Incorporating biological structure into machine learning models in biomedicine

This manuscript (permalink) was automatically generated from greenelab/biopriors-review@a5e7ab2 on September 13, 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 models

Early neural network models primarily used hand-engineered sequence features as input to a fully connected neural network [8,9]. As convolutional neural network (CNN) approaches matured for image processing and computer vision, researchers were able to use similar ideas to leverage biological sequence proximity in modeling. CNNs are a neural network variant in which input data are grouped by spatial context to extract features for prediction. The definition of “spatial context” is specific to the input. For example, for images pixels that are nearby in 2D space might be grouped, or for genomic sequences base pairs that are nearby in the linear genome might be grouped.

These approaches work by first encoding input into a numeric matrix (for DNA “one-hot” encoding is often used: A=[1,0,0,0], C=[0,1,0,0], G=[0,0,1,0], T=[0,0,0,1]). They then apply spatial filters to the encoded input, which are adjusted based on the data during model training. In this way, CNNs are able to consider context without making strong assumptions about exactly how much context is needed or how it should be encoded; the data informs the encoding. A detailed description of how CNNs are applied to sequence data can be found in [10].

Applications in regulatory biology

Many of the first applications of deep learning to biological sequence data were in regulatory biology. Example early applications of CNNs on sequence data include prediction of binding protein sequence

specificity from DNA or RNA sequence [11], prediction of variant effects from noncoding DNA sequence [12], and prediction of chromatin accessibility from DNA sequence [13].

Recent sequence models take advantage of hardware advances and methodological innovation to incorporate more sequence context and rely on fewer modeling assumptions. BPNet, a CNN used to predict transcription factor binding profiles from raw DNA sequences, was able to accurately map known locations of binding motifs in mouse embryonic stem cells [14]. The BPNet model considers 1000 base pairs of context around each position when predicting binding probabilities, using a technique called dilated convolutions [15] to make a large input field feasible. This context is particularly important because motif spacing and periodicity can influence protein function. cDeepbind [16] combines RNA sequences with information about secondary structure to predict RNA binding protein affinities. Its convolutional model acts on a feature vector combining sequence and structural information, simultaneously using context for both to inform predictions. APARENT [17] is a CNN used to predict alternative polyadenylation (APA) from a training set of over 3 million synthetic APA reporter sequences. These diverse applications underscore the power of modern deep learning models to synthesize large sequence datasets.

Models that consider sequence context have also been applied to impute and make predictions from epigenetic data. DeepSignal [18] is a CNN that uses contextual electrical signals from Oxford Nanopore single-molecule sequencing data to predict 5mC or 6mA DNA methylation status. MRCNN [19] uses sequences of length 400, centered at CpG sites, to predict 5mC methylation status. Deep learning models have also been used to predict gene expression from histone modifications [20,21]. Here, a neural network model consisting of long short-term memory (LSTM) units was used to encode the long-distance interactions of histone marks in both the 3′ and 5′ genomic directions. In each of these cases, proximity in the linear genome proved useful for modeling the complex interactions between DNA sequence and epigenome.

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 ^)
- 10.1093/bioinformatics/bty554 (similar to above algorithms but for off-target predictions, need to read in more detail)
- 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

8. **DEEP: a general computational framework for predicting enhancers**
Dimitrios Kleftogiannis, Panos Kalnis, Vladimir B. Bajic
Nucleic Acids Research (2014-11-05) https://doi.org/gcgk83
DOI: 10.1093/nar/gku1058 · PMID: 25378307 · PMCID: PMC4288148

9. **The human splicing code reveals new insights into the genetic determinants of disease**
H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, … B. J. Frey
Science (2014-12-18) https://doi.org/f6wzj2
DOI: 10.1126/science.1254806 · PMID: 25525159 · PMCID: PMC4362528

10. **Deep learning for computational biology**
Christof Angermueller, Tanel Pärnamaa, Leopold Parts, Oliver Stegle
Molecular Systems Biology (2016-07) https://doi.org/f8xtvh
DOI: 10.15252/msb.20156651 · PMID: 27474269 · PMCID: PMC4965871

11. **Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning**
Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey
Nature Biotechnology (2015-07-27) https://doi.org/f7mkrd
DOI: 10.1038/nbt.3300 · PMID: 26213851

12. **Predicting effects of noncoding variants with deep learning–based sequence model**
Jian Zhou, Olga G Troyanskaya
Nature Methods (2015-08-24) https://doi.org/gcgk8g
DOI: 10.1038/nmeth.3547 · PMID: 26301843 · PMCID: PMC4768299

13. **Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks**
David R. Kelley, Jasper Snoek, John L. Rinn
Genome Research (2016-05-03) https://doi.org/f8sw35
DOI: 10.1101/gr.200535.115 · PMID: 27197224 · PMCID: PMC4937568

14. **Deep learning at base-resolution reveals motif syntax of the cis-regulatory code**
Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Amr Alexandari, Sabrina Krueger, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, Julia Zeitlinger
Cold Spring Harbor Laboratory (2019-08-21) https://doi.org/gf64fc
DOI: 10.1101/737981

15. **Multi-Scale Context Aggregation by Dilated Convolutions**
Fisher Yu, Vladlen Koltun
arXiv (2015-11-23) https://arxiv.org/abs/1511.07122v3

16. **cDeepbind: A context sensitive deep learning model of RNA-protein binding**
Shreshth Gandhi, Leo J. Lee, Andrew Delong, David Duvenaud, Brendan J. Frey
Cold Spring Harbor Laboratory (2018-06-12) https://doi.org/gf68nk
DOI: 10.1101/345140

17. **A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation**
Nicholas Bogard, Johannes Linder, Alexander B. Rosenberg, Georg Seelig
Cell (2019-06) https://doi.org/gf66nc
DOI: 10.1016/j.cell.2019.04.046 · PMID: 31178116 · PMCID: PMC6599575

18. **DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning**
Peng Ni, Neng Huang, Zhi Zhang, De-Peng Wang, Fan Liang, Yu Miao, Chuan-Le Xiao, Feng Luo, Jianxin Wang
Bioinformatics (2019-04-17) https://doi.org/gf66qw
DOI: 10.1093/bioinformatics/btz276 · PMID: 30994904

19. **MRCNN: a deep learning model for regression of genome-wide DNA methylation**
Qi Tian, Jianxiao Zou, Jianxiong Tang, Yuan Fang, Zhongli Yu, Shicai Fan
BMC Genomics (2019-04) https://doi.org/gf48g6
DOI: 10.1186/s12864-019-5488-5 · PMID: 30967120 · PMCID: PMC6457069

20. **Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin**
Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi
Cold Spring Harbor Laboratory (2018-05-25) https://doi.org/gf66qz
DOI: 10.1101/329334

21. **DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications**
Arshdeep Sekhon, Ritambhara Singh, Yanjun Qi
Bioinformatics (2018-09-01) https://doi.org/gd9mk4
DOI: 10.1093/bioinformatics/bty612 · PMID: 30423076