Deep Learning for Computational Biology

CS260

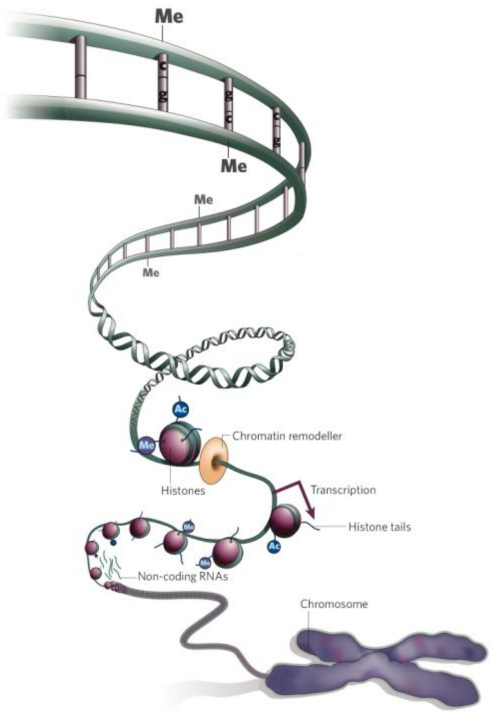
UC RIVERSIDE
UNIVERSITY OF CALIFORNIA

January 15, 2018

Epigenetics

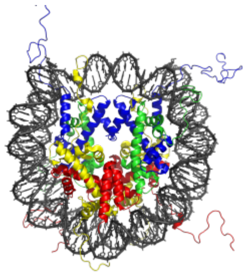
DNA methylation
Nucleosome positioning
Histones tail modifications
Chromatin

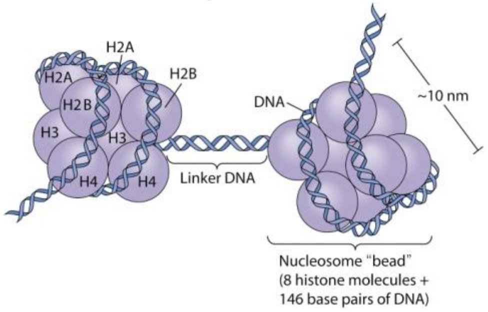
Epigenome



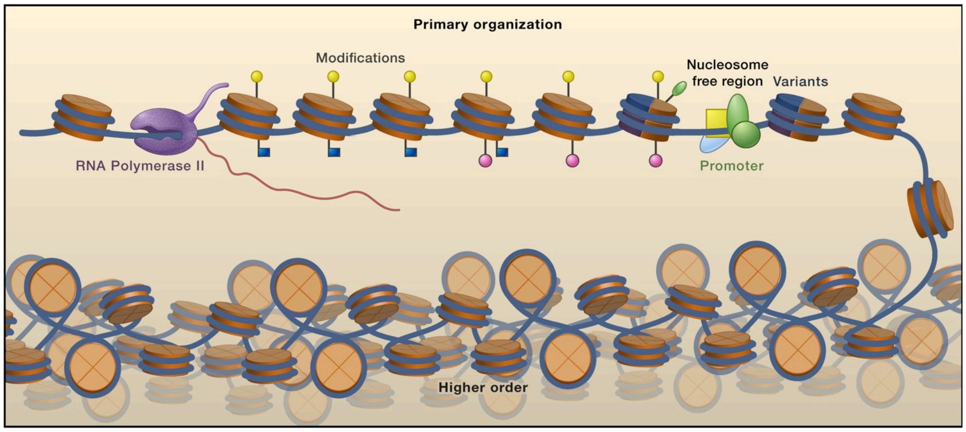
Epigenetics: Nucleosomes/Histones

- Eight core histones (2 x H2A, 2 x H2B, 2 x H3, and 2 x H4) and ~147bps of DNA wrapped around form the ‘nucleosome’
- Nucleosomes are thought to carry epigenetically inherited information in the form of covalent modifications of their core histones (*histone code)*
- Nucleosome “slide” (?) and their position regulate gene expression





Nucleosomes organization



DNA methylation

- Cytosines (5mC) can be methylated, which “turns” them in a “fifth” type of nucleotide
- In mammals 60%-90% of all CpGs are 5mC
- The rate of cytosine DNA methylation differs strongly between species, e.g. 14% of cytosines are methylated in *Arabidopsis thaliana*, 8% in *Mus musculus*, 2.3% in *Escherichia coli*, 0.03% in *Drosophila*, and virtually none (< 0.0002%) in yeast
- In plants and stem cells, the cytosine can be methylated also CHG, and CHH sites, where H={A,C,T} (*asymmetric methylation*) in addition to CG (*symmetric methylation*)

DNA methylation

- 5mC is associated with gene *silencing* (decreased gene expression)
- 5mC also plays a crucial role in the development of nearly all types of cancer and *imprinting*
- 5mC can be inherited through cell division (in somatic cells, patterns of DNA methylation are transmitted to daughter cells with a high fidelity)
- 5mC is dynamic, but DNA demethylation is poorly understood
- New types of methylation in mammalian genomes: hydroxymethylation (5hmC), 5-formylcytosine (5fC), 5-carboxycytosine (5caC)

