Supplemental Material

Protocol S1

The PromPredict algorithm considers the average free energy over a 100nt window and compares it with the the average free energy over a downstream 100nt window separated by 50bp. Hence, to predict whether a nucleotide 'n+50' is a promoter, the following parameters are determined and E1(n+50) and D(n+50) are compared with their corresponding cutoffs from Table S1.

$$E1(n+50) = \frac{\sum_{n}^{n+100} \Delta G^0}{100} \tag{1}$$

$$E2(n+50) = \frac{\sum_{n+150}^{n+250} \Delta G^0}{100} \tag{2}$$

$$D(n+50) = E1(n+50) - E2(n+50) \tag{3}$$

All such predicted consecutive nucleotides are grouped to form the same prediction if they lie within 50bp of each other. The procedure of arriving at cut-off values and prediction method is described in detail by Rangannan and Bansal (2009).

Table S1: *Cutoff values for E1 and D:* The cutoff values are applied according to the GC content of surrounding 1000nt fragment. The cutoffs have been derived by training on transcription start site (TSS) and translation start site (TLS) data of prokaryotes (Rangannan and Bansal, 2010) and match well with the AFE values obtained for upstream promoter and downstream non-promoter regions in Arabidopsis and rice, as shown in Fig. S3.

GC% Range	E1-cutoff	D-cutoff	GC% Range	E1-cutoff	D-cutoff
15 – 20	-14.5	1.0	50 – 55	-19.2	1.4
20 – 25	-15.1	1.0	55 – 60	-20.0	1.4
25 – 30	-15.6	1.0	60 – 65	-21.0	1.2
30 – 35	-16.3	1.0	65 – 70	-22.0	1.2
35 – 40	-16.8	1.2	70 – 75	-23.0	1.2
40 – 45	-17.3	1.4	75 – 80	-24.0	1.2
45 – 50	-18.3	1.4	80 – 85	-25.0	1.2

