The ENCODE-DREAM Challenge to Predict Genome-Wide Binding of Regulatory Proteins to DNA

Akshay Balsubramani[†*], Nathan Boley[†], James C. Costello[‡], Laura M. Heiser[§], Tim Jeske[¶], Robert Kueffner[£], Jin Wook Lee[†], Rani K. Powers[‡], Anshul Kundaje[†$] [*]

Transcription factors (TFs) are regulatory proteins that bind specific DNA elements in the genome to activate or repress transcription of target genes. The binding landscapes of TFs are highly variable across different cellular contexts. While experimental strategies to profile genome-wide TF binding are widely used, their application is limited to a small fraction of TFs in a few cellular contexts due to material and cost constraints. Hence, accurate and high-resolution computational approaches are necessary to close this gap and complement experimental results [1].

The Encyclopedia of DNA Elements (ENCODE) Consortium [2] has experimentally assayed binding landscapes of 100s of TFs in 10s of cell types. Related assays such as those profiling DNA accessibility and gene expression have also been used to provide complementary information about regulatory genomic activity in the same cell types. These reference datasets provide a unique opportunity to train and uniformly evaluate computational TF binding prediction methods across diverse TFs and cell types/tissues. We organized the ENCODE-DREAM Transcription Factor Binding Site Prediction Challenge to evaluate the state of the art for these prediction tasks openly and reproducibly.

The challenge evaluated TF binding prediction as a multitask binary classification problem, with each task being a (TF, cell type) pair. On each task, participants were required to predict the probability of a binding event of the TF in the cell type at locations (200-base intervals) across the entire genome. The ground-truth labels were derived from high quality chromatin immunoprecipitation sequencing (ChIP-seq) experiments, which provide a genome-wide track of binding enrichment scores for each TF. We used statistical methods based on standardized ENCODE pipelines [3] to identify high-confidence binding events across the genome, resulting in a binary genome-wide track indicating whether each location in the genome is bound or unbound by the TF.

Teams participating in the challenge used multiple modalities of high-quality data from genome-wide biochemical assays to classify genomic locations as TF-bound or TF-unbound across tissues and TFs. The data modalities reflect biochemical mechanisms involved in TF binding.

Sequence. The DNA sequence (DNA-seq) is one of the primary determinants of binding, wherein regions of sequence are recognized by specific TFs [4].
DNA accessibility. The assay DNase-seq identifies regions where DNA is less compacted and more accessible for TF binding, in a tissue-specific manner [5].
DNA shape. Twisting, bending, and shearing of DNA influence binding in a TF-specific fashion [6].
Gene expression. Expression levels of all human genes are provided using RNA-seq results from ENCODE, for possible inference of e.g. cofactor proteins that recruit TFs for binding [7].
Binding in other tissues. ChIP-seq data are provided for a number of training TFs and tissues.

Prediction and evaluation were done across the whole genome, on 200-base (bp) windows every 50 bp, to incorporate the effects of local spatial context on TF binding. This yields 60,519,747 windows (data points); 8,843,011 were from preselected *validation chromosomes* used solely for validation, with labels hidden from participants. The quantity of data exceeds 500 GB. 40 international teams participated in the challenge, including developers of some well-used binding prediction methods [8].

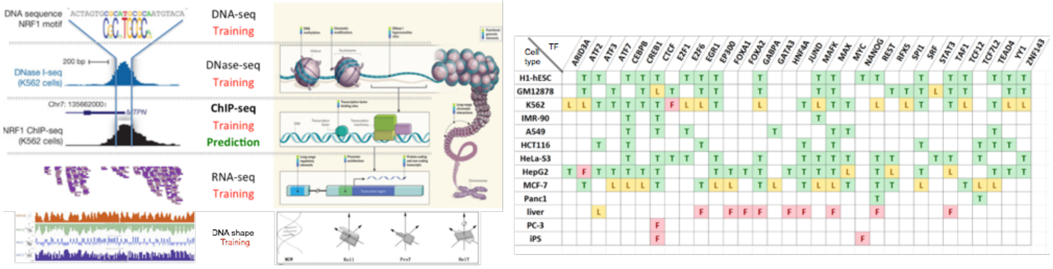
[*] Corresponding author. Email address: {abalsubr, akundaje}@stanford.edu. †Dept. of Genetics, Stanford University. ‡Dept. of Pharmacology, University of Colorado. §Dept. of Biomedical Engineering, Oregon Health and Science University. ¶Dept. of Pediatrics, Ludwig Maximilians University. £Dept. of Genetics, Icahn School of Medicine at Mount Sinai. $ Dept. of Computer Science, Stanford University.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Details

Design. The challenge was carefully conducted to keep the participants blind to the validation data and benchmark the effect of cross-tissue prediction relative to prediction within the same cell type (i.e., domain adaptation). We used a set of previously unpublished ChIP-seq experiments that have been performed across a set of different tissues and TFs. The ChIP-seq data were partitioned into training, leaderboard, and final-round tasks (Fig. 1) corresponding to different phases of the challenge. Initially, only ChIP-seq data from the training tasks was made public for the leaderboard round, in which participants were allowed to test their methods on each leaderboard task with a limited number of submissions whose performance metrics were updated live on a public leaderboard. This was followed by the scoring phase of the challenge, in which the participants submitted cross-tissue predictions on the final-round tasks for evaluation to produce the challenge results. All final-round data except for the validation labels was then released, and teams again submitted predictions for the final-round tasks with this more predictive within-tissue data.