RNA methylation

- Adenosine (6mA) is the most prevalent mammalian mRNA post-transcriptional modification
- It is also found in tRNA, rRNA, and small nuclear RNA (snRNA) as well as several long non-coding RNA
- Recent studies have discovered protein ‘writers’, ‘erasers’ and ‘readers’ of this RNA chemical mark
- It has been show to affect gene expression, but its function it is poorly understood

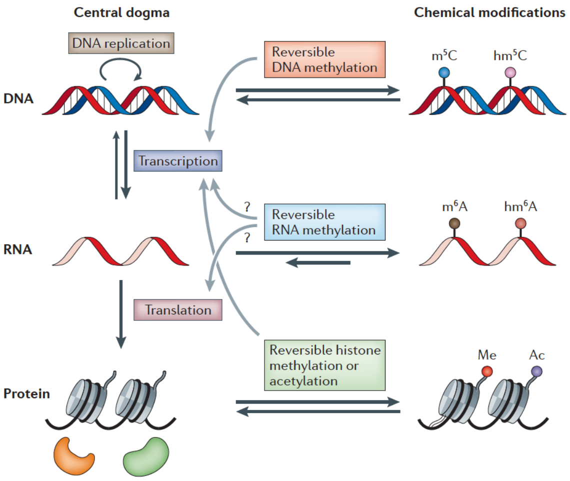
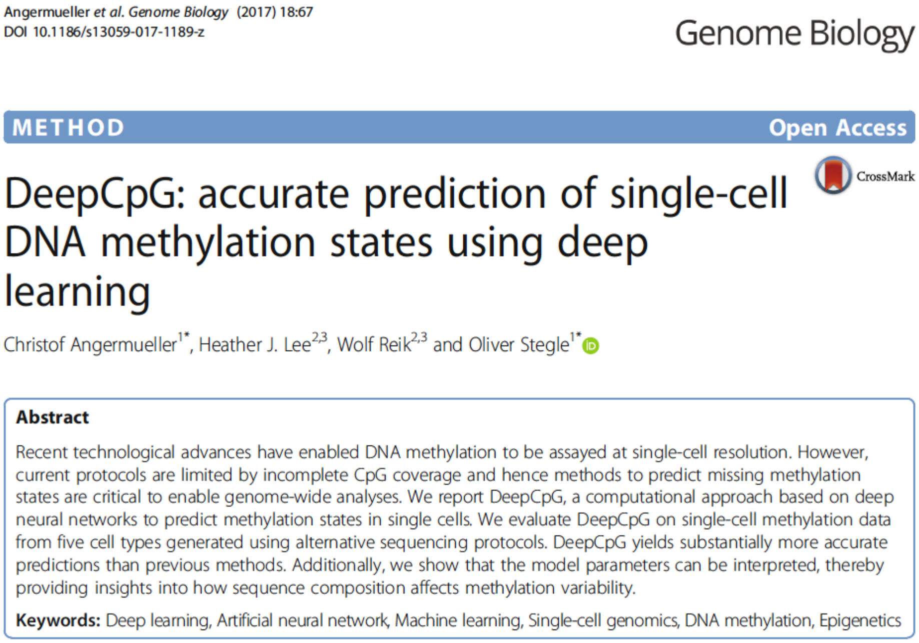


Figure 1 | **Reversible chemical modifications that regulate the flow of genetic information.** In the central dogma, genetic information is passed from DNA to RNA and then to protein. Epigenetic DNA modifications (for example, the formation of 5-methylcytosine (m^{5}C; also known as 5mC) and 5-hydroxymethylcytosine (hm^{5}C; also known as 5hmC)) and histone modifications (for example, methylation (me) and acetylation (ac)) are known to have important roles in regulating cell differentiation and development. Reversible RNA modifications (for example, the formation of *N*^{6}-methyladenosine (m^{6}A) and *N*^{6}-hydroxymethyladenosine (hm^{6}A)) add an additional layer of dynamic regulation of biological processes.

Fu *et al.*, *Nat Rev Genetics*, 2014

Prediction problems

- DNA methylation state
- Impact of sequence variation in methylation
- TF motifs from ChIP-seq data
- Non-coding function prediction from sequence
- Gene expression from histone tail modifications
- Regulatory code of the accessible genome



Angermueller *et al. Genome Biology* (2017) 18:67
DOI 10.1186/s13059-017-1189-z

Genome Biology

METHOD Open Access

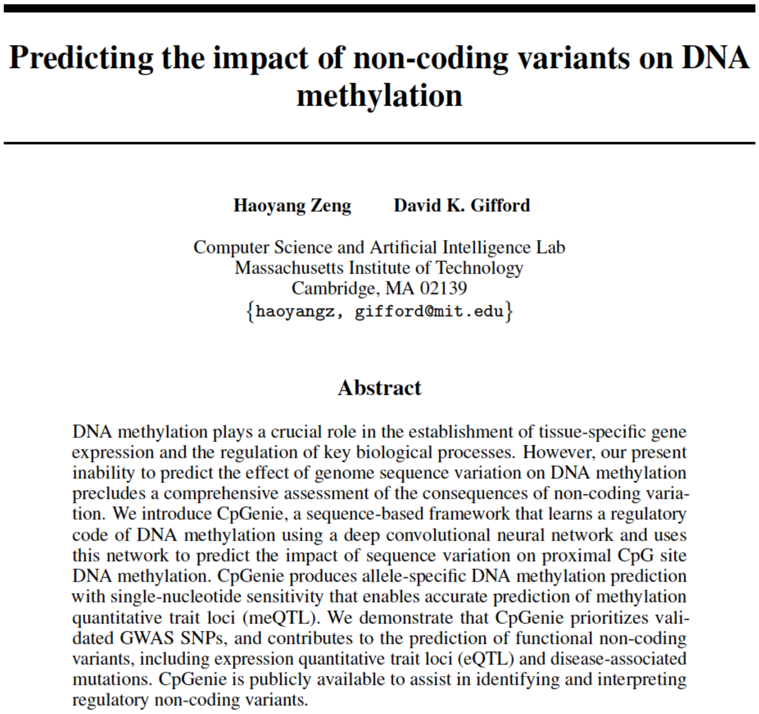
DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning