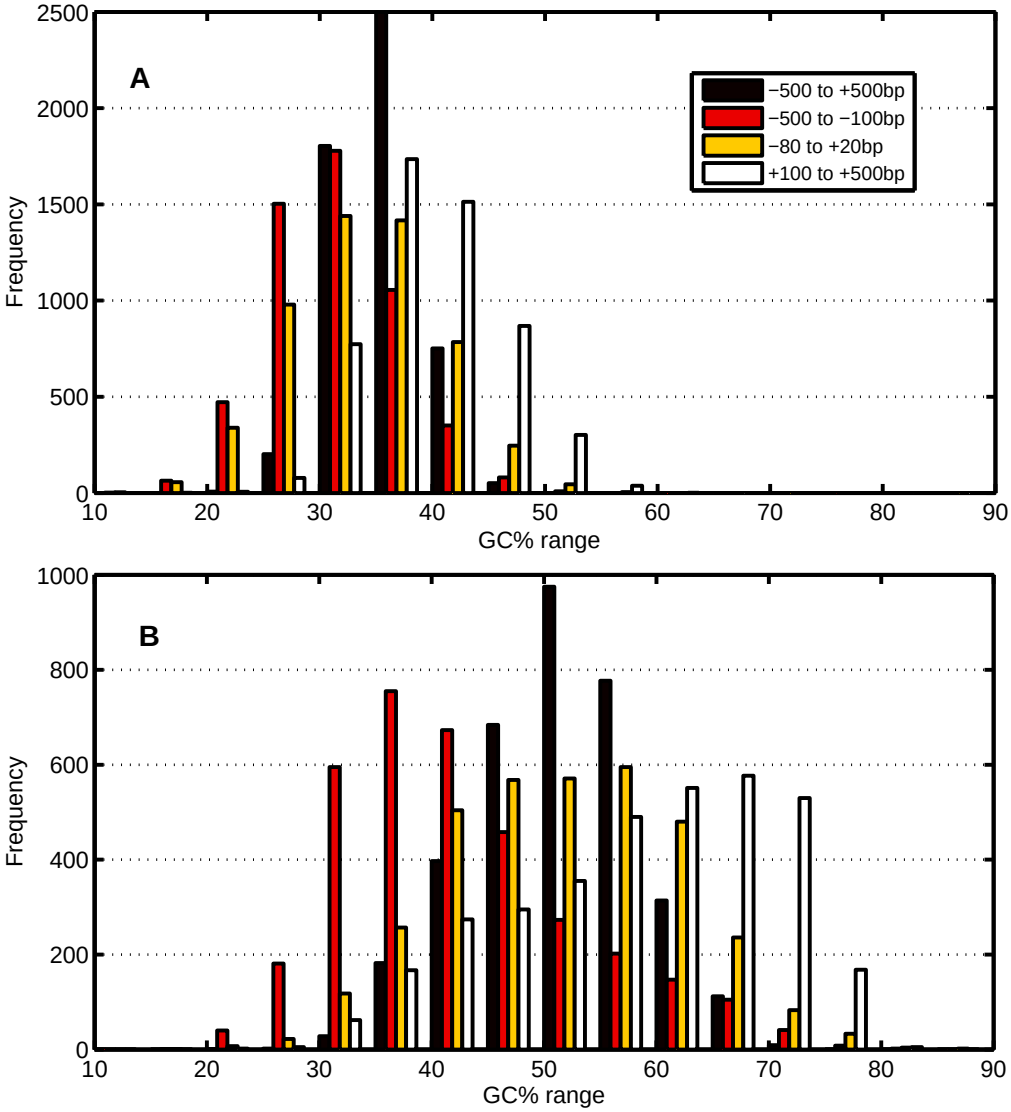

Figure S1: *GC content distribution for (A) Arabidopsis and (B) rice sequences in the vicinity of transcription start site (TSS):* The frequency distribution of GC content in a 1000bp segment (−500bp to +500bp) around TSS, upstream sequence (−500bp to −100bp), downstream sequence (+100bp to +500bp) and sequence adjacent to the TSS (−80bp to +20bp) are shown. The GC content of Arabidopsis promoters shows a narrow range within each subclass and small differences between the subclasses. The GC content for the rice promoters shows a broad range with large differences between subclasses especially upstream and downstream regions.

