Supplementary material for "A dynamic Bayesian network for identifying protein binding footprints from single molecule based sequencing data"

Xiaoyu Chen [1], Michael M. Hoffman [2], Jeff A. Bilmes [3],
Jay R. Hesselberth [2*] and William S. Noble [2,1*]

[1]Department of Computer Science and Engineering,
[2]Department of Genome Sciences,
[3]Department of Electrical Engineering,
University of Washington,
Seattle, WA, USA

January 8, 2010

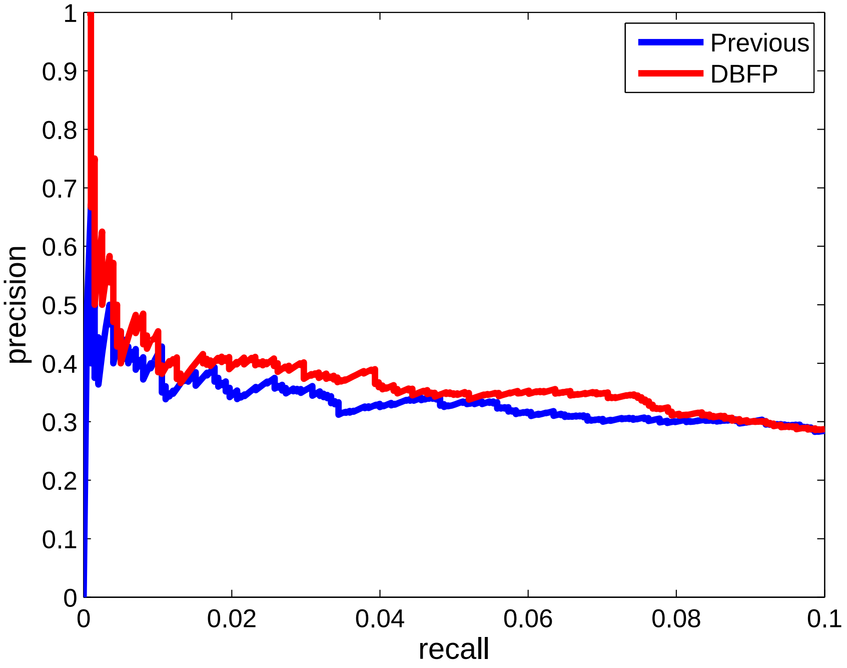


Figure 1: **Precision-recall curves for footprint detection.** The figure plots the precision-recall curves (up to a recall of 0.1) for DBFP and our previously described method, using the *MacIsaac binding sites* as the gold standard.

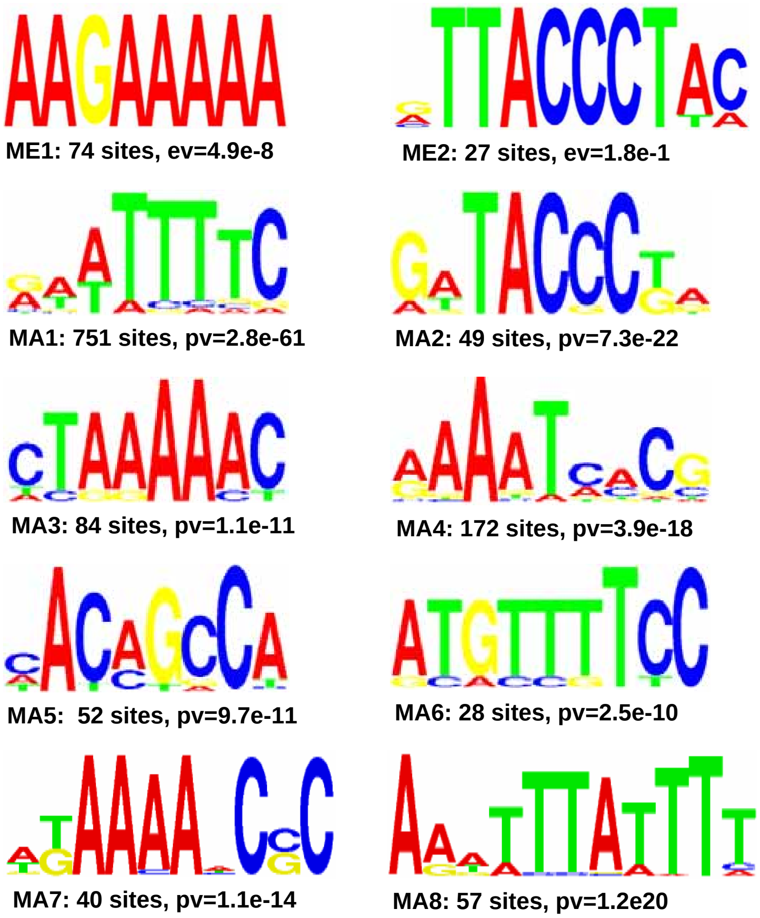
Searching for novel motifs

In order to search novel motifs in our identified footprints, we considered only the footprints that do not overlap any known motif, using a scanning threshold of q value < 0.05. We ran the MEME motif discovery algorithm [1] on the resulting collection of 2,955 footprints, using the ZOOPS (zero-or-one occurrence per sequence) model and masking out low complexity sequence. MEME found two motifs at an E-value threshold of 1.0. The motif logos are shown in Figure 2.