Christof Angermueller[1*], Heather J. Lee[2,3], Wolf Reik[2,3] and Oliver Stegle[1*]

Abstract

Recent technological advances have enabled DNA methylation to be assayed at single-cell resolution. However, current protocols are limited by incomplete CpG coverage and hence methods to predict missing methylation states are critical to enable genome-wide analyses. We report DeepCpG, a computational approach based on deep neural networks to predict methylation states in single cells. We evaluate DeepCpG on single-cell methylation data from five cell types generated using alternative sequencing protocols. DeepCpG yields substantially more accurate predictions than previous methods. Additionally, we show that the model parameters can be interpreted, thereby providing insights into how sequence composition affects methylation variability.

Keywords: Deep learning, Artificial neural network, Machine learning, Single-cell genomics, DNA methylation, Epigenetics



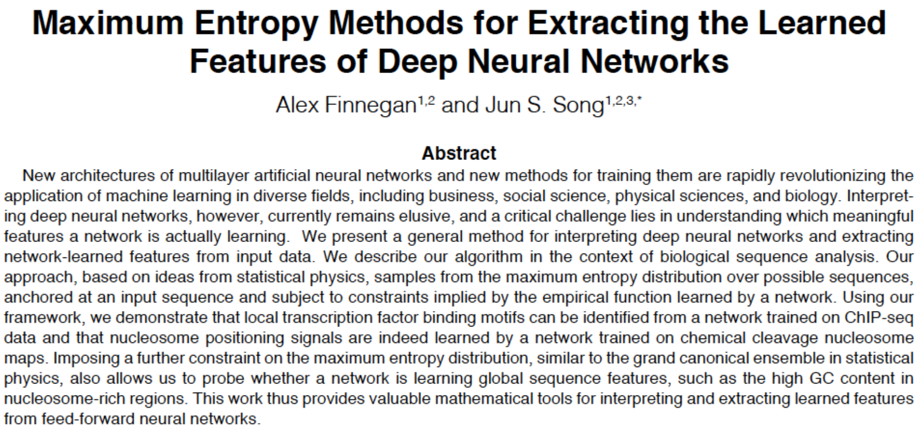
Predicting the impact of non-coding variants on DNA methylation

Haoyang Zeng **David K. Gifford**

Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA 02139
{haoyangz, gifford@mit.edu}

Abstract

DNA methylation plays a crucial role in the establishment of tissue-specific gene expression and the regulation of key biological processes. However, our present inability to predict the effect of genome sequence variation on DNA methylation precludes a comprehensive assessment of the consequences of non-coding variation. We introduce CpGenie, a sequence-based framework that learns a regulatory code of DNA methylation using a deep convolutional neural network and uses this network to predict the impact of sequence variation on proximal CpG site DNA methylation. CpGenie produces allele-specific DNA methylation prediction with single-nucleotide sensitivity that enables accurate prediction of methylation quantitative trait loci (meQTL). We demonstrate that CpGenie prioritizes validated GWAS SNPs, and contributes to the prediction of functional non-coding variants, including expression quantitative trait loci (eQTL) and disease-associated mutations. CpGenie is publicly available to assist in identifying and interpreting regulatory non-coding variants.

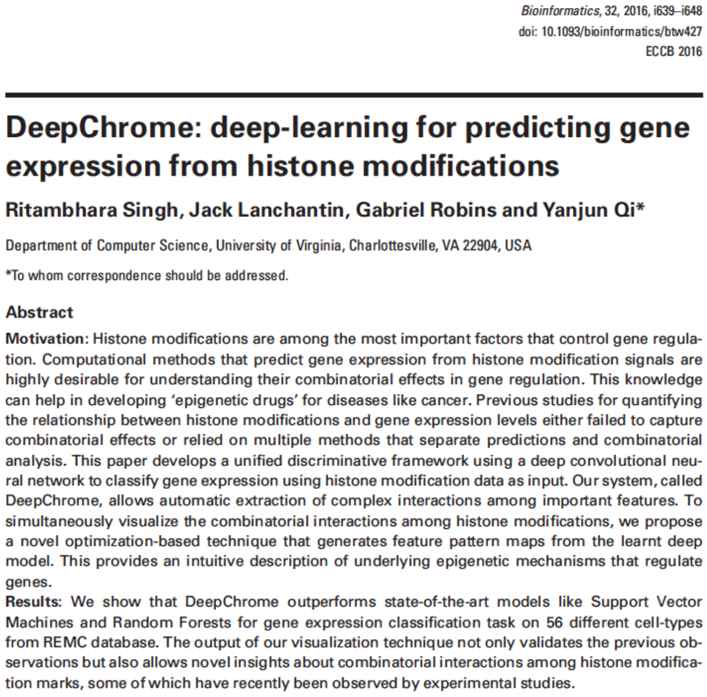


Maximum Entropy Methods for Extracting the Learned Features of Deep Neural Networks

