Published in final edited form as:

Nat Biotechnol. ; 29(6): 483–484. doi:10.1038/nbt.1892.

Jury remains out on simple models of transcription factor specificity

Quaid Morris^{1,2,3}, Martha L. Bulyk^{4,5,6,7}, and Timothy R. Hughes^{1,2}

¹The Donnelly Centre, University of Toronto, 160 College St., Toronto, ON, Canada M5S 3E1

²Department of Molecular Genetics, University of Toronto, 160 College St., Toronto, ON, Canada M5S 3E1

³Department of Computer Science, University of Toronto, 160 College St., Toronto, ON, Canada M5S 3E1

⁴Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

⁵Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

⁶Harvard -MIT Division of Health Sciences and Technology (HST); Harvard Medical School, Boston, MA 02115

⁷Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138

In this issue, Zhao and Stormo introduce a new method for deriving position weight matrices (PWMs) from Protein Binding microarrays (BEEML-PBM). Using this method, they challenge a central claim of our 2009 paper¹, by “[concluding] that the widespread phenomenon of secondary binding preference identified by Badis *et al.* is not supported by the data and is likely due to suboptimal estimation of the PWM.” BEEML-PBM is simple, elegant and corrects for a pronounced positional effect of TF binding in the PBM assay; however we do not agree with their overall conclusion and believe that it is based on incomplete and biased analysis of our data. Zhao and Stormo's conclusions are based on comparing the performance of BEEML-PBM PWMs and our methods on held-out data. However, they over-estimate the performance of their PWMs and under-estimate the performance of our methods.

First, their claims of suboptimality of our PWMs are based on results from only one of the three motif finders that we employed, Seed-and-Wobble (SnW). SnW was not developed to predict probe intensities and does not attempt to produce a summary PWM that optimizes performance over all probes in predicting probe intensities. Instead, it was developed for the purpose of summarizing the 8-mer data, seeding with the highest scoring 8-mer, in a compact way for use in visual depiction as sequence logos. In contrast, another of the methods we employed, RankMotif++, is designed to produce summary PWMs and we have previously reported² that it, like BEEML-PBM, better predicts probe intensities than SnW. So we suspect its performance would be much more competitive. In fact, RankMotif++ is very similar to the BEEML base method³; it fits a PWM model using a regression-like

procedure to optimally predict PBM intensity data. RankMotif++ differs from BEEML primarily in that it regresses on a partial preference ordering of probes inferred from their PBM intensities rather than on their actual intensities themselves. We acknowledge that comparisons with RankMotif++ PWMs would have been difficult because, although the source code for RankMotif++ has been available for three years, the PWMs we learned for Badis et al. were until recently only available as sequence logos. However, we made the motifs available to Zhao and Stormo when we were notified of this oversight and before the final submission of their paper. The motifs are available here: http://the_brain.bwh.harvard.edu/suppl105/.

Second, we note that Zhao and Stormo use a positional effect model when training their PWMs but do not allow the methods that they are comparing against the same opportunity to correct this bias during training. We propose that this correction is a major cause of BEEML-PBM's success and that both the multiple PWM methods and the 8-mer affinity estimates we employed would greatly benefit from a similar correction, thus restoring our reported gain in performance. For example, the 8-mer median intensities used in Figure 2A are not corrected for positional biases and this leads to the counter-intuitive claim that for the 15–20 (of 41) data points that lie above the diagonal, BEEML-PBM PWMs capture more than 100% of the replicate reproducibility. A more appropriate comparison would either employ PWMs uncorrected for positional bias (as we did in our original paper) or to compare against similarly corrected 8-mer median intensities. Zhao and Stormo do neither and as such, we believe that their prediction accuracy estimates are inflated.

Finally, we note that explaining 90% of the reproducible binding signal is not the same as explaining 100%, and proteins that we and others have confirmed have multiple binding modes do not satisfy Zhao and Stormo's 90% cut-off. For example, we reported that Jundm2 (Jdp2) binds two half-sites with variable spacing between them; this is clearly observed in the top-scoring 8-mers¹. This mode of binding is common among other bZIP proteins. Furthermore, Zhao and Stormo do not consider the PBM data for Bcl6b, a C2H2 zinc finger for which we obtained two very different PWMs; these are also clearly observed in the top-scoring 8-mers, and moreover enrichment for motif matches can be observed in associated ChIP-chip data¹. In general, variable spacing in long C2H2 zinc finger array seems to be common; for example, ChIP-seq for REST also supports use of partial vs. full sites, and different spacings⁴. Single summary PWMs cannot capture these binding modes, and it is important to do so, as C2H2 zinc fingers are the most common domain in metazoa, and long arrays of these domains are common in human and mouse genomes.

We agree that simple and accurate representation of TF sequence specificity on the basis of PBM data is an important problem. We ourselves have been working on extensions to our algorithms to capture the PBM positional and orientation effects (which we have previously reported⁵). We also recently conducted a DREAM competition in which the goal was to predict PBM probe intensities using a two-array framework and evaluation criteria similar to those employed in¹ and Zhao et al. A manuscript describing these new data, the competition, the methods of ~20 groups, their evaluation, and a web site that allows benchmarking any method to the DREAM results is in preparation (M. Weirauch, G. Stolovitsky et al.). We have now obtained the BEEML-PBM code, and we look forward

References

1. Badis G, et al. Diversity and complexity in DNA recognition by transcription factors. *Science*. 2009; 324:1720–1723. [PubMed: 19443739]
2. Chen X, Hughes TR, Morris Q. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics*. 2007; 23:i72–79. [PubMed: 17646348]
3. Zhao Y, Granas D, Stormo GD. Inferring binding energies from selected binding sites. *PLoS Comput Biol*. 2009; 5:e1000590. [PubMed: 19997485]
4. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007; 316:1497–1502. [PubMed: 17540862]
5. Berger MF, et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*. 2006; 24:1429–1435. [PubMed: 16998473]