Table 4. Comparison of our method with COMPAREPROSPECTOR and PHYLOCON on 53 yeast coregulated gene datasets

Tf	~No.	Literature	COMPAREPROSPECTOR	PHYLOCON	Our method
ABF1	210	RTCAYTNNNNACGW	TCGTATAAAGTGATA	TCGTATAAAGTGATA	TCACTTTATACGAA
ACE2	69	TGCTGGT	NULL	NULL	TGCTGGCCCA
AZF1	12	TTTTTCTT	CATTTCTTTTTCTGC	TTTTTCTT	AAAAAGAAAAAAAAA
BAS1	31	TGACTC	AGCTGACTACAACC	TGACTCTG	GACTCTGCCTAA
CAD1	23	TTACTAA	ATGATTAGTAAGCAA	GCACGCGATGCTGACT AATG	TTACTGAAAAAAA
CBF1	11	RTCACRTGA	ATCACGTGACCATCA	TCACGTGACC	GGTCACGTGGTCA
CIN5	109	TTACRTAA	TGATTATGTAATCAT	NULL	NULL
FKH1	122	GGTAAACAA	TTTTTGTTTACATTT	TGTTTAC	AAAAGATAAACAAA
FKH2	112	GGTAAACAA	GAAAAAGGTAAACAA	NULL	GCGGGTAAAT
GAL4	10	CGG(N11)CC	CGGGCGACAGTGCTCCG A	CGGTCAACTGTTGTCC G	TTCGGAACAAGGCGG ACC
GCN4	61	TGACTCAT	GGCAATGACTCATCC	TGACTC	TGACTCAT
GCR1	7	GGCTTCCWC	GAAAAAGTGGAAGCT	GAAGGGAA	GGAAGCCACACAC
GLN3	12	GATAAG	ATGATATGGCTACGC	NULL	GAGAGATAAAA
HAC1	14	KGMCAGCGTGTC	TTAACCAGGCGCGGC	GACAGCG	TGCCACGTAA
HAP4	41	ATTGG YCNNCCAATNANM	CGGACCAATCAGATG	NULL	TTTCTTCCAAT
HAP5	8	CCAAT	AGACTCCATTATTTT	CCAATGAG	TTTTTCCAAT
HSF1	7	CTTCTAGAAG	AGTTATATATTGATG	TTTTCCTT	NULL
INO2	32	ATTTCACATC	TGCCACTCTTCCAAT	NULL	AGCATGTGAAAAC
INO4	27	TATTCATATGC	TTTTCACATGCTGTC	TTTTCACATG	CATGTGAAAAT
LEU3	19	GCCGGAACCGG	GACCGGTACCGGCAT	CGGTACCGGCCT	CGGAACCGGAAGA
MAC1	33	GAGCAAA	AGAAGATAAAGTAAA	GAAAAAA	TTTTTTTTGCTCA
MBP1	96	ACGCGT	GTGTGTATGGGTGTG	AAAAAAGACGCGT	AACGCGTCACGA
MCM1	59	TTACCNAATTNGGAAA *	TTTCCAAATTAGGAA	CCTAATTAGGAAA	CCAAAATAGGAAAT
MET31	18	AAACTGTGG	TCACGTGACTAGCAA	CACGTGATAT(cbf1)	TGTGGCGTA CACGTGAAAAAC (cbf1)

MET32	41	AAACTGTGG	GCACGTGATATTACA (cbf1)	AGCACGAGAAAAAA (cbf1)	AGCACGAGAAAAAA (cbf1)
MET4	9	AAACTGTGG	CCACAACTGTGGCTG	CGTTTCTTTTTT CTTTTTT	CTGTGGCAA GCACGTGAAAA(cbf1)
MOT3	5	(C/A/T)AGG(T/C)A	ACTTAAGAAGACATG	TTATTTTT GCTCGC	GAAGGGAAA
MSN4	28	MAGGGGN	AGATGAACTAAAAAC	NULL	CCCCTGAAAA
NRG1	48	TGTCCCCTAATG TCCCTCATTTC	GGGCCAAGTGCCAAG	NULL	CCCCTCCTCT
PDR1	45	TTCCGCGGAA TCCGTGGA TCCGCGGGA	ACACACCCATACACC (rap1)	TGAAAAATTT GCGATGAG	CACACCCACACACC (rap1)
PDR3	18	TCCGCGGA	TCCACGACAACTGCA	NULL	NULL
PHO4	60	NNVCACGTKBGN	TTCATTTTTGTCACC	TGGTCACACAGCACG GT	TCTTTCTTG
RAP1	106	ACACCCATACATCT	TTACACCCATACATT	NULL	AGATGTATGGGTGT
RCS1	24	AMTGCACCCADTT	TAGTATTAAGCCTCG	CACGTGGCTTA	TTTTTCAACTT
REB1	114	NTTACCCGG	TTGTTACCCGGATTG	TTGTTACCCGGATTG	CCGGGTAACAAAAA
RFX1	16	CTATTGCTGCAAC GTTGCCATGGCG	CGTTGCCATGGCAAC	CATGGCAAC	GCCATGGCAACGGA
RLM1	41	CTATTTATAG	CTATTTTTAGATTAG	NULL	TTTGCCGAG
ROX1	54	YSYATTGTT	TATATAACTTAACTA	NULL	GCGTGGGGTAA
RPH1	7	GTgAAAGTAtGctTACTT TgAC	GCATACCTGGGTGGG	TTGCCCTGA	CGCGCGCTCAGGAG
RPN4	25	GGTGGCAAA	GGGAAACAGGAGGTG	NULL	CGGCTCACAAA
SKN7	41	ATTTGGCYGGSCC	GCCTGGCCCGGCACA	NULL	GCGGCTGGCCA
SKO1	12	TGACGTCA AGTACGTCAT	AATGGCGTTAACGGT	NULL	TGACGTAA
SIP4	7	CCTTTAATCCG CCATTCGGCCG CCGTTCGACCG	TGCATATTACTGTGT	TTTTGTAACCA	ACCCGGAA
SMP1	52	ACTACTAwwwwTAG	GGATGAGTGGGTCTG	AGGACCC	NULL
STE12	45	TGAAACA	ATGTGTTTCAAATTG	TGAAAC	TTGTTTCAAA
SUM1	33	ATTTGTGACactt	ATATTTACTGACACT	ATCAGTAA	TTTGTGTCACT