Alex Finnegan[1,2] and Jun S. Song[1,2,3,*]

Abstract

New architectures of multilayer artificial neural networks and new methods for training them are rapidly revolutionizing the application of machine learning in diverse fields, including business, social science, physical sciences, and biology. Interpreting deep neural networks, however, currently remains elusive, and a critical challenge lies in understanding which meaningful features a network is actually learning. We present a general method for interpreting deep neural networks and extracting network-learned features from input data. We describe our algorithm in the context of biological sequence analysis. Our approach, based on ideas from statistical physics, samples from the maximum entropy distribution over possible sequences, anchored at an input sequence and subject to constraints implied by the empirical function learned by a network. Using our framework, we demonstrate that local transcription factor binding motifs can be identified from a network trained on ChIP-seq data and that nucleosome positioning signals are indeed learned by a network trained on chemical cleavage nucleosome maps. Imposing a further constraint on the maximum entropy distribution, similar to the grand canonical ensemble in statistical physics, also allows us to probe whether a network is learning global sequence features, such as the high GC content in nucleosome-rich regions. This work thus provides valuable mathematical tools for interpreting and extracting learned features from feed-forward neural networks.



Bioinformatics, 32, 2016, i639–i648
doi: 10.1093/bioinformatics/btw427
ECCB 2016

DeepChrome: deep-learning for predicting gene expression from histone modifications

Ritambhara Singh, Jack Lanchantin, Gabriel Robins and Yanjun Qi*

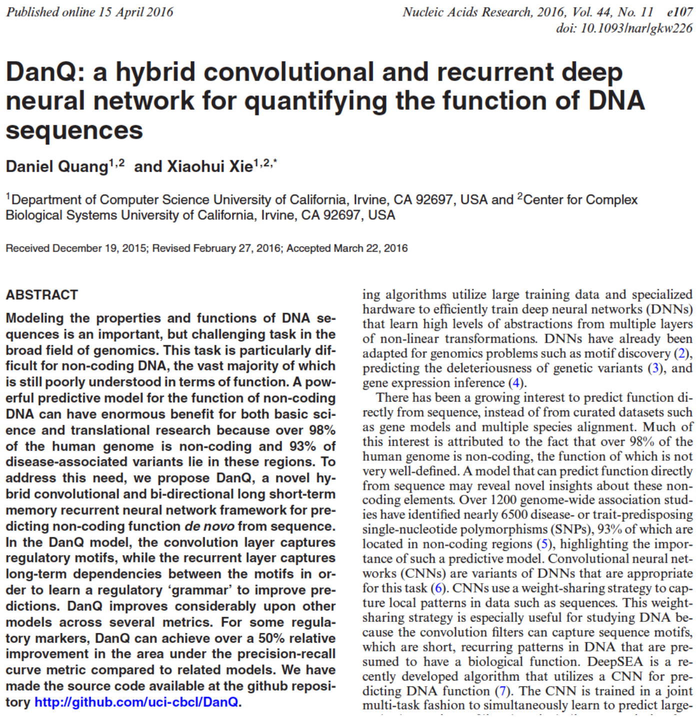
Department of Computer Science, University of Virginia, Charlottesville, VA 22904, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Histone modifications are among the most important factors that control gene regulation. Computational methods that predict gene expression from histone modification signals are highly desirable for understanding their combinatorial effects in gene regulation. This knowledge can help in developing 'epigenetic drugs' for diseases like cancer. Previous studies for quantifying the relationship between histone modifications and gene expression levels either failed to capture combinatorial effects or relied on multiple methods that separate predictions and combinatorial analysis. This paper develops a unified discriminative framework using a deep convolutional neural network to classify gene expression using histone modification data as input. Our system, called DeepChrome, allows automatic extraction of complex interactions among important features. To simultaneously visualize the combinatorial interactions among histone modifications, we propose a novel optimization-based technique that generates feature pattern maps from the learnt deep model. This provides an intuitive description of underlying epigenetic mechanisms that regulate genes.

Results: We show that DeepChrome outperforms state-of-the-art models like Support Vector Machines and Random Forests for gene expression classification task on 56 different cell-types from REMC database. The output of our visualization technique not only validates the previous observations but also allows novel insights about combinatorial interactions among histone modification marks, some of which have recently been observed by experimental studies.



