Saturn team write-up

We used xgboost (T. Chen and Guestrin 2016) to fit a gradient tree boosting model to a set of features derived from the called DNase peaks and a genome wide scan for known motifs and *de novo* motifs. We used the DGF method Wellington to create new features of just those binding sites overlapping a footprint. We found that using a kernel smoother on our predictions increased our cross-validation performance.

Features

The most predictive feature we used were the p-values of the relaxed DNase peaks.

We compiled a list of motifs that represented the binding specificities of all the TFs in the challenge from the following public motif databases:

- Human and Mouse high-throughput SELEX motifs from (Jolma et al. 2010)
- Human and Mouse HT-SELEX motifs from (Jolma et al. 2013)
- The Swiss Regulon human and mouse motifs (Pachkov et al. 2013)
- From (Zhao and Stormo 2011)
- JASPAR CORE (Mathelier et al. 2016)
- Direct and inferred motifs for *Homo sapiens* from (Weirauch et al. 2014)

We then perfomed a genome-wide scan of all these motifs using the STEME software (Reid and Wernisch 2014) to generate a set of putative binding sites. These were summarised as region-level features by the maximum of their log-odds ratio.

We used a discriminative motif finder, DREME (Bailey 2011), to find motifs that discriminated between the bound sequences and those that were unbound for each TF. In exactly the same way as for the known motifs, we perfomed a genome-wide scan and summarised the putative binding sites as region-level features on a per-TF basis.

We used Wellington (Piper et al. 2013), a DNase footprint detection algorithm to determine TF binding footprints in the DNase peaks. We used these to filter both the known motif binding sites and those the *de novo* sites. We included both the filtered binding sites and the unfiltered sites as features.

Predictions

We applied a kernel smoothing method to the predictions. We adjusted the log odds ratio for each region using a Gaussian kernel. We used TF-specific length-scales between 0 and 200 base pairs that we chose using cross-validation.

References

Bailey, T. L. 2011. "DREME: Motif Discovery in Transcription Factor ChIP-Seq Data." *Bioinformatics* 27 (12): 1653–9.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System," March. doi:10.1145/2939672.2939785[1].

Jolma, Arttu, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, et al. 2010. "Multiplexed Massively Parallel SELEX for Characterization of Human Transcription Factor Binding Specificities." *Genome Research* 20 (6): 861–73. doi:10.1101/gr.100552.109[2].

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. "DNA-Binding Specificities of Human Transcription Factors." *Cell* 152 (1–2): 327–39. doi:10.1016/j.cell.2012.12.009[3].

Mathelier, Anthony, Oriol Fornes, David J. Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, et al. 2016. "JASPAR 2016: A Major Expansion and Update of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 44 (D1): D110–D115. doi:10.1093/nar/gkv1176[4].

Pachkov, Mikhail, Piotr J. Balwierz, Phil Arnold, Evgeniy Ozonov, and Erik van Nimwegen. 2013. "SwissRegulon, a Database of Genome-Wide Annotations of Regulatory Sites: Recent Updates." *Nucleic Acids Research* 41 (D1): D214–D220. doi:10.1093/nar/gks1145[5].

Piper, Jason, Markus C. Elze, Pierre Cauchy, Peter N. Cockerill, Constanze Bonifer, and Sascha Ott. 2013. "Wellington: A Novel Method for the Accurate

[1] https://doi.org/10.1145/2939672.2939785

[2] https://doi.org/10.1101/gr.100552.109

[3] https://doi.org/10.1016/j.cell.2012.12.009

[4] https://doi.org/10.1093/nar/gkv1176

[5] https://doi.org/10.1093/nar/gks1145

Identification of Digital Genomic Footprints from DNase-Seq Data.” *Nucleic Acids Research*, September, gkt850. doi:10.1093/nar/gkt850[6].

Reid, John E., and Lorenz Wernisch. 2014. “STEME: A Robust, Accurate Motif Finder for Large Data Sets.” *PLoS ONE* 9 (3): e90735. doi:10.1371/journal.pone.0090735[7].

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. “Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.” *Cell* 158 (6): 1431–43. doi:10.1016/j.cell.2014.08.009[8].

Zhao, Y., and G. D. Stormo. 2011. “Quantitative Analysis Demonstrates Most Transcription Factors Require Only Simple Models of Specificity.” *Nature Biotechnology* 29 (6): 480–83. http://www.nature.com/nbt/journal/v29/n6/abs/nbt.1893.html.

[6] https://doi.org/10.1093/nar/gkt850

[7] https://doi.org/10.1371/journal.pone.0090735

[8] https://doi.org/10.1016/j.cell.2014.08.009