Incorporating biological structure into machine learning models in biomedicine

This manuscript (permalink) was automatically generated from greenelab/biopriors-review@e7b3047 on August 29, 2019.

Authors

- **Jake Crawford**
  0000-0001-6207-0782 · jjc2718 · jjc2718
  Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA · Funded by Grant XXXXXXXX

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe
  Department of Something, University of Whatever; Department of Whatever, University of Something

Abstract

Introduction

When applying machine learning techniques to biomedical datasets, it can be challenging to distinguish signal from noise, particularly in the presence of limited amounts of data. Biological knowledge can take many forms, including genomic sequences, pathway databases, gene interaction networks, and knowledge hierarchies such as the Gene Ontology [1]. Incorporating these resources in machine learning models can be helpful in identifying patterns in noisy data [2] and in interpreting model predictions [3]. However, there is often no canonical way to encode these structures as real-valued predictors. This means modelers must be creative when deciding how to encode biological knowledge that they expect will be relevant to the task in models.

Biomedical datasets often contain more input predictors than data samples [4,5]. For example, a genetic study may genotype millions of single nucleotide polymorphisms (SNPs) in hundreds of patients, or a gene expression study may profile the expression of thousands of genes in only a handful of samples. Thus, it can be useful to include prior information describing the relationships between the predictors to inform the representation learned by the model. This stands in contrast to non-biological applications of machine learning, where one might fit a model on millions of images [6] or tens of thousands of documents [7], making inclusion of prior information unnecessary.

In this review, we survey approaches to learning models from biomedical data that incorporate external information about the structure of desirable solutions. One class of commonly used approaches involves using raw sequence data to learn a representation that considers the context of each base pair. For models that operate on gene input, such as gene expression data or genetic variants, it can be useful to incorporate networks or pathways describing relationships between genes. We also consider other examples in this review, such as neural network architectures that are constrained based on biological knowledge.

Sequence-Constrained Models

Applications in regulatory biology

1. Early examples
   - 10.1126/science.1254806 (can cite as an early example of using hand-engineered features as input to a 2-layer MLP, preceded most CNN examples on sequence data -> splicing exon inclusion prediction)
   - 10.1038/nbt.3300 (CNN for prediction of DNA/RNA binding protein sequence specificity)
   - 10.1038/nmeth.3547 (CNN for multi-task variant effect prediction from noncoding genome, one of the earlier examples of CNN on genomic sequence data)
   - 10.1101/gr.200535.11 (CNN for prediction of chromatin accessibility from DNA sequence, evaluated to predict effects of SNPs)
2. More recent examples
   - 10.1101/737981 (CNN for prediction of TF binding probability and binding strength from sequence, used to identify and interpret cis-regulatory code i.e. order and spacing of TF binding motifs, no need for strong modeling assumptions + allow lots of context ~1000 base windows)
   - 10.1016/j.cell.2019.04.046 (prediction of alternative polyadenylation from sequence, used to predict probability of poly-A cleavage and engineer new PASs)
   - DNA methylation models typically operate on microarray beta values (proportion of DNA methylated at each CpG locus), so these wouldn't be considered sequence-constrained. An exception: DeepSignal (10.1093/bioinformatics/btz276) - CNN that works on electrical signals around CpG sites from Nanopore sequencing reads, to predict 5mC or 6mA methylation status

    - 10.1101/329334 and 10.1093/bioinformatics/bty612: prediction of gene expression (or differential gene expression) from histone modification signals. Signals are discretized using bins (no raw sequence input) but bidirectional LSTM considers context in both directions and allows for learning complex histone mark dependencies.

Applications in variant calling and mutation detection

- 10.1038/nbt.4235 (actually works on images of read pileups, but uses sequence context)
- 10.1038/s41467-019-09025-z (single molecule sequencing: PacBio/ONT, works on featurization of sequence defined in paper, outperforms ^ on these)
- 10.1101/097469 (similar to DeepVariant, but works on raw sequences rather than images, + less preprocessing/claims to learn more info directly from sequences)
- 10.1101/601450 (calling of short indels using sequence information)
- 10.1038/s41467-019-09027-x (mutation detection in tumors)

Applications in CRISPR guide selection

- 10.1371/journal.pcbi.1005807 (shows that a wide variety of sequence attributes are useful for predicting cleavage efficiency, didn't use contextual info?)
- 10.1038/nbt.3437, 10.1038/s41551-017-0178-6 (models for predicting on-target efficacy and off-target effects, uses contextual info about sequence i.e. order 2+ features)
- 10.1038/nbt.4061 (CNN-based model for Cpf1 activity prediction, convolutions of length 5 on one-hot encoded sequence)
- 10.1186/s13059-018-1459-4 (similar to ^ but for Cas9, also uses ChIP-seq data)
- 10.1101/636472 (similar to ^)
- 10.1101/505602 (uses STRING -> gene network-based features, in addition to sequence features, for on-target efficacy prediction)

Network- and Pathway-Constrained Models

Applications in bulk transcriptomics

- 10.1016/j.cels.2019.04.003 (pathways -> transfer learning for rare diseases)
- 10.1038/s41540-019-0086-3 (PPI + pathways -> phenotype prediction/cancer subtyping)
- https://arxiv.org/pdf/1906.10670 (HumanBase -> AML drug response prediction)
- 10.1093/bioinformatics/bty429 (HINT PPI)
- 10.1038/s41598-017-03141-w (PPI -> mutated cancer gene detection)
- 10.1371/journal.pcbi.1006657 (PPI + network optimization -> cancer outcome prediction)
- 10.1186/s12859-018-2500-z (pathways -> NN structure, for predicting patient outcomes in GBM)
- 10.1186/s12859-017-1984-2 (gene regulatory network -> NN structure, for patient outcomes in kidney transplantation and ulcerative colitis)
- 10.1038/s41598-018-19635-0 (gene regulatory network -> group lasso, same datasets as ^)
- 10.1093/bioinformatics/bty945 (gene regulatory network -> NN structure, for predicting gene expression given gene knockouts)

