Tractable and expressive generative models of genetic variation data

Meihua Dang[1], Anji Liu[1], Xinzhu Wei[1], Sriram Sankararaman*[1,2,3], and Guy Van den Broeck*[1]

[1] Department of Computer Science, UCLA, Los Angeles, USA
{mhdang,liuanji,aprilwei,sriram,guyvdb}@cs.ucla.edu
[2] Department of Human Genetics, School of Medicine, UCLA, Los Angeles, USA
[3] Department of Computational Medicine, School of Medicine, UCLA, Los Angeles, USA

Generative models of genetic sequence data play a central role in population genomics. By modeling dependencies across individuals and sites, these models have empowered genomic analyses such as genotype imputation, haplotype phasing, and ancestry inference. Such models also form the basis for programs that simulate artificial genomes (AGs) that, in turn, have played a critical role in testing evolutionary hypothesis, inferring population genetic models, validating empirical results, and benchmarking methods. The ability to accurately and efficiently simulate AGs has been important. Classical probabilistic models based on hidden Markov model (HMM) are tractable in computing likelihoods and thus widely applied in genotype imputation, but they are not accurate enough in modeling dependencies. While recently popularized deep learning methods such as deep generative adversarial networks (GANs), variational autoencoders (VAEs), and restricted Boltzmann machines (RBMs) are more expressive than HMMs, they are limited in computing exact probabilistic likelihoods and challenging to train.

We propose a class of probabilistic models that can both give us those tractability advantages and keep high expressiveness. To model the distribution over a sequence of variants, we propose a class of latent variable models where each hidden variable is associated with a SNP and the hidden variables are connected via a tree-structured graphical model. This model, termed the *hidden Chow-Liu tree* (HCLT), generalizes previously proposed HMMs. Although HMMs also associate each hidden random variable with a SNP, the hidden variables are related by a chain (a special type of tree) with the restriction that the only edges are present between consecutive SNPs along the genome. By allowing for more general tree structures, the HCLT model can potentially capture long-range correlations or linkage-disequilibrium (LD) among SNPs. While the HCLT model is more expressive than HMMs, it is unclear if such a model can be efficiently learned from data. A second contribution of our work is an affirmative answer to this question by representing HCLTs as Probabilistic Circuits (PCs), a large class of probabilistic models encoded using circuit representations. PCs have been shown to permit tractable inference tasks (e.g., marginal likelihood computation) which are beyond the reach of most deep generative models. The

* Equal contribution.

representation of HCLTs as PCs enables us to leverage recent advances in deep learning such as stochastic learning algorithms and the use of GPUs to enable efficient parameter estimation and inference. HCLT can scale to around 10,000 SNPs and 5,000 genotypes and learning converges in less than 2 hours, and single SNP imputation for all sites on such model takes around 5 seconds. We also leverage the framework of PCs to explore more restrictive models including Markov models.

Finally, we perform extensive experiments to show that HCLTs generate more accurate AGs relative to more restrictive models (fully factorized models and Markov models) suggesting that the structure encoded by the HCLT captures dependencies in genetic variation data. More interestingly, we find that HCLTs as well as deep generative models (GANs and RBMs) preserve LD structure among SNPs. When trained on a subset of individuals from the 1000 Genomes Project (1KGP) across 805 SNPs that are distributed across the genome (and chosen to capture global population structure) as well as a second dataset of 10K SNPs from a contiguous region on chromosome 15, HCLTs greatly improved over existing methods. Compared to HMMs, averaged log-likelihoods of HCLTs improved from -438 to -389 on the 805 SNPs setting and from -633 to -357 on 10K SNPs setting, while results are evaluated on a distinct set of individuals not used in training. We also evaluate the AGs generated by different models by comparing the PCA plots, allele frequencies, and linkage disequilibrium (LD) patterns and observe that the AGs generated by the HCLTs are substantially closer to the patterns observed in real data. In comparison to the next-best method, GANs, the Wasserstein 2D distances between the PCA representations of real versus generated individuals are 0.0015 with a 62.5% improvement and 0.0029 with a 55.4% improvement on 805 and 10K data respectively. The R-squared correlations between real and generated LDs are 0.99 and 0.95, while RBMs achieve 0.98 and 0.95 respectively. Our results suggest that the increased expressivity of HCLTs leads to more accurate models of genetic variation. Furthermore, recent advances in learning and inference enabled by PCs allows us to fully exploit this increased capacity.