Evaluation. Computational ChIP-seq positives (putative bound sites) typically require follow-up whole-genome biological validation experiments to be useful in a biological context. This has two consequences motivating evaluation in our challenge. First, the negative "background" set was drawn from the whole genome, even though TFs only bind to a tiny fraction of the genome, leading to highly imbalanced classification problems – the $\frac{\text{negative}}{\text{positive}}$ ratio ranged from 250 to 1200 across tasks. Second, the recall at low FDR was a significant metric in our evaluation, [2] since assays following up on false positives represent significant wasted effort.



TF Cell type	ARID3A	ATF2	ATF3	ATF7	CEBPB	CREB1	CTCF	E2F1	E2F6	EGR1	EP300	FOXA1	FOXA2	GABPA	GATA3	HNF4A	JUND	MAFK	MAX	MYC	NANOG	REST	RFX5	SPI1	SRF	STAT3	TAF1	TCF12	TCF7L2	TEAD4	YY1	ZNF143
H1-hESC		T	T		T	T	T		T	T	T			T				T	T		T	T			T		T	T		T	T	T
GM12878		T		T		T	L	T		T	T			T				T	T				T	T	T	L	T	T			T	T
K562	L	L	T	T	T	T	T	F	L	L	T			L			T	L	T	T		L		L	T		T	L		T	L	L
IMR-90					T		T											T														
A549					T		T		T						T				T	T										T		
HCT116			T		T					T							T		T						T				T	T	T	
HeLa-S3					T		T	T	T		T			T			T	T	T	T		T	T			T	T		T			T
HepG2	T	F	T	T	T	T	T				T	T	T	T		T	T	T	T	L		T	L		T		L			T	T	T
MCF-7		T		L	L	L	T			T	L	L		T	L		T	L	L	T		T	T		L			T	L	L		
Panc1																						T							T			
liver			L							F		F	F	F		F	F		F			F					F					
PC-3							F																									
iPS							F														F											

Figure 1: *Left*: The challenge used different data modalities corresponding to various biochemical determinants of binding. *Right*: Data from 32 TFs and 13 cell types were used in the challenge (**T**raining, **L**eaderboard, **F**inal-round), with final evaluation on 13 tasks (12 TFs, 4 cell types).

Results. By common consensus, top teams faced some ubiquitous issues which were reported to be very significant to final-round performance. Cross-cell-type prediction was a pervasive domain adaptation difficulty, with top teams scoring an average of $> 10\%$ higher in auPRC on the within-tissue benchmarks. The top teams employed various strategies to avoid overreliance on cell-type-independent *sequence* features, as well as iterative data sampling strategies to attempt to correct for the covariate shift between cell types and class imbalance. The domain adaptation and sampling methods used by top teams were heuristic, and the teams reported room for improvement in this area.

Summary

We are analyzing the challenge predictions from a computational biology perspective, and developing an open benchmarking engine incorporating the design and evaluation insights gained during the challenge. The 40 teams participating in the challenge did not generate predictions of sufficient quality to supplant biological assays for downstream genome-wide analysis on most TFs assayed. But the wide variation in performance across tasks (from 25% to 80% auPRC) suggest significant performance improvements still to be made with high biological impact. These would involve basic topics in learning research: multitask information sharing (between the challenge's 41 prediction tasks) and domain adaptation (between the 13 cell types). We hope the challenge data and analysis help catalyze such work, and standardize method development for genome-wide TF binding prediction.

[2] We aggregated rankings of participants according to auROC, auPRC, recall@FDR10, and recall@FDR50 to arrive at the final challenge results.

Acknowledgments

We would like to thank the participating teams, as well as the DREAM challenge network [9] and the Synapse portal [10] for hosting challenge submissions and data.

References

[1] Mark B Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, et al. Architecture of the human regulatory network derived from encode data. *Nature*, 489(7414):91–100, 2012.

[2] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[3] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831, 2012.

[4] Gary D Stormo and Yue Zhao. Determining the specificity of protein-dna interactions. *Nature Reviews Genetics*, 11(11):751, 2010.

[5] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75, 2012.

[6] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein-dna recognition. *Nature*, 461(7268):1248, 2009.

[7] Mark Ptashne and Alexander Gann. Transcriptional activation by recruitment. *Nature*, 386(6625):569, 1997.

[8] Matthew T Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–134, 2013.

[9] Dream challenges. `http://dreamchallenges.org/`, 2017.

[10] Synapse | sage bionetworks. `https://www.synapse.org/`, 2017.