# Associating transcription factor binding site motifs with target GO terms and target genes: Supplementary Material

Mikael Bodén Timothy L Bailey

May 14, 2008

Table 1 provides AUC50 for each GO term prediction method when applied on the yeast Gold standard set at each of the three target levels. Table 2 similarly allows a comparison of methods at all target levels when applied on the yeast data, here at a fixed $E = 10$.

Of 51 JASPAR motifs for human, 21 were analysed anecdotally by referring to UniProt entries of transcription factors. A maximum of two predicted GO terms and a few keywords extracted from UniProt are provided in Tables 3-4.

Results from applying the proposed aggregate target gene score, for individual TFs, are presented in Tables 5-6. The change in AUC50 caused by including the predicted GO terms into the gene score correlates weakly but positively with the number of predicted terms (0.19).

Of the genes ranked highest by Avg-Odds-GO for each of the 63 transcription factor binding motifs, 54 were included as target genes in the MacIsaac et al. low-confidence map. 25/63 of the top-ranked target genes were in their high-confidence map.

Illustrating typical target gene predictions, the top-20 genes for DIG1 and REB1 appear in Tables 7-8.

DIG1 has a small set of predicted GO terms (3) and consequently identifies a smaller set of genes (217) to receive a doubled target gene score. To put this into context, there are 82 target genes in the HC-set which associate significantly with 41 GO terms (at level 1).

REB1, on the other hand, does not have any significant target GO terms among the 217 putative target genes. The target gene predictions here, however, incorporate *predicted* GO terms. For REB1, the predicted GO terms are very generic and apply to all but 1180 genes. Tables 7-8 provide case-based support for the positive influence of predicted GO terms on target gene predictions.

Target GO terms statistically inferred from transcription target sets are provided in a separate MS Excel Workbook (Target.xls).

GO term predictions by Avg-Odds Global in both the yeast and human genomes (1000-bp upstream regions) are provided in a separate MS Excel Workbook (Predict.xls).

| *Method* | *Mean AUC50* | | |
|---|---|---|---|
| | *Level 1* | *Level 2* | *Level 3* |
| **Hit-count** | | | |
| $10^{-3}$ Local | 0.26 | 0.30 | 0.31 |
| $10^{-3}$ Global | 0.30 | 0.34 | 0.39 |
| $10^{-4}$ Local | **0.46** | **0.51** | **0.57** |
| $10^{-4}$ Global | 0.45 | 0.48 | 0.54 |
| $10^{-5}$ Local | 0.22 | 0.23 | 0.24 |
| $10^{-5}$ Global | 0.19 | 0.19 | 0.21 |
| **Avg-Odds** | | | |
| Local | **0.46** | **0.51** | 0.56 |
| Global | 0.45 | 0.50 | 0.56 |
| ZS Local | 0.43 | 0.48 | 0.56 |
| ZS Global | 0.45 | 0.49 | **0.57** |
| ZM Local | 0.44 | 0.49 | 0.56 |
| ZM Global | 0.44 | 0.48 | 0.55 |
| **Max-Odds** | | | |
| Local | 0.44 | 0.49 | 0.54 |
| Global | 0.44 | 0.48 | 0.54 |

Table 1: **Average AUC50 values for each of the methods in Yeast.** Results are for the 43 "Target Level 3" motifs. All three target levels in the gold standard are tested. Higher AUC50 value is better. For the Hit-Count method, the specified motif-match $p$-value threshold is used (see Methods). For Avg-Odds, variations with a sequence-shuffled Z-score (ZS) and a motif-shuffled Z-score (ZM) are trialed.