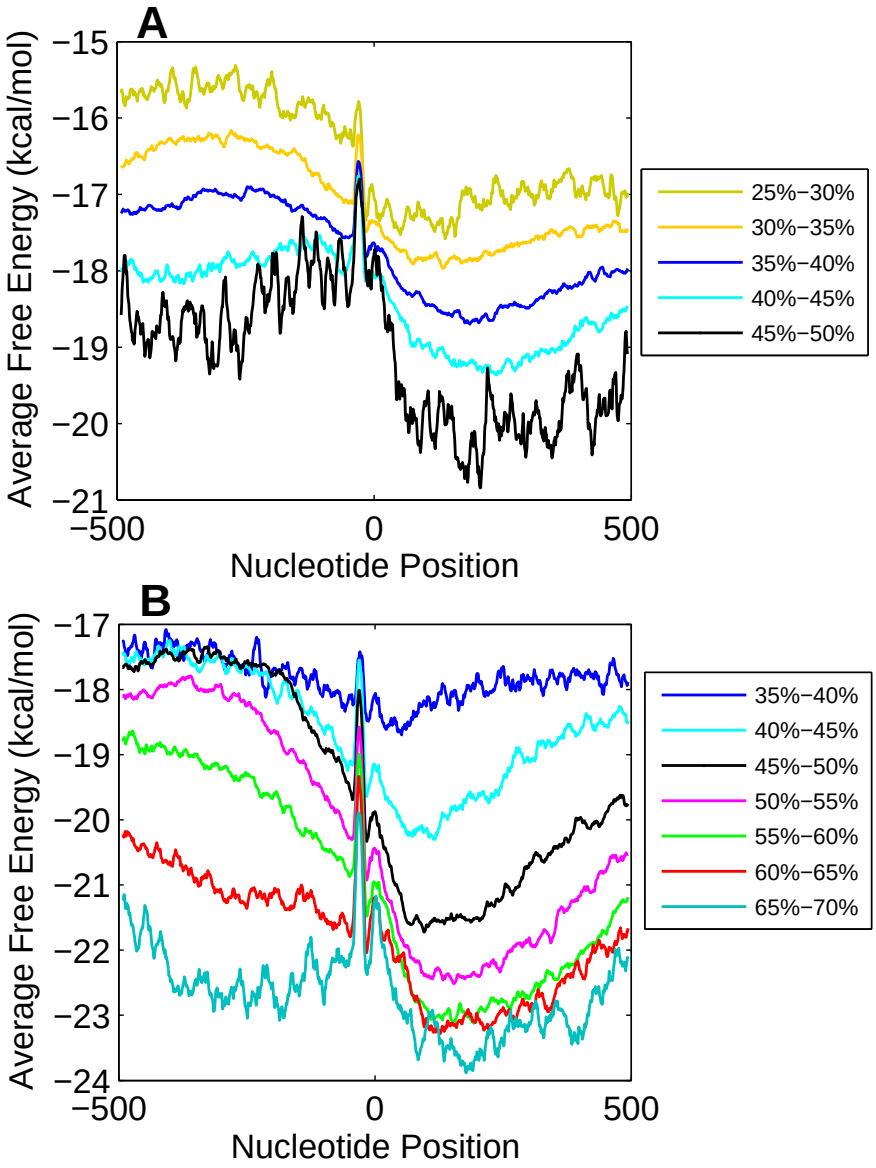

Figure S2: AFE profiles for 1000bp sequences (−500 to +500bp with respect to TSS) in different ranges of GC content are shown for genes from chr.1 of (A) Arabidopsis and (B) rice.

