A Supplementary Material for the GCB 2008 submission KIRMES: Kernel-based Identification of Regulatory Modules in Euchromatic Sequences

A.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.

A.2 Conservation Data

We obtain the sequence conservation information on *Arabidopsis thaliana* aligning every gene (according to the TAIR7 annotation [Ara07]), 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* [CSR$^+$06], *Oryza sativa* [ITB$^+$07], and *Populus trichocarpa* [TDJ$^+$06]. To compute these alignments, we first do an initial BLASTn search of every *A. thaliana* gene against a database of each plant's chromosome sequences [AMS$^+$97]. The best BLAST hit is used as a seed for an optimal local alignment by means of the Smith-Waterman implementation of the EMBOSS suite of tools [RLB00].

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.

With this method, we align a total of 7 732 genes of *A. thaliana* with *Medicago truncatula*, 5 524 with *Oryza sativa* and 11 946 with *Populus trichocarpa*. Of these alignments, 55–58% of the bases are matches, 10% are gaps and 32–35% are mismatches. For the conservation information in the K_{WDSC} kernel, only matches contribute favorably, other conditions are ignored.

References

[AMS+97] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, September 1997.

[Ara07] Arabidopsis Information Resource (TAIR). Arabidopsis Genome Annotation TAIR7. `http://arabidopsis.org`, April 2007.

[CSR+06] S.B. Cannon, L. Sterck, S. Rombauts, S. Sato, and et al. Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. *Proc Natl Acad Sci U S A*, 103(40):14959–14964, 2006.

[ITB+07] T. Itoh, T. Tanaka, R.A. Barrero, C. Yamasaki, Y. Fujii, and et al. Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res*, 17(2):175–183, 2007.

[RLB00] P Rice, I Longden, and A Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276–277, 2000.

[TDJ+06] G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, and et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, 313(5793):1596–1604, September 2006.