| *Method* | *Average predicted* | *Level 1 target* | | | *Level 2 target* | | | *Level 3 target* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *1 TP* | *Rec* | *Prec* | *1 TP* | *Rec* | *Prec* | *1 TP* | *Rec* | *Prec* |
| **Hit-count** | | | | | | | | | | |
| $10^{-3}$ Local | 38.4 | 0.71 | 0.31 | 0.17 | 0.71 | 0.35 | 0.15 | 0.64 | 0.33 | 0.12 |
| $10^{-3}$ Global | 34.2 | 0.84 | 0.33 | 0.21 | 0.84 | 0.38 | 0.19 | 0.76 | 0.44 | 0.16 |
| $10^{-4}$ Local | 16.3 | 0.73 | 0.36 | *0.47 | 0.73 | 0.41 | *0.45 | 0.73 | 0.49 | *0.37 |
| $10^{-4}$ Global | 13.0 | 0.69 | 0.31 | *0.48 | 0.69 | 0.35 | *0.47 | 0.64 | 0.40 | *0.37 |
| $10^{-5}$ Local | 3.1 | 0.31 | 0.09 | *0.72 | 0.29 | 0.09 | *0.72 | 0.29 | 0.11 | *0.67 |
| $10^{-5}$ Global | 2.3 | 0.22 | 0.07 | *0.66 | 0.22 | 0.07 | *0.66 | 0.22 | 0.09 | *0.63 |
| **Avg-Odds** | | | | | | | | | | |
| Local | 47.8 | 0.89 | 0.53 | 0.23 | 0.89 | 0.59 | 0.21 | **0.87** | 0.63 | 0.17 |
| Global | 54.0 | **0.91** | **0.54** | 0.21 | **0.91** | **0.60** | 0.19 | **0.87** | **0.66** | 0.15 |
| ZS Local | 39.1 | 0.84 | 0.47 | 0.25 | 0.84 | 0.53 | 0.23 | 0.84 | 0.63 | 0.20 |
| ZS Global | 39.9 | 0.84 | 0.48 | 0.26 | 0.84 | 0.54 | 0.24 | 0.84 | 0.62 | 0.20 |
| ZM Local | 36.7 | 0.89 | 0.46 | **0.28** | 0.87 | 0.51 | **0.26** | **0.87** | 0.60 | **0.21** |
| ZM Global | 37.1 | 0.87 | 0.45 | 0.27 | 0.84 | 0.49 | 0.25 | 0.84 | 0.59 | 0.20 |
| **Max-Odds** | | | | | | | | | | |
| Local | 39.9 | 0.87 | 0.49 | 0.26 | 0.87 | 0.54 | 0.23 | 0.82 | 0.60 | 0.19 |
| Global | 44.2 | 0.89 | 0.51 | 0.23 | 0.89 | 0.57 | 0.21 | 0.84 | 0.63 | 0.17 |

Table 2: **Accuracy of predicted TF-GO term associations in Yeast.** Averaged outcomes for searches using $E = 10$ with the 43 motifs from the *S. cerevisiae* gene regulatory network, and for each of the three GO terms target levels. The columns indicate the average number of predictions returned by each method, the probability of predicting at least one true positive (1 TP, higher is better), the recall, and the precision. When one or more TFs render no predictions they are excluded from the precision average (marked with '*').

| *JASPAR motif* | *Factor (UniProt)* | *Number of GO terms* | *Example of predicted* GO terms *Transcriptional target keywords* |
|---|---|---|---|
| RREB1 | RREB1_HUMAN | 5 | Positive regulation of cell differentiation (GO:0045597)<br>Multicellular organismal development (GO:0007275)<br>*Ras/Raf-mediated cell differentiation*<br>*Represses angiotensinogen gene* |
| NFIL3 | NFIL3_HUMAN | 23 | Defense response (GO:0006952)<br>transcription factor activity (GO:0003700)<br>*Protects pro-B cells from programmed cell death*<br>*Component of the circadian clock* |
| SPIB | SPIB_HUMAN | 8 | Immune response (GO:0006955)<br>Chemokine receptor binding (GO:0042379)<br>*Promotes development of plasmacytoid dendritic cells*<br>*B-cell receptor (BCR) signaling* |
| NFKB1 | NFKB1_HUMAN | 18 | Signal transduction (GO:0007165)<br>Protein binding (GO:0005515)<br>*Immune response, Apoptosis* |
| Pax6 | PAX6_HUMAN | 8 | Anatomical structure morphogenesis (GO:0009653)<br>Wnt receptor signaling and calcium modulat... (GO:0007223)<br>*Development of the eye, nose, ...*<br>*Specification of the ventral neuron subtypes* |
| SRF | SRF_HUMAN | 10 | Extracellular space (GO:0005615)<br>Inflammatory response (GO:0006954)<br>*Cardiac differentation and maturation* |
| SPI1 | SPI1_HUMAN | 10 | Nucleus (GO:0005634)<br>Macromolecule metabolic process (GO:0043170)<br>*Activation of macrophages or B-cells*<br>*Pre-mRNA splicing* |
| IRF1 | IRF1_HUMAN | 12 | Antigen processing and presentation (GO:0019882)<br>Immune response (GO:0006955)<br>*Represses type I IFN and*<br>*IFN-inducible MHC class I genes* |
| IRF2 | IRF2_HUMAN | 7 | Response to biotic stimulus (GO:0009607)<br>Anatomical structure morphogenesis (GO:0009653)<br>*(same as IRF1)* |
| Pbx | PBX1_HUMAN | 19 | Embryonic development (GO:0009790)<br>Glucuronosyltransferase activity (GO:0015020)<br>*Role in steroidogenesis,*<br>*sexual development and differentiation* |