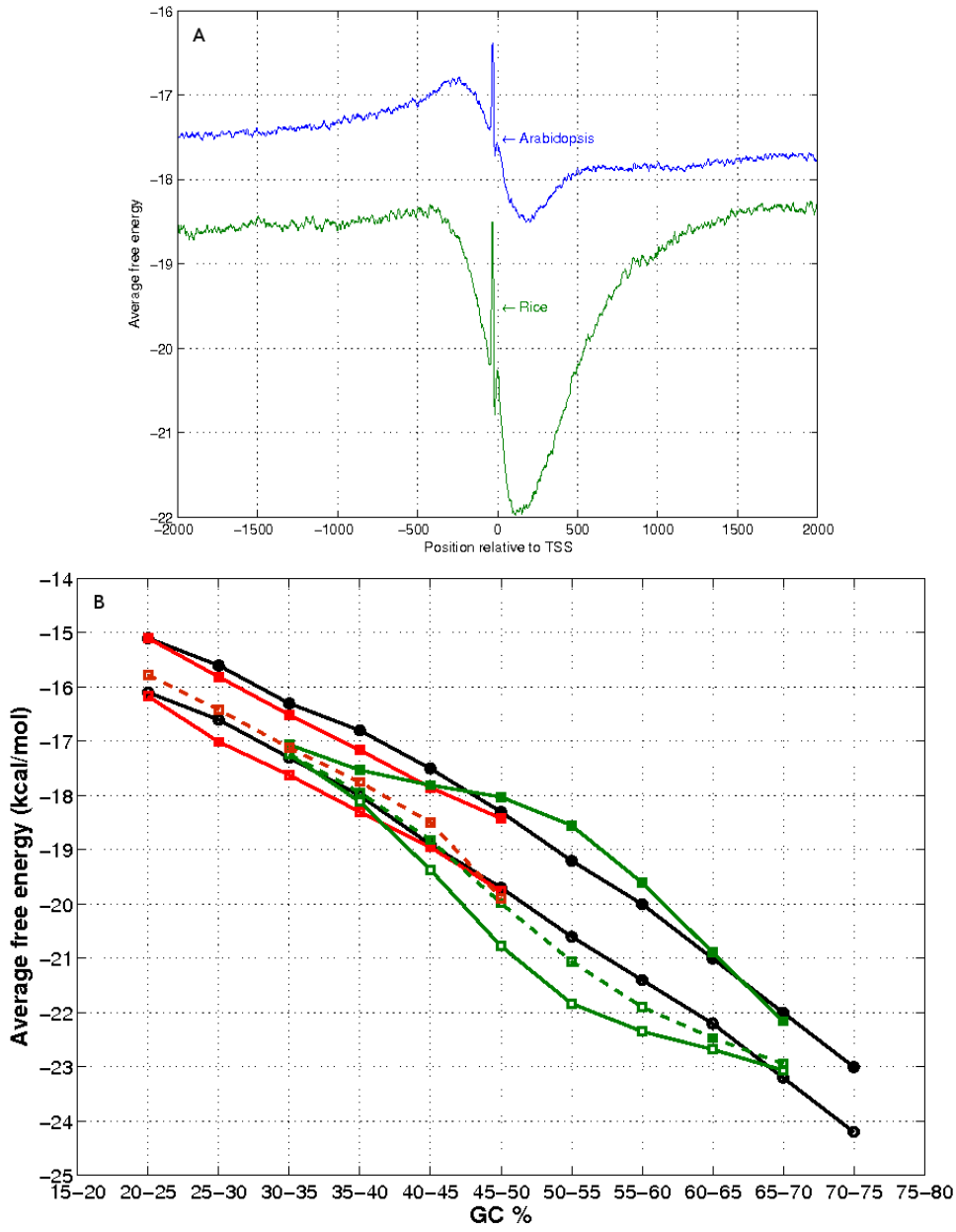

Figure S3: (A) AFE profiles in the region −2000bp to +2000bp in the vicinity of the TSS for chr1 Arabidopsis and rice. (B) The threshold values of free energy used to predict promoters in genomic DNA with GC content varying between 20 to 80% as derived from prokaryotic genomic data are shown in black. The filled circles correspond to the minimal free energy threshold (E1) for a particular fragment and hollow circles to the free energy (E2) in its downstream region, for it to be assigned as a putative promoter. The average free energy values in a representative set of 1001 nt long sequences flanking the TSSs in Arabidopsis and rice are shown as red and green squares respectively. The AFE values for the +500 to +1000 downstream regions in Arabidopsis and rice are shown as hollow squares connected by a dashed lines.

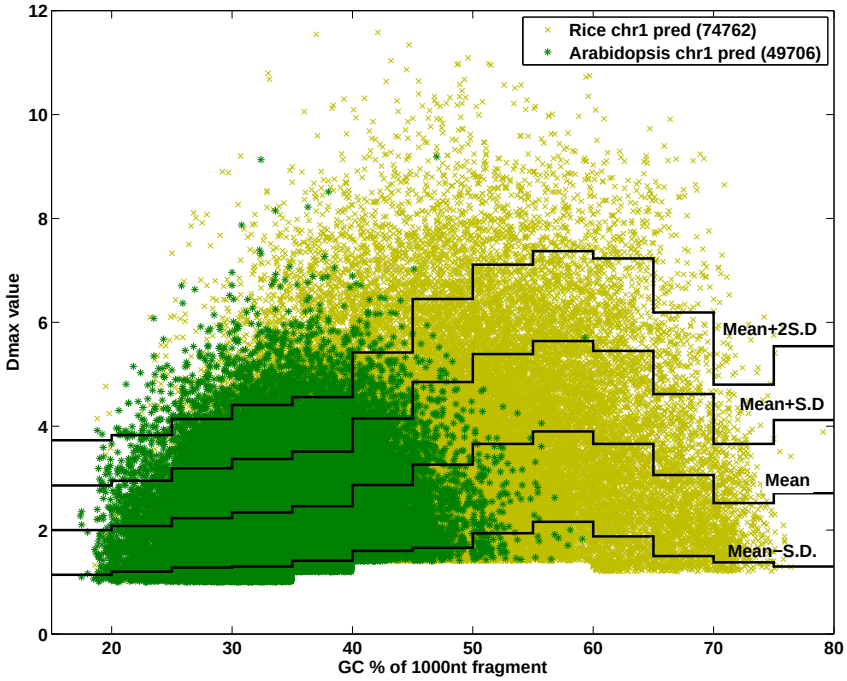

Figure S4: Cutoffs values for score classes were determined from the mean and standard deviation of Dmax values from Arabidopsis and rice predictions.