Published online 15 April 2016

Nucleic Acids Research, 2016, Vol. 44, No. 11 e107
doi: 10.1093/nar/gkw226

DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences

Daniel Quang[1,2] and Xiaohui Xie[1,2,*]

[1]Department of Computer Science University of California, Irvine, CA 92697, USA and [2]Center for Complex Biological Systems University of California, Irvine, CA 92697, USA

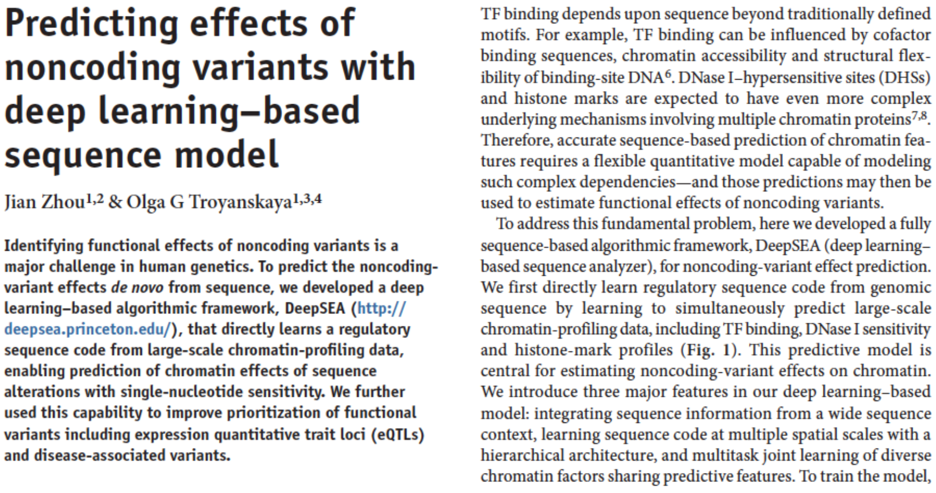
Received December 19, 2015; Revised February 27, 2016; Accepted March 22, 2016

ABSTRACT

Modeling the properties and functions of DNA sequences is an important, but challenging task in the broad field of genomics. This task is particularly difficult for non-coding DNA, the vast majority of which is still poorly understood in terms of function. A powerful predictive model for the function of non-coding DNA can have enormous benefit for both basic science and translational research because over 98% of the human genome is non-coding and 93% of disease-associated variants lie in these regions. To address this need, we propose DanQ, a novel hybrid convolutional and bi-directional long short-term memory recurrent neural network framework for predicting non-coding function *de novo* from sequence. In the DanQ model, the convolution layer captures regulatory motifs, while the recurrent layer captures long-term dependencies between the motifs in order to learn a regulatory 'grammar' to improve predictions. DanQ improves considerably upon other models across several metrics. For some regulatory markers, DanQ can achieve over a 50% relative improvement in the area under the precision-recall curve metric compared to related models. We have made the source code available at the github repository http://github.com/uci-cbcl/DanQ.

ing algorithms utilize large training data and specialized hardware to efficiently train deep neural networks (DNNs) that learn high levels of abstractions from multiple layers of non-linear transformations. DNNs have already been adapted for genomics problems such as motif discovery (2), predicting the deleteriousness of genetic variants (3), and gene expression inference (4).

There has been a growing interest to predict function directly from sequence, instead of from curated datasets such as gene models and multiple species alignment. Much of this interest is attributed to the fact that over 98% of the human genome is non-coding, the function of which is not very well-defined. A model that can predict function directly from sequence may reveal novel insights about these non-coding elements. Over 1200 genome-wide association studies have identified nearly 6500 disease- or trait-predisposing single-nucleotide polymorphisms (SNPs), 93% of which are located in non-coding regions (5), highlighting the importance of such a predictive model. Convolutional neural networks (CNNs) are variants of DNNs that are appropriate for this task (6). CNNs use a weight-sharing strategy to capture local patterns in data such as sequences. This weight-sharing strategy is especially useful for studying DNA because the convolution filters can capture sequence motifs, which are short, recurring patterns in DNA that are presumed to have a biological function. DeepSEA is a recently developed algorithm that utilizes a CNN for predicting DNA function (7). The CNN is trained in a joint multi-task fashion to simultaneously learn to predict large-



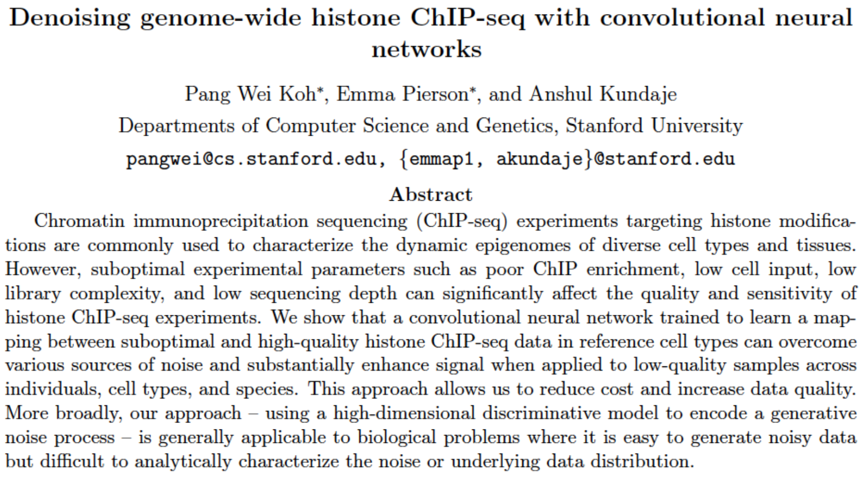
Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou[1,2] & Olga G Troyanskaya[1,3,4]

Identifying functional effects of noncoding variants is a major challenge in human genetics. To predict the noncoding-variant effects *de novo* from sequence, we developed a deep learning–based algorithmic framework, DeepSEA (http://deepsea.princeton.edu/), that directly learns a regulatory sequence code from large-scale chromatin-profiling data, enabling prediction of chromatin effects of sequence alterations with single-nucleotide sensitivity. We further used this capability to improve prioritization of functional variants including expression quantitative trait loci (eQTLs) and disease-associated variants.