Table 3: **Subjective analysis of TF-GO term predictions for JASPAR Human TFs (part I).** The table shows results for the 21 Jaspar motifs for Human TFs for which the Avg-Odds Global method predicts 25 or fewer TF-GO associations at an $E$-value cutoff of 0.05. For each JASPAR motif, we show the UniProt name of the TF gene, the number of predicted GO terms, and up to two predicted GO terms and the UniProt annotation for the TF that (subjectively) matches them. GO terms with extremely low $E$-values are also shown, even if no UniProt annotation matches them.

| *JASPAR motif* | *Factor (UniProt)* | *Number of GO terms* | *Example of predicted* GO terms *Transcriptional target keywords* |
|---|---|---|---|
| RUNX1 | RUNX1_HUMAN | 2 | Immune response (GO:0006955)<br>Immune system process (GO:0002376)<br>*T-cell receptor enhancers*<br>*Development of normal hematopoiesis* |
| FOXF2 | FOXF2_HUMAN | 10 | Anatomical structure morphogenesis (GO:0009653)<br>Anatomical structure development (GO:0048856)<br>*Lung-specific genes* |
| PPARG-RXRA | RXRA_HUMAN<br>PPARG_HUMAN | 3 | Lipid transporter activity (GO:0005319)<br>Extracellular region (GO:0005576)<br>*Retinoic acid response pathway*<br>*Beta-oxidation pathway of fatty acids* |
| PPARG | PPARG_HUMAN | 12 | Regulation of metabolic process (GO:0019222)<br>Extracellular region (GO:0005576)<br>*Hypolipidemic drugs and fatty acids*<br>*Adipocyte differentiation and glucose homeostasis* |
| TP53 | P53_HUMAN | 2 | Inflammatory response (GO:0006954)<br>Defense response (GO:0006952)<br>*Tumor suppressor, apoptosis induction* |
| MYC-MAX | MYC_HUMAN | 6 | Cell-cell adhesion (GO:0016337)<br>Proteinaceous extracellular matrix (GO:0005578)<br>*Transcription of growth-related genes* |
| NR2F1 | COT1_HUMAN | 12 | Regulation of transcription, DNA-dependent (GO:0006355)<br>Vitamin binding (GO:0019842)<br>*Initiation of transcription* |
| FOXD1 | FOXD1_HUMAN | 21 | Anatomical structure morphogenesis (GO:0009653)<br>Activation of plasma proteins during inflam... (GO:0002541)<br>*Development retina, the optic chiasm*<br>*and morphogenesis of the kidney* |
| RORA1 | RORA_HUMAN | 13 | Response to xenobiotic stimulus (GO:0009410)<br>Transcription regulator activity (GO:0030528)<br>*Organogenesis and differentiation*<br>*(source NCBI entry NP_599023)* |
| TAL1-TCF3 | TFE2_HUMAN | 14 | Multi-organism process (GO:0051704)<br>Defense response to bacterium (GO:0042742)<br>*Tissue-specific cell fate during embryogenesis*<br>*Muscle or early B-cell differentiation* |
| FOXI1 | FOXI1_HUMAN | 1 | GTPase regulator activity (GO:0030695)<br>*Hearing, sense of balance and kidney function*<br>*Epithelium of distal renal tubules* |