Table S2: Cutoff values based on mean and standard deviation of maximum D value of Arabidopsis and rice predictions

GC% Range	**No. of seq**[a]	**Dmax cutoff**			
		Mean−S.D.	Mean	Mean+S.D.	Mean + 2S.D.
15-20	113	1.14	2.00	2.86	3.73
20-25	1789	1.20	2.08	2.95	3.83
25-30	9405	1.28	2.23	3.19	4.14
30-35	26073	1.30	2.34	3.37	4.41
35-40	37942	1.41	2.46	3.51	4.56
40-45	20018	1.60	2.87	4.15	5.42
45-50	11709	1.66	3.26	4.85	6.45
50-55	7195	1.94	3.66	5.39	7.11
55-60	4725	2.16	3.90	5.64	7.37
60-65	2903	1.88	3.66	5.45	7.23
65-70	1836	1.50	3.06	4.62	6.19
70-75	726	1.38	2.52	3.66	4.80
75-80	24	1.30	2.71	4.12	5.54

a: Sequences pooled from both rice and Arabidopsis predictions

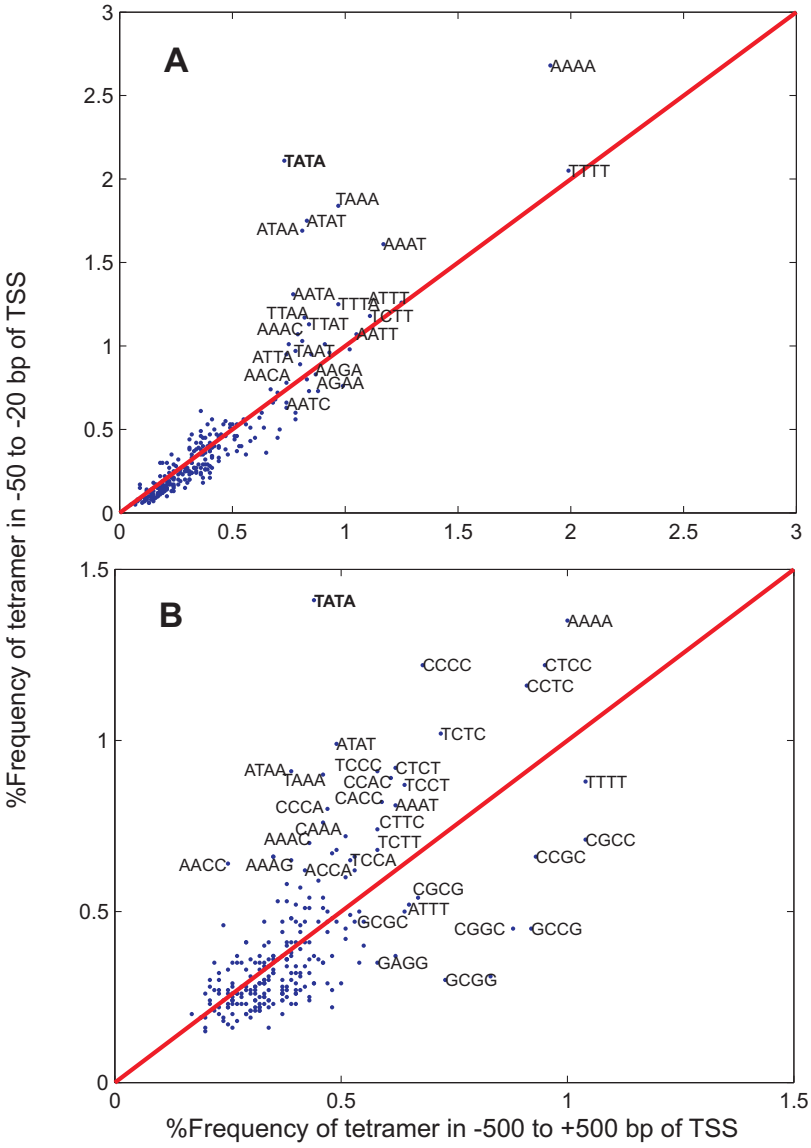

Figure S5: Percentage frequency of occurrence of tetramers in core promoter (−50 to −20 bp) as compared to that in the 1000nt (−500 to +500 bp) region in the vicinity of TSS for (A) Arabidopsis and (B) rice.

