Creative probabilistic programming for biology

MIRIAM SHIFFMAN shiffman@{mit.edu,broadinstitute.org}

The "meaningfulness" of a learned representation in biology can only be measured with respect to a particular biological context or question. Modeling is the structure that provides this context and endows latent representations with meaning. *Probabilistic* modeling is often the most suitable choice — not only for its decision theoretic properties and coherent handling of measurement noise, but because biology itself is probabilistic. And *probabilistic programming languages* are one tool missing from widespread adoption in biology, with the potential to more naturally and holistically meld the modeling process with the process of wet lab science.

Probabilistic programming languages (PPLs) add random variables to the long list of built-in types that we expect in a language (strings, ints, and the like). Fundamental operations in probability — like sampling, conditioning, and inference — are fundamental (automated) features of a PPL. In other words: the edit distance between writing down the (mathematical) model and coding up the (executable) model is amazingly low.

PPLs make Bayesian methods accessible to non-experts. Better yet, PPLs do for creativity in generative modeling what differentiable languages like TensorFlow and PyTorch have done for neural networks: promote a flowering of experimentation through assembly of complex architectures out of legolike, high-level abstractions. In short: tweak the model but not the algorithm. Inference is generally harder than backpropagation, so efficiency may suffer compared to model-specific algorithms. However, wall time is distinct from user time — and the involved process of deriving and implementing custom inference can follow the valuable experimentation phase.

A recent example: dropout (observed abundance of zeros in single-cell RNA sequencing) has been explained as zero inflation since scRNA-seq's inception. Several papers this year[1,2,3] independently contradict this long-held assumption, showing that zero counts in droplet single-cell data closely mirror the expected pattern from count models alone. In a probabilistic program, the model comparison to draw this conclusion is as simple as changing a few words or lines of code.

PPLs are useful for building and *extending* current workhorses in computational biology, like latent factor models and variational autoencoders. They also enable straightforward implementation of hierarchical models, reaping inferential power (and interpretability) by sharing parameters among genes in a common pathway or single cells from a common individual.

---

[1] Townes FW, Hicks SC, Aryee MJ, Irizarry RA (2019) Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. *bioRxiv*: 574574

[2] Svensson V (2019) Droplet scRNA-seq is not zero-inflated. *bioRxiv*: 582064

[3] Silverman JD, Roche K, Mukherjee S, David LA (2018) Naught all zeros in sequence count data are the same. *bioRxiv*: 477794

Could PPLs be more intimately integrated into the wet lab process, like optimizing experimental protocols? Could probabilistic programs of biological processes be synthesized automatically from experimental data? Could useful structures like Gene Ontology and KEGG pathways be encoded as PPL primitives? Could uncertainty quantification inform the next gene to perturb or tissue to sequence?

A call to action for scientists at the intersection of machine learning, language design, and biology: we need better support for discrete structures like trees, a common regime in biology. This is a hard problem since existing black-box methods like variational inference and Hamiltonian Monte Carlo require differentiability of the posterior with respect to its parameters (and so exclude uncollapsed discrete variables).

In tandem, to interpret the dense information contained in high-dimensional, multimodal posteriors, we need new methods for intuitive visualization of uncertainty. And until journals accept graphics with interactivity and animation, we would benefit from new publishing venues in the spirit of machine learning's distill.pub, where in-depth, manipulable graphics (often with inventive interfaces) are the centerpiece and conduit for insight.

We should be exploring how experimental biology can be restructured around probabilistic modeling — as an ongoing part of data collection and experimental design, beyond a post hoc analysis — and how PPLs can be extended to meet the particular challenges of biology and promote model-tinkering in new and creative ways. Bring generative models out of the silo of the lengthy appendix!