We also ran the Amadeus motif discovery program [3] on the set of 2,955 footprints (foreground sequences) and a set of 14,775 background sequences that were randomly selected from the *background* and *hypersensitive* segments. Because Amadeus can only be run each time for one particular motif width between 6 and 12, the motifs that are detected with varied width from different runs may be redundant. We first collected 20 motifs returned by Amadeus with width ranging from 8 to 12. Next, we used Tomtom [2] to compare these 20 motifs and divided them into subsets. Two motifs will be put in the same subset if Tomom detects significant similarity between them. We then chose the most significant motif (i.e. the one of the lowest p value reported by Amadeus) from each subset. We obtained eight subsets and therefore eight distinct motifs in the end. The logos of these motifs are also shown in Figure 2.

We compared the nine motifs detected by MEME and Amadeus to the known MacIsaac/Zhu/TRANSFAC motifs using Tomtom. Only two motifs, ME1 and MA1, match significantly with known motifs. Motif ME1 is similar to two TRANSFAC motifs of Ste11, while motif MA1 is similar to some T-rich known motifs, including MacIsaac motifs Mcm1 and Stb1, and Zhu motifs Sfp1, Stb3, and Ypr196w. The matching between these two motifs and the known motifs provides evidence that a significant percentage of the footprints contain weak sequence motifs that were simply not identified by FIMO.

For motifs ME2 and MA2, we observe that their cores are similar to MacIsaaac motif Reb1, but with differences at the flanking positions. Positions 2-7 of ME2 highly resemble positions 1-6 of Reb1, but the eighth position of ME2 contains a nucleotide (T) that is much less frequent at the seventh position of Reb1; positions 3-7 of MA2 match well with positions 2-6 of Reb1, but the second and the eighth position of ME2 are more degenerate than the first and the seventh position of Reb1, respectively. To



ME1: 74 sites, ev=4.9e-8

ME2: 27 sites, ev=1.8e-1

MA1: 751 sites, pv=2.8e-61

MA2: 49 sites, pv=7.3e-22

MA3: 84 sites, pv=1.1e-11

MA4: 172 sites, pv=3.9e-18

MA5: 52 sites, pv=9.7e-11

MA6: 28 sites, pv=2.5e-10

MA7: 40 sites, pv=1.1e-14

MA8: 57 sites, pv=1.2e20

Figure 2: **Motifs discovered from footprints that do not overlap any known motifs.** Motifs detected by MEME have a name starting with *ME*, while those detected by Amadeus have a name starting with *MA*. For each motif, the number of sites used to construct that motif is listed; its significance score is also listed (E-value for MEME motifs and p-value for Amadeus motifs).

further compare ME2 and MA2 with Reb1, we searched the 4,679 identified footprints for occurrences of these three motifs. The comparison of motif sites are described in detail as follows:

- At a scanning q value threshold 0.05, Reb1 has 551 occurrences, while ME2 and MA2 have 28 and 132 occurrences, respectively. None of the ME2 sites overlaps a Reb1 site, whereas MA2 has 103 sites overlapping a Reb1 site. Moreover, ME2 and MA2 have 11 overlapping sites.

- For each of the three motifs, we calculated the distribution of distances from motif sites to their closest TSSs. All three motifs are mostly enriched around 75-100 bp upstream of the TSSs (data not shown).

- We also used g:Profiler [4] to analyze the GO term enrichment in the set of downstream genes of each motif (i.e. the genes located at most 250 bp away from a site of that motif). ME2 has 19 downstream genes which are enriched for two GO terms, *BP:localization* and *CC:organelle membrane*. MA2 has 91 downstream genes, and these genes are enriched for several GO terms, such as *BP:intracellular transport*, *BP:establishment of localization*, *CC:intracellular part*, *CC:protein complex*, and *MF:protein transporter activity*. Most of the enriched terms for ME2 and MA2 are also shared by Reb1.

In conclusion, both ME2 and MA2 are similar to Reb1 in various ways. We therefore conjecture that ME2 and MA2 represent respectively an alternative form of the Reb1 motif.

The other six motifs detected by Amadeus (i.e., MA3-MA8) are candidate novel binding motifs, for which Tomtom detected no significant similarity to any known motif.

References

[1] T. Bailey, N. Williams, C. Misleh, and W. Li. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web server issue):W369–W373, 2006.

[2] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble. Quantifying similarity between motifs. *Genome Biology*, 8:R24, 2007.

[3] C. Linhart, Y. Halperin, and R. Shamir. Transcription factor and microrna motif discovery: the amadeus platform and a compendium of metazoan target sets. *Genome Res*, 18(7):1180–1189, Jul 2008.

[4] J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*, 35(Web Server issue):W193–W200, Jul 2007.