SUT1	46	AACGCGCAGG AACGCGTGCC ATCGCGCAATT	NULL	NULL	GCGCGGAAAA
SWI4	113	CGCGAAA CACGAAAA	NULL	TTCGCGTCGCGTTT	TTTCGCGT
SWI5	72	ACCAGCA	NULL	NULL	GGTGCGGGTA
SWI6	111	CGCGAAA CACGAAAA	NULL	GACGCGT CGCGAAA	ACGCGTCGCGA GCGTCGCGAAA
UME6	101	TGCCGCCGA TAGCCGCCGA	ACTTCGGCGGCTAAA	CCTCGGCGGCTAA	CCTCGGCGGCTAA
YAP1	49	NTGASTCAG actcTTAGTAAagga	TGATTAGTAATCATA	TTAGTCAGCATC	GCCGCTACTAAA
ZAP1	15	ACCCTAAAGGT	CATAACCTTTAGGGT	ACCTTTAGGGT	CCCTCAAGGTCAAA
53			29+1 correct prediction, 18 wrong prediction	24+1 correct prediction, 10 wrong prediction	41 correct prediction, 8 wrong prediction

Ref. 1 points out that the consensus site of CIN5 is TTACRTAA. We use TTACRTAA here instead of TRANSFAC one TTACTAA.

1. All transcription factors with known sites in ref. 1 and with experimentally verified sites in TRANSFAC (we can find the sites in TRANSFAC or the paper cited by TRANSFAC) will be used. But if the binding sites provided by TRANSFAC are too long or contradict with each other, we will not use them. By this criterion, we obtained the 53 transcription factors (Tf) tabulated above.

2. The criterion for matching with TRANSFAC motifs is that there should be at most one mismatch in the orange regions. Those orange regions must be continuous except the

ambiguous positions, such as SIP4 (colored orange according to ref. 1). The length of the orange parts must be at least 6 unless the motif is shorter than 6.

3. We know SWI4 works with SWI6, and MET4 works with MET31 and MET32. So if we find any sites matching either of them for the corresponding transcription factors, we will claim matching.

4. We output 10 motifs from COMPAREPROSPECTOR and our methods. For PHYLOCON, we manually check all the output motifs (on average, there are 60 output motifs for every transcription factor. Although there are some motifs appearing a few times in the output, we still give PHYLOCON a great advantage here, which may result in its high specificity). For our method, the motifs in Table 4 are the best motif in our output in more than 80% of the time.

5. The NULL in the last three columns means the outputs contain only AAAAA or TTTTT or TATATA- or CACACA-like segments. In the AAAA and TTTTT cases, the outputs must have more than 90% of A or T. In the CACACA and TATATA cases, all of them must be CA or TA except the two boundary nucleotides. For those outputs as NULL, we treat them as no prediction.

6. Ambiguity Codes: D = A or T; B = C, G ,or T; S = C or G; W = A or T; R = A or G; Y = C or T; K = G or T; M = A or C; V = A, C, or G; and N = A, C, G, or T.

7. The 29+1 in the last row of the table means there are 29 motifs output from COMPAREPROSPECTOR satisfying the criteria of matching defined above and there is another motif (SUM1) output from COMPAREPROSPECTOR that does not satisfy the criteria of matching defined above. 24+1 in the last row of the table has a similar meaning.

8. The true-positive rates of the tree Gibbs sampler, COMPAREPROSPECTOR, and PHYLOCON are 41/53 = 77.4%, 30/53 = 56.6%, and 25/53 = 47.2%, respectively. The false-positive rates of tree Gibbs sampler, COMPAREPROSPECTOR, and PHYLOCON are 8/49 = 16.3%, 18/48 = 37.5%, and 10/35 = 28.6%, respectively. Although we may not be so good as the authors at tuning parameters in using PHYLOCON and COMPAREPROSPECTOR, our method is better.

1. Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al*. (2004) *Nature* **431,** 99-104.