CSE 590 C
Comp Bio Seminar

Organizational Meeting

10/5/2020

https://courses.cs.washington.edu/courses/cse590c/20au/

Larry Ruzzo

1 MB Walter Costa, C Höner Zu Siederdissen, M Dunjić, PF Stadler, K Nowick, "SSS-test: a novel test for detecting positive selection on RNA secondary structure." *BMC Bioinformatics*, 20, #1 (2019) 151.

2a C Robert, M Watson, "The incredible complexity of RNA splicing." *Genome Biol*, 17, #1 (2016) 265.

2b A Nellore, AE Jaffe, JP Fortin, J Alquicira-Hernández, L Collado-Torres, S Wang, RA Phillips, N Karbhari, KD Hansen, B Langmead, JT Leek, "Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive." *Genome Biol*, 17, #1 (2016) 266.

3a K Choi, Y Chen, DA Skelly, GA Churchill, "Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics." *Genome Biol*, 21, #1 (2020) 183.

3b TH Kim, X Zhou, M Chen, "Demystifying "drop-outs" in single-cell UMI data." *Genome Biol*, 21, #1 (2020) 196.

Sara Mostafavi

4. Avsec, Weilert, Shrikumar, Krueger, Alexandari, Dalal, Fropf, Mc Anany, Gagneur, Kundaje, Zeitlinger "Base-resolution models of transcription factor binding reveal soft motif syntax", BioRxiv, 2020.

The arrangement of transcription factor (TF) binding motifs (syntax) is an important part of the cis-regulatory code, yet remains elusive. We introduce a deep learning model, BPNet, that uses DNA sequence to predict base-resolution ChIP-nexus binding profiles of pluripotency TFs. We develop interpretation tools to learn predictive motif representations and identify soft syntax rules for cooperative TF binding interactions. Strikingly, Nanog preferentially binds with helical periodicity, and TFs often cooperate in a directional manner, which we validate using CRISPR-induced point mutations. Our model represents a powerful general approach to uncover the motifs and syntax of cis-regulatory sequences in genomics data.

Yuliang Wang

5. BH Hristov, B Chazelle, M Singh, "uKIN Combines New and Prior Information with Guided Network Propagation to Accurately Identify Disease Genes". *Cell Syst*, 10, #6 (2020) 470-479.e3.

   This paper described a novel guided network propagation approach to prioritize putative disease genes using protein-protein interaction networks. This method uses known disease genes to guide random walks initiated at newly implicated genes, which allow for network-based integration of prior and new data. The authors showed its effectiveness for both cancer genomics data and GWAS data.

- Free all Mondays except the next one (Oct 12th).

Bill Noble

6. Z He, A Brazovskaja, S Ebert, JG Camp, B Treutlein, "CSS: cluster similarity spectrum integration of single-cell genomics data." *Genome Biol*, 21, #1 (

   It is a major challenge to integrate single-cell sequencing data across experiments, conditions, batches, time points, and other technical considerations. New computational methods are required that can integrate samples while simultaneously preserving biological information. Here, we propose an unsupervised reference-free data representation, cluster similarity spectrum (CSS), where each cell is represented by its similarities to clusters independently identified across samples. We show that CSS can be used to assess cellular heterogeneity and enable reconstruction of differentiation trajectories from cerebral organoid and other single-cell transcriptomic data, and to integrate data across experimental conditions and human individuals.2020) 224.

Sheng Wang

7. Rao, Bhattacharya, Thomas, Duan, Chen, et al. (2019). Evaluating protein transfer learning with TAPE. Adv. in Neural Info. Processing Systems (pp. 9689-9701).

   Machine learning applied to protein sequences is an increasingly popular area of research. Semi-supervised learning for proteins has emerged as an important paradigm due to the high cost of acquiring supervised protein labels, but the current literature is fragmented when it comes to datasets and standardized evaluation techniques. To facilitate progress in this field, we introduce the Tasks Assessing Protein Embeddings (TAPE), a set of five biologically relevant semi-supervised learning tasks spread across different domains of protein biology. We curate tasks into specific training, validation, and test splits to ensure that each task tests biologically relevant generalization that transfers to real-life scenarios. We bench- mark a range of approaches to semi-supervised protein representation learning, which span recent work as well as canonical sequence learning techniques. We find that self-supervised pretraining is helpful for almost all models on all tasks, more than doubling performance in some cases. Despite this increase, in several cases features learned by self-supervised pretraining still lag behind features ex- tracted by state-of-the-art non-neural techniques. This gap in performance suggests a huge opportunity for innovative architecture design and improved modeling paradigms that better capture the signal in biological sequences. TAPE will help the machine learning community focus effort on scientifically relevant problems.

Chris Thachuk

8. AJ Simon, S d'Oelsnitz, AD Ellington, "Synthetic evolution." *Nat Biotechnol*, 37, #7 (2019) 730-743.

   The combination of modern biotechnologies such as DNA synthesis, λ red recombineering, CRISPR-based editing and next-generation high-throughput sequencing increasingly enables precise manipulation of genes and genomes. Beyond rational design, these technologies also enable the targeted, and potentially continuous, introduction of multiple mutations. While this might seem to be merely a return to natural selection, the ability to target evolution greatly reduces fitness burdens and focuses mutation and selection on those genes and traits that best contribute to a desired phenotype, ultimately throwing evolution into fast forward.

Fall Schedule

- Larry -

  1 RNA Evolution

  2 Splicing

  3 sc dropout

4 Sara - Kundaje

5 Yuliang - NetProp

6 Bill – sc clustering

7 Sheng – prot “TAPE”

8 Chris – synth evo

10/12	#8, Jason
10/19	#3, Ayse
10/26	#4, Joe
11/2	#7, Nicasia
11/9	#5, Alyssa
11/16	
11/23	
11/30	
12/7	

Volunteers Still Needed!

Students – Please email ruzzo@uw.edu & I’ll put you in touch with faculty for your paper(s).