Table 4: **Subjective analysis of TF-GO term predictions for JASPAR Human TFs (part II).**

| *Motif* | *Bits* | *Genes w/terms* | *Target genes* | *AUC50* | | | |
|---|---|---|---|---|---|---|---|
| | | | | *Avg-Odds-GO* | *Avg-Odds* | *Max-Odds* | *Change* |
| ABF1 | 13.98 | 4703 | 199 | 0.07 | 0.06 | 0.05 | 0.00 |
| ADR1 | 6.84 | 426 | 15 | 0 | 0 | 0 | 0.00 |
| AFT2 | 7.94 | 1331 | 60 | 0.08 | 0.05 | 0.02 | 0.03 |
| ARG80 | 8.42 | 364 | 13 | 0.11 | 0 | 0 | 0.11 |
| ARG81 | 10.74 | 364 | 8 | 0.18 | 0 | 0 | 0.18 |
| BAS1 | 9.77 | 315 | 32 | 0.23 | 0.25 | 0.18 | -0.01 |
| CBF1 | 12.78 | 337 | 167 | 0.15 | 0.19 | 0.17 | -0.04 |
| CHA4 | 9.63 | 1277 | 13 | 0.02 | 0 | 0.11 | 0.02 |
| DIG1 | 8.27 | 118 | 82 | 0.1 | 0.05 | 0.04 | 0.04 |
| FHL1 | 11.32 | 3717 | 95 | 0.19 | 0.18 | 0.15 | 0.01 |
| FKH1 | 11.26 | 286 | 67 | 0.13 | 0.12 | 0.08 | 0.01 |
| FKH2 | 11.08 | 637 | 78 | 0.17 | 0.14 | 0.13 | 0.03 |
| GAL80 | 10.33 | 2896 | 2 | 0.79 | 0.92 | 0.87 | -0.13 |
| GAT1 | 10.57 | 7 | 11 | 0.41 | 0.38 | 0 | 0.03 |
| GCN4 | 10.57 | 422 | 125 | 0.21 | 0.14 | 0.07 | 0.07 |
| GCR2 | 9.21 | 42 | 32 | 0.03 | 0.03 | 0.03 | 0.00 |
| GLN3 | 8.16 | 611 | 44 | 0.03 | 0.13 | 0.06 | -0.10 |
| GZF3 | 10.33 | 7 | 3 | 0.18 | 0.2 | 0 | -0.02 |
| HAP1 | 7.96 | 43 | 72 | 0.09 | 0.05 | 0.04 | 0.04 |
| HAP2 | 8.53 | 510 | 28 | 0.07 | 0.03 | 0.04 | 0.03 |
| HAP4 | 10.87 | 231 | 39 | 0.26 | 0.21 | 0.21 | 0.04 |
| HSF1 | 14.14 | 58 | 32 | 0.61 | 0.59 | 0.56 | 0.02 |
| IME1 | 12.73 | 303 | 4 | 0.06 | 0.1 | 0.37 | -0.04 |
| INO2 | 11.73 | 303 | 27 | 0.26 | 0.27 | 0.22 | -0.01 |
| LEU3 | 12.92 | 13 | 11 | 0.72 | 0.71 | 0.68 | 0.01 |
| MBP1 | 9.15 | 1538 | 107 | 0.19 | 0.15 | 0.1 | 0.04 |
| MCM1 | 15.39 | 540 | 56 | 0.44 | 0.45 | 0.42 | -0.01 |
| MET31 | 9.2 | 62 | 21 | 0.15 | 0.1 | 0 | 0.05 |
| MET32 | 14.54 | 84 | 19 | 0.19 | 0.17 | 0.17 | 0.01 |
| MET4 | 13.21 | 62 | 9 | 0.37 | 0.23 | 0 | 0.14 |