Applications in single cell transcriptomics

- 10.1038/s41592-019-0456-1 (pathways -> cell type deconvolution/pathway-level eQTLs)
- 10.1093/nar/gkx681 (PPI; protein-DNA interactions)
- 10.1101/544346 (PPI)

Applications in genetics

1. Genotype-phenotype associations (GWAS/eQTL)
    - 10.1093/bioinformatics/btu293

    - 10.1093/bioinformatics/btx677 (multivariate version of ^)
    - 10.1111/biom.13072 (phenotype prediction using variant and GE data)
2. Using pathways as a prior to study SNP-SNP or gene-gene interactions (not sure if this is truly a constraint on the model, more just a strategy to reduce the number of hypothesis tests)
    - 10.1101/182741, 10.1371/journal.pgen.1006973 (pathways + GWAS data -> GIs)
    - 10.1371/journal.pgen.1006516 (pathways + GWAS data -> GIs)
3. Using SNP-SNP or gene-gene networks to reduce hypothesis testing burden for detecting GIs
    - can find some examples if we decide this is a direction worth going

Other constrained models

Ontology-constrained models

1. Phenotype prediction
    - 10.1038/nmeth.4627 (gene ontology -> NN structure, to predict effects of mutations on growth in yeast)
2. Function prediction
    - 10.1093/bioinformatics/btx624 (GO hierarchy relationships -> NN structure, for function prediction)
    - 10.1093/bioinformatics/btx252 (PPI network + tissue ontology -> function prediction)

Otherwise constrained models

1. Cell cycle information
    - 10.1038/nbt.3102
    - 10.1101/526848 (in principle the denoising method could be generalized to other gene sets, but here they used cell cycle-relevant gene sets and emphasized the utility of this)
2. Circadian rhythms: 10.1073/pnas.1619320114 (circular node autoencoder for modeling periodic gene expression)
3. TAD/3D chromatin structure information? Can probably find some examples of this

Conclusions

1. What is outside of the scope of this review? (this can also go in introduction?)
    a. Biological “constraints” vs. feature selection or feature extraction from heterogeneous biological data (e.g. network embedding approaches)
        - Example: one could use a network-based feature extraction method (e.g. Node2Vec) to convert each gene in a PPI network into a set of real-valued features, then use those + gene expression as input to a model
        - For purposes of keeping this review short enough, I’m trying to stay away from papers like ^, but still unclear to me where exactly the line should be drawn. Almost any ML model that operates on sequence data can be viewed as having a feature extraction component, for example.

    b. Could kind-of consider many single-cell dimension reduction methods biologically constrained (e.g. dropout/zero inflation modeling approaches, etc), but this is way too broad for this review - maybe refer the reader to other recent reviews of these methods.

    c. Could also consider omics integration methods (combining, for example, gene expression and epigenetic data) to be biologically constrained, but we refer the reader to 10.1016/j.inffus.2018.09.012 for further detail on these methods.

References

1. **The Gene Ontology Resource: 20 years and still GOing strong***Nucleic Acids Research* (2018-11-05) https://doi.org/gf63mb
DOI: 10.1093/nar/gky1055 · PMID: 30395331 · PMCID: PMC6323945

2. **Network propagation: a universal amplifier of genetic associations**
Lenore Cowen, Trey Ideker, Benjamin J. Raphael, Roded Sharan
Nature Reviews Genetics (2017-06-12) https://doi.org/gbhkwn
DOI: 10.1038/nrg.2017.38 · PMID: 28607512

3. **Visible Machine Learning for Biomedicine**
Michael K. Yu, Jianzhu Ma, Jasmin Fisher, Jason F. Kreisberg, Benjamin J. Raphael, Trey Ideker
Cell (2018-06) https://doi.org/gdqcd8
DOI: 10.1016/j.cell.2018.05.056 · PMID: 29906441 · PMCID: PMC6483071

4. **Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets**
Michael K. K. Leung, Andrew Delong, Babak Alipanahi, Brendan J. Frey
Proceedings of the IEEE (2016-01) https://doi.org/f75grb
DOI: 10.1109/jproc.2015.2494198

5. **Diet Networks: Thin Parameters for Fat Genomics**
Adriana Romero, Pierre Luc Carrier, Akram Erraqabi, Tristan Sylvain, Alex Auvolat, Etienne Dejoie, Marc-André Legault, Marie-Pierre Dubé, Julie G. Hussin, Yoshua Bengio
arXiv (2016-11-28) https://arxiv.org/abs/1611.09340v3

6. **ImageNet: A large-scale hierarchical image database**
Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei
2009 IEEE Conference on Computer Vision and Pattern Recognition (2009-06) https://doi.org/cvc7xp
DOI: 10.1109/cvpr.2009.5206848

7. **NewsWeeder: Learning to Filter Netnews**
Ken Lang
Machine Learning Proceedings 1995 (1995) https://doi.org/gf63mh
DOI: 10.1016/b978-1-55860-377-6.50048-7