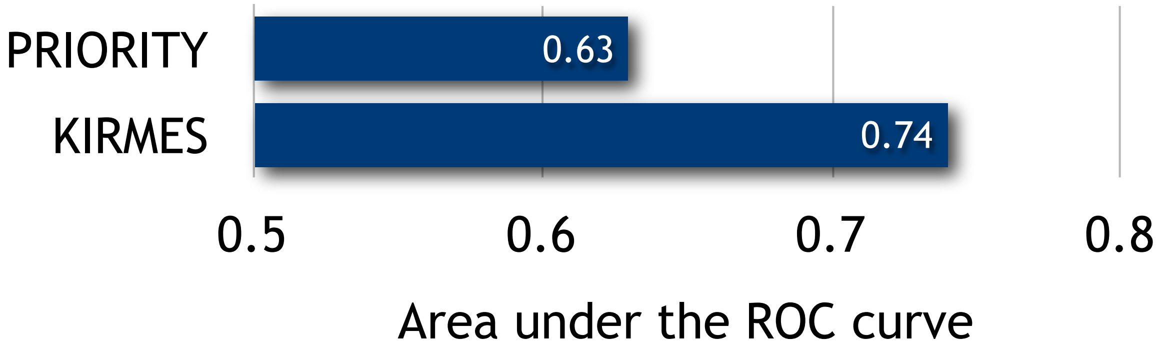

Fig. S.2. omparison between the Gibbs sampler PRIORITY and the KIRMES approach. The rows show a percentage of successes on the 158 gene sets; the definition of both a success and the gene sets can be found in the text (*cf.* Section S.3).