Protocol S2

Gene Ontology: The genes categorized in the Gene Ontology GO SLIM categories were downloaded from TAIR (http://www.arabidopsis.org) for Arabidopsis and AmiGO (The Gene Ontology Consortium, 2000; Carbon et al., 2009) for rice. The percentage of genes that have TP predictions in region −500 to +100bp with respect to the TSS (TP_{genes}) and that don't have any predictions in this region (FP_{genes}) were determined.

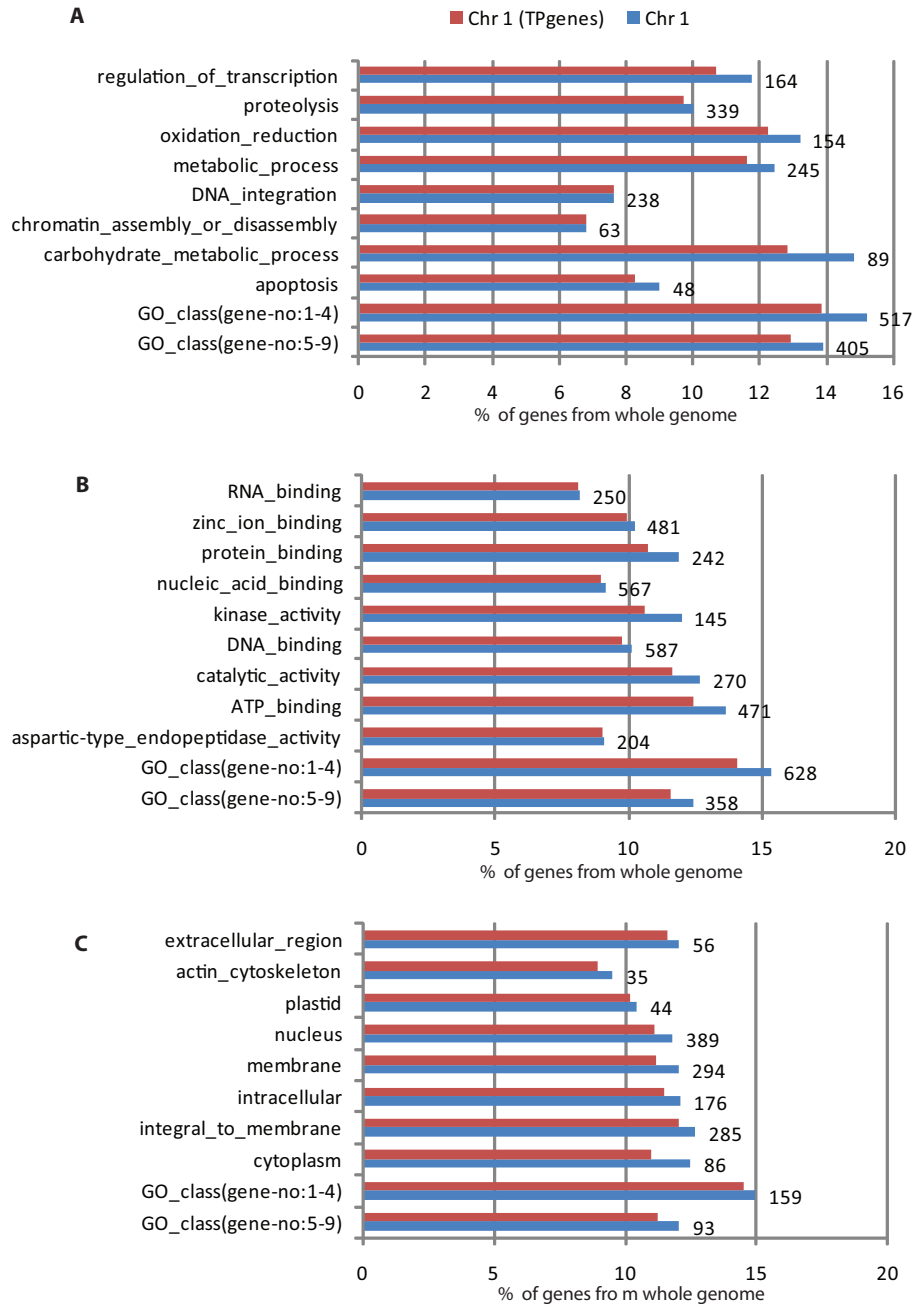

Figure S6: Gene representation in GO categories for (A) Biological Process, (B) Molecular Function and (C) Cellular Component. The figure shows genes from rice chr 1 for each GO category as a percentage of the genes in the whole genome present in that category. The red bar indicates the TP genes while the blue bar indicates all genes. The number of chr 1 genes assigned to each category are mentioned adjacent to each bar.

Table S3: Distribution of Arabidopsis TP and FN genes in GO categories as a percentage of the total genes in the genome present in that category.

Functional Category	Total genes	TP: % of Total	FN: % of Total
Cellular Component			
extracellular	359	95.265	4.178
cell wall	485	94.639	5.361
other cellular components	2406	93.724	6.359
plasma membrane	1827	93.596	6.568
nucleus	2079	93.362	6.590
plastid	969	92.570	7.946
other membranes	2528	92.366	7.872
Golgi apparatus	207	92.271	9.179
chloroplast	2635	91.879	8.008
ER	348	91.667	7.759
other cytoplasmic components	2696	91.320	8.828
other intracellular components	3415	91.274	8.726
cytosol	664	91.114	9.337
unknown cellular components	5050	91.030	8.772
mitochondria	881	90.919	9.535
ribosome	388	86.856	13.402
Molecular Function			
transcription factor activity	1188	94.737	5.343
nucleic acid binding	615	94.181	5.360
other enzyme activity	2481	93.764	6.652
other binding	2579	93.748	6.361