Table 5: **Accuracy of three target gene prediction methods in Yeast (part I).** Table shows the prediction accuracy (AUC50) of target gene predictions using the Avg-Odds-GO, Avg-Odds and Max-Odds scoring methods. Results are shown for each of the 63 Yeast transcription factor binding motifs for which target genes and predicted GO terms are available. Provided is also the information content (Bits), the number of genes that are annotated with any of the predicted GO terms, the number of known target genes (HC-set) of each motif. The final column contains the AUC50 difference between Avg-Odds-GO and Avg-Odds (a positive value means that we gain accuracy by including predicted GO terms).

| *Motif* | *Bits* | *Genes w/terms* | *Target genes* | *AUC50* | | | |
|---|---|---|---|---|---|---|---|
| | | | | *Avg-Odds-GO* | *Avg-Odds* | *Max-Odds* | *Change* |
| MIG1 | 11.91 | 33 | 3 | 0 | 0 | 0 | 0.00 |
| MSN2 | 9.03 | 1365 | 63 | 0.06 | 0.07 | 0.01 | -0.01 |
| MSN4 | 8.77 | 1359 | 49 | 0.04 | 0.08 | 0.04 | -0.04 |
| NDD1 | 9.39 | 918 | 60 | 0 | 0.03 | 0.03 | -0.03 |
| NRG1 | 8.73 | 2834 | 76 | 0.05 | 0.1 | 0.04 | -0.06 |
| PHD1 | 7.22 | 33 | 79 | 0.04 | 0.04 | 0.01 | 0.01 |
| PHO4 | 10.7 | 1781 | 23 | 0.2 | 0.27 | 0.09 | -0.08 |
| RAP1 | 11.72 | 3459 | 79 | 0.04 | 0.04 | 0.05 | 0.00 |
| RCS1 | 7.71 | 178 | 42 | 0.2 | 0.17 | 0.01 | 0.03 |
| RDS1 | 11.54 | 2909 | 10 | 0.3 | 0.28 | 0.1 | 0.02 |
| REB1 | 11.35 | 4703 | 217 | 0.05 | 0.05 | 0.04 | 0.00 |
| RGT1 | 15.64 | 232 | 3 | 0.33 | 0.33 | 0.19 | 0.00 |
| RIM101 | 12.05 | 21 | 2 | 0 | 0 | 0 | 0.00 |
| ROX1 | 7.64 | 232 | 33 | 0.03 | 0.03 | 0.02 | 0.00 |
| RPN4 | 13.84 | 4026 | 67 | 0.32 | 0.31 | 0.3 | 0.01 |
| RTG3 | 9.9 | 364 | 31 | 0.1 | 0.05 | 0.02 | 0.04 |
| SFP1 | 12.16 | 3721 | 17 | 0.16 | 0.13 | 0.07 | 0.04 |
| SKN7 | 6.27 | 3219 | 112 | 0.04 | 0.05 | 0.02 | -0.01 |
| SKO1 | 9.94 | 4703 | 19 | 0.04 | 0.03 | 0 | 0.00 |
| STB2 | 10.38 | 4446 | 8 | 0 | 0 | 0 | 0.00 |
| STB5 | 9.28 | 4703 | 17 | 0.35 | 0.34 | 0.26 | 0.01 |
| STE12 | 8.76 | 502 | 121 | 0.12 | 0.08 | 0.02 | 0.04 |
| STP4 | 20.24 | 3022 | 1 | 0 | 0 | 0 | 0.00 |
| SUM1 | 11.29 | 195 | 49 | 0.09 | 0.06 | 0.06 | 0.03 |
| SUT1 | 5.69 | 2413 | 55 | 0.06 | 0.09 | 0.01 | -0.02 |
| SWI4 | 10.04 | 33 | 97 | 0.14 | 0.15 | 0.05 | -0.01 |
| SWI5 | 5.94 | 4395 | 56 | 0.05 | 0.03 | 0.01 | 0.01 |
| SWI6 | 7.42 | 1605 | 108 | 0.15 | 0.13 | 0.02 | 0.01 |
| THI2 | 17.54 | 20 | 8 | 0.79 | 0.78 | 0.79 | 0.01 |
| TYE7 | 8.32 | 1201 | 39 | 0.13 | 0.1 | 0.06 | 0.02 |
| UGA3 | 10.33 | 3279 | 1 | 0 | 0 | 0 | 0.00 |
| UME6 | 12.22 | 75 | 97 | 0.19 | 0.19 | 0.18 | 0.00 |
| YAP7 | 10.21 | 609 | 66 | 0.17 | 0.18 | 0.15 | 0.00 |
| *Mean* | 10.55 | 1336.4 | 50.5 | 0.17 | 0.16 | 0.12 | 0.01 |