TF binding depends upon sequence beyond traditionally defined motifs. For example, TF binding can be influenced by cofactor binding sequences, chromatin accessibility and structural flexibility of binding-site DNA[6]. DNase I–hypersensitive sites (DHSs) and histone marks are expected to have even more complex underlying mechanisms involving multiple chromatin proteins[7,8]. Therefore, accurate sequence-based prediction of chromatin features requires a flexible quantitative model capable of modeling such complex dependencies—and those predictions may then be used to estimate functional effects of noncoding variants.

To address this fundamental problem, here we developed a fully sequence-based algorithmic framework, DeepSEA (deep learning–based sequence analyzer), for noncoding-variant effect prediction. We first directly learn regulatory sequence code from genomic sequence by learning to simultaneously predict large-scale chromatin-profiling data, including TF binding, DNase I sensitivity and histone-mark profiles (**Fig. 1**). This predictive model is central for estimating noncoding-variant effects on chromatin. We introduce three major features in our deep learning–based model: integrating sequence information from a wide sequence context, learning sequence code at multiple spatial scales with a hierarchical architecture, and multitask joint learning of diverse chromatin factors sharing predictive features. To train the model,

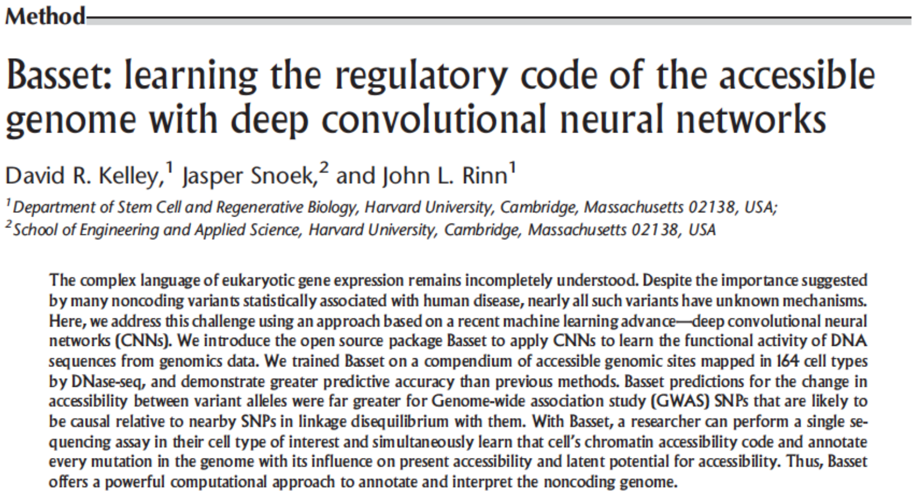


Denoising genome-wide histone ChIP-seq with convolutional neural networks

Pang Wei Koh*, Emma Pierson*, and Anshul Kundaje
Departments of Computer Science and Genetics, Stanford University
pangwei@cs.stanford.edu, {emmap1, akundaje}@stanford.edu

Abstract

Chromatin immunoprecipitation sequencing (ChIP-seq) experiments targeting histone modifications are commonly used to characterize the dynamic epigenomes of diverse cell types and tissues. However, suboptimal experimental parameters such as poor ChIP enrichment, low cell input, low library complexity, and low sequencing depth can significantly affect the quality and sensitivity of histone ChIP-seq experiments. We show that a convolutional neural network trained to learn a mapping between suboptimal and high-quality histone ChIP-seq data in reference cell types can overcome various sources of noise and substantially enhance signal when applied to low-quality samples across individuals, cell types, and species. This approach allows us to reduce cost and increase data quality. More broadly, our approach – using a high-dimensional discriminative model to encode a generative noise process – is generally applicable to biological problems where it is easy to generate noisy data but difficult to analytically characterize the noise or underlying data distribution.



Method

Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks

David R. Kelley,[1] Jasper Snoek,[2] and John L. Rinn[1]

[1] *Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA;*
[2] *School of Engineering and Applied Science, Harvard University, Cambridge, Massachusetts 02138, USA*

The complex language of eukaryotic gene expression remains incompletely understood. Despite the importance suggested by many noncoding variants statistically associated with human disease, nearly all such variants have unknown mechanisms. Here, we address this challenge using an approach based on a recent machine learning advance—deep convolutional neural networks (CNNs). We introduce the open source package Basset to apply CNNs to learn the functional activity of DNA sequences from genomics data. We trained Basset on a compendium of accessible genomic sites mapped in 164 cell types by DNase-seq, and demonstrate greater predictive accuracy than previous methods. Basset predictions for the change in accessibility between variant alleles were far greater for Genome-wide association study (GWAS) SNPs that are likely to be causal relative to nearby SNPs in linkage disequilibrium with them. With Basset, a researcher can perform a single sequencing assay in their cell type of interest and simultaneously learn that cell's chromatin accessibility code and annotate every mutation in the genome with its influence on present accessibility and latent potential for accessibility. Thus, Basset offers a powerful computational approach to annotate and interpret the noncoding genome.