hydrolase activity	1700	93.253	6.802
transporter activity	864	93.204	6.580
other molecular functions	998	93.010	6.710
transferase activity	1775	92.641	8.038
nucleotide binding	1443	92.322	7.550
DNA or RNA binding	1428	92.248	8.075
kinase activity	877	92.122	8.298
protein binding	1787	91.735	8.419
receptor binding or activity	118	91.473	6.977
unknown molecular functions	5108	91.443	8.342
structural molecule activity	387	87.755	12.245
Biological Process			
transcription	1231	94.801	5.280
signal transduction	771	94.034	6.096
other biological processes	1396	93.625	6.734
response to stress	1529	93.525	6.867
response to abiotic or biotic stimulus	1487	93.208	7.061
developmental processes	1386	92.857	7.937
other metabolic processes	6503	92.773	7.335
other cellular processes	6949	92.488	7.541
protein metabolism	2378	92.010	7.906
transport	1364	91.862	7.478
cell organization and biogenesis	976	91.496	7.992
unknown biological processes	6499	91.429	8.340
electron transport or energy pathways	220	91.364	9.545
DNA or RNA metabolism	271	90.775	8.487

Table S4: Geneids for orthologous genes selected for comparison

Gene	Arabidopsis geneids	Rice geneids
Aspartate aminotransferase	At2g30970	Os02g0236000
Cu/Zn superoxide dismutase	At5g18100	Os03g0219200
Dof genes	At3g52400	Os03g0787000
P-type ATPase	At1g07670	Os03g0281600
FAD2	At3g12120	Os02g0716500
PRF1	At2g19760	Os10g0323600

References

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., the AmiGO Hub, and the Web Presence Working Group (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25:288–9.

Rangannan, V. and Bansal, M. (2009). Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. *Mol. Biosyst.*, 5:1758–69.

Rangannan, V. and Bansal, M. (2010). High quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics*, 26:3043–50.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25:25–9.