Table 6: **Accuracy of three target gene prediction methods in Yeast (part II).**

| Gene | Rank | | Target of DIG1 in | |
|---|---|---|---|---|
| | Avg-Odds-GO | Avg-Odds-Std | HC-map | LC-map |
| YNL279W | 1 | 2 | * | * |
| YNL280C | 2 | 1 | * | * |
| YCL027W | 3 | 17 | * | * |
| YNL042W-B | 4 | 3 | | * |
| YOL111C | 5 | 19 | | * |
| YGL194C-A | 6 | 4 | | |
| YDR085C | 7 | 35 | * | * |
| YIL015W | 8 | 42 | * | * |
| YPL192C | 9 | 52 | * | * |
| YCR089W | 10 | 80 | | * |
| YGL053W | 11 | 82 | | |
| YLL043W | 12 | 83 | | * |
| YKL116C | 13 | 87 | | |
| YGL095C | 14 | 5 | | * |
| YIL071C | 15 | 6 | | * |
| YDR309C | 16 | 7 | * | * |
| YER068W | 17 | 127 | | * |
| YGR282C | 18 | 8 | | * |
| YFR031C-A | 19 | 9 | | * |
| YAL064W-B | 20 | 10 | | * |

Table 7: **An example target gene prediction for DIG1 using Avg-Odds-GO and Avg-Odds-Standard.** Scores by each method are used to rank each target gene, here sorted by their Avg-Odds-GO score. In addition, memberships of the MacIsaac et al. high-confidence (HC) and low-confidence (LC) maps are noted ('*' indicates membership). Avg-Odds-standard predictions that rank 11-16, 18, and 19-20 do not appear in the Avg-Odds-GO top-20. None of these appear in the 82-gene HC-map.

| Gene | Rank | | Target of REB1 in | |
|---|---|---|---|---|
| | Avg-Odds-GO | Avg-Odds-Std | HC-map | LC-map |
| YCR032W | 1 | 2 | * | * |
| YEL012W | 2 | 3 | | * |
| YOL006C | 3 | 4 | * | * |
| YKL059C | 4 | 5 | * | * |
| YJL106W | 5 | 6 | | * |
| YKL058W | 6 | 7 | * | * |
| YML043C | 7 | 8 | | * |
| YML042W | 8 | 9 | | * |
| YLR162W-A | 9 | 1 | | * |
| YPR035W | 10 | 11 | | * |
| YMR001C | 11 | 14 | | * |
| YMR002W | 12 | 15 | | * |
| YOR349W | 13 | 16 | | * |
| YOR348C | 14 | 17 | | * |
| YHL004W | 15 | 18 | | |
| YOL005C | 16 | 19 | * | * |
| YOL004W | 17 | 20 | * | * |
| YNL079C | 18 | 21 | | * |
| YNL078W | 19 | 22 | | * |
| YPL266W | 20 | 23 | | * |

Table 8: An example target gene prediction for REB1 using Avg-Odds-GO and Avg-Odds-Standard. Scores by each method are used to rank each target gene, here sorted by their Avg-Odds-GO score. In addition, memberships of the MacIsaac et al. high-confidence (HC) and low-confidence (LC) maps are noted ('*' indicates membership). It should be noted that genes ranked 10, 12-13 by Avg-Odds-Standard appear as ranks 253, 255-256, respectively, by Avg-Odds-GO. All three cases appear in the LC-map but not in the 217-gene HC-map.