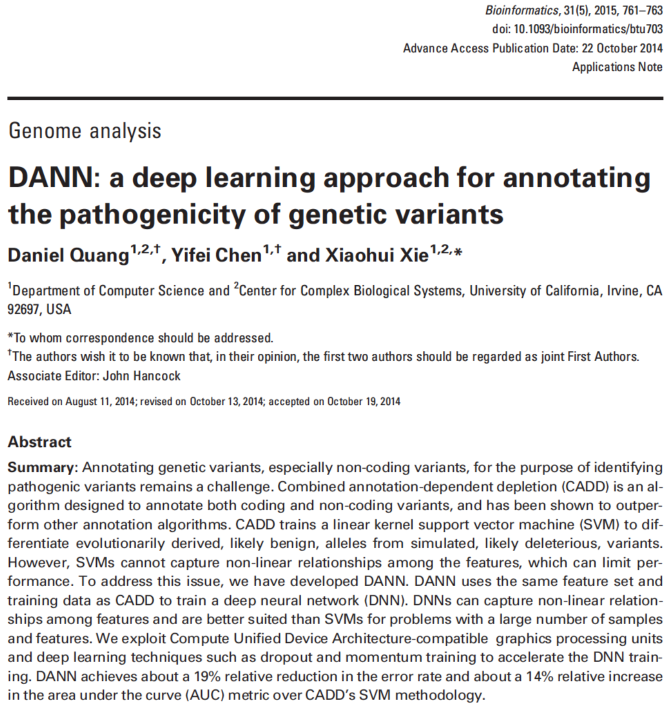
Structural variations
SNPs

Variations in the Human Genome

- *Single nucleotide polymorphisms* (SNP) occur on average between every 1 in 100 and 1 in 300 bases in the human genome
- Large-scale structural variations range from a few thousand to a few million bps: these variations include differences in the number of copies individuals have of a particular gene, deletions, translocations and inversions (*copy number variations* or CNV)
- A high proportion of the genome (currently estimated at up to 12%) is subject to CNV
- SNPs and CNVs may either be inherited or caused by *de novo* mutation
- Many SNPs and CNVs are associated with genetic diseases and cancer

Prediction problems

- Pathogenicity of genetic variants
- Modeling of SNPs using Boltzmann machines
- Calling SNP variants



Bioinformatics, 31(5), 2015, 761–763
doi: 10.1093/bioinformatics/btu703
Advance Access Publication Date: 22 October 2014
Applications Note

Genome analysis

DANN: a deep learning approach for annotating the pathogenicity of genetic variants

Daniel Quang[1,2,†], **Yifei Chen**[1,†] **and Xiaohui Xie**[1,2,*]

[1]Department of Computer Science and [2]Center for Complex Biological Systems, University of California, Irvine, CA 92697, USA

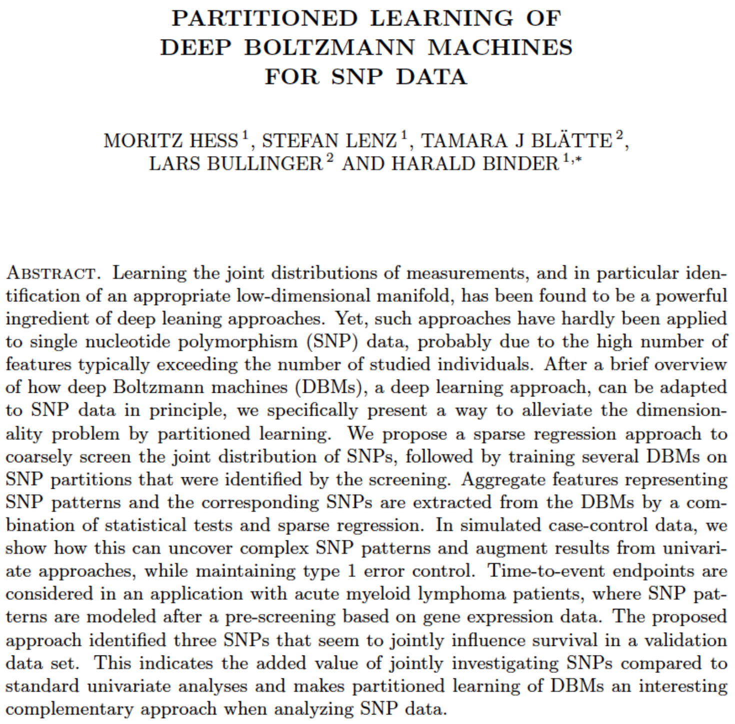
*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Associate Editor: John Hancock

Received on August 11, 2014; revised on October 13, 2014; accepted on October 19, 2014

Abstract

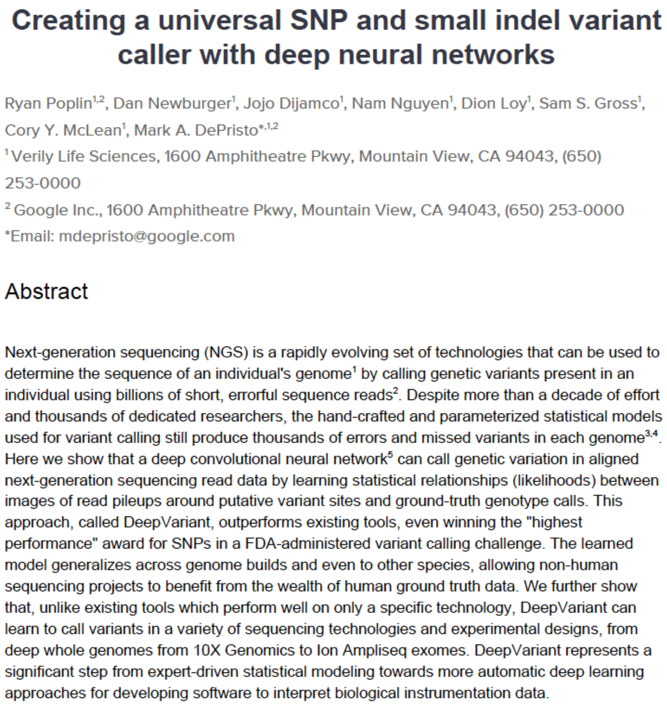
Summary: Annotating genetic variants, especially non-coding variants, for the purpose of identifying pathogenic variants remains a challenge. Combined annotation-dependent depletion (CADD) is an algorithm designed to annotate both coding and non-coding variants, and has been shown to outperform other annotation algorithms. CADD trains a linear kernel support vector machine (SVM) to differentiate evolutionarily derived, likely benign, alleles from simulated, likely deleterious, variants. However, SVMs cannot capture non-linear relationships among the features, which can limit performance. To address this issue, we have developed DANN. DANN uses the same feature set and training data as CADD to train a deep neural network (DNN). DNNs can capture non-linear relationships among features and are better suited than SVMs for problems with a large number of samples and features. We exploit Compute Unified Device Architecture-compatible graphics processing units and deep learning techniques such as dropout and momentum training to accelerate the DNN training. DANN achieves about a 19% relative reduction in the error rate and about a 14% relative increase in the area under the curve (AUC) metric over CADD's SVM methodology.



PARTITIONED LEARNING OF DEEP BOLTZMANN MACHINES FOR SNP DATA

MORITZ HESS[1], STEFAN LENZ[1], TAMARA J BLÄTTE[2],
LARS BULLINGER[2] AND HARALD BINDER[1,*]

Abstract. Learning the joint distributions of measurements, and in particular identification of an appropriate low-dimensional manifold, has been found to be a powerful ingredient of deep leaning approaches. Yet, such approaches have hardly been applied to single nucleotide polymorphism (SNP) data, probably due to the high number of features typically exceeding the number of studied individuals. After a brief overview of how deep Boltzmann machines (DBMs), a deep learning approach, can be adapted to SNP data in principle, we specifically present a way to alleviate the dimensionality problem by partitioned learning. We propose a sparse regression approach to coarsely screen the joint distribution of SNPs, followed by training several DBMs on SNP partitions that were identified by the screening. Aggregate features representing SNP patterns and the corresponding SNPs are extracted from the DBMs by a combination of statistical tests and sparse regression. In simulated case-control data, we show how this can uncover complex SNP patterns and augment results from univariate approaches, while maintaining type 1 error control. Time-to-event endpoints are considered in an application with acute myeloid lymphoma patients, where SNP patterns are modeled after a pre-screening based on gene expression data. The proposed approach identified three SNPs that seem to jointly influence survival in a validation data set. This indicates the added value of jointly investigating SNPs compared to standard univariate analyses and makes partitioned learning of DBMs an interesting complementary approach when analyzing SNP data.



Creating a universal SNP and small indel variant caller with deep neural networks

Ryan Poplin[1,2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Dion Loy[1], Sam S. Gross[1], Cory Y. McLean[1], Mark A. DePristo[*,1,2]

[1] Verily Life Sciences, 1600 Amphitheatre Pkwy, Mountain View, CA 94043, (650) 253-0000

[2] Google Inc., 1600 Amphitheatre Pkwy, Mountain View, CA 94043, (650) 253-0000

*Email: mdepristo@google.com

Abstract

Next-generation sequencing (NGS) is a rapidly evolving set of technologies that can be used to determine the sequence of an individual's genome[1] by calling genetic variants present in an individual using billions of short, errorful sequence reads[2]. Despite more than a decade of effort and thousands of dedicated researchers, the hand-crafted and parameterized statistical models used for variant calling still produce thousands of errors and missed variants in each genome[3,4]. Here we show that a deep convolutional neural network[5] can call genetic variation in aligned next-generation sequencing read data by learning statistical relationships (likelihoods) between images of read pileups around putative variant sites and ground-truth genotype calls. This approach, called DeepVariant, outperforms existing tools, even winning the "highest performance" award for SNPs in a FDA-administered variant calling challenge. The learned model generalizes across genome builds and even to other species, allowing non-human sequencing projects to benefit from the wealth of human ground truth data. We further show that, unlike existing tools which perform well on only a specific technology, DeepVariant can learn to call variants in a variety of sequencing technologies and experimental designs, from deep whole genomes from 10X Genomics to Ion Ampliseq exomes. DeepVariant represents a significant step from expert-driven statistical modeling towards more automatic deep learning approaches for developing software to interpret biological instrumentation data.