Enformer: Effective Gene Expression Prediction from Sequence by Integrating Long-range Interactions[1]

Ding Bai

January 26th, 2023

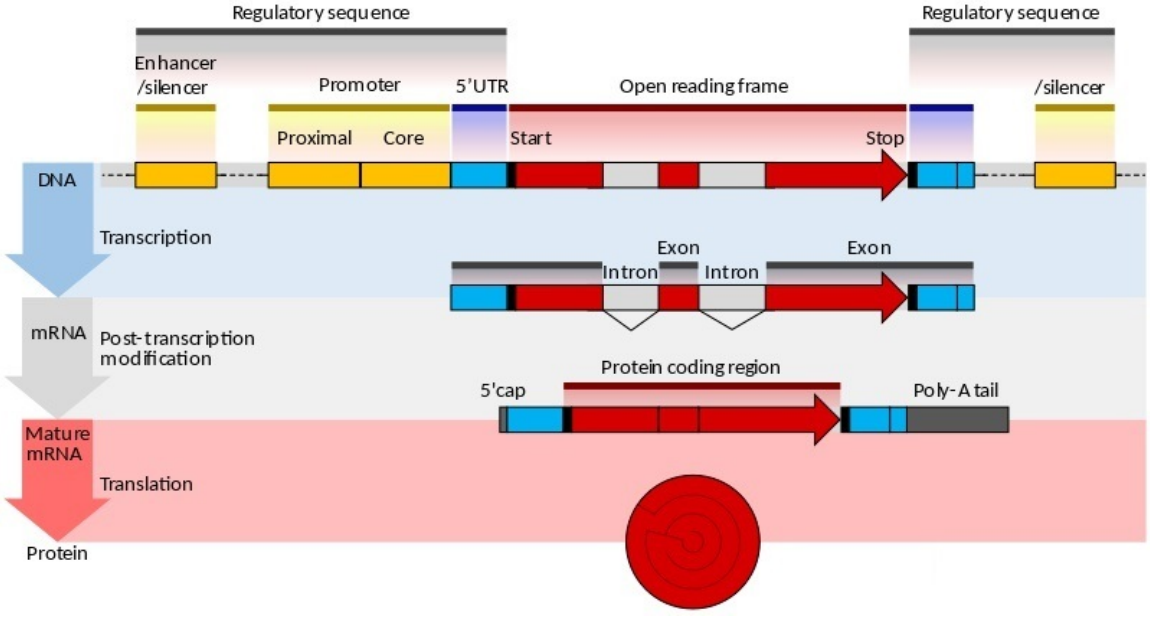
[1] Avsec et al., "Effective gene expression prediction from sequence by integrating long-range interactions".

Outline

Genome Sequences

- ”ATCGTGCAT...ACG”; Every 3 of the sequence determines an amino acid in protein: ”ATC”, ”TCG”, ”CGT”, ... , ”ACG”
- Not every piece of a gene sequence decides the protein.
- In fact, Only about 1 percent of DNA is made up of protein-coding genes; the other 99 percent is noncoding.
- They are called ”Regulatory sequence” if they are useful.
- ”kb”: not kilobyte, but kilobase (Nucleobase base pairs).

From DNA to Protein

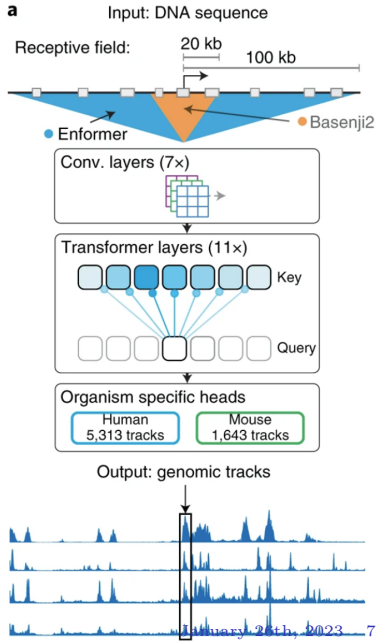


Regulatory sequence

- "TSS": transcription start site (5'UTR shown before)
- Deep CNNs achieve the current state of the art at predicting **gene expression** from **DNA sequences** for the human and mouse genomes[1,2,3,4].
- These models are only able to consider sequence elements up to 20 kb away from the TSS, because the locality of convolutions limits information flow in the network between distal elements.
- Many well-studied regulatory elements (enhancers, repressors, and insulators) can influence gene expression from far greater than 20 kb away [5].
- Thus, increasing information flow between distal elements is a promising path to increase predictive accuracy.

Overall Structure

- Enformer: based on self-attention.
- The machine learning problem: predict thousands of epigenetic and transcriptional datasets in a multitask setting across long DNA sequences.
- Training on most of the human and mouse genomes;
- Observed improved correlation between predictions and measured data.
- More effective use of long-range information, as benchmarked by CRISPRi enhancer assays.
- More accurate predictions of mutation effects, as measured by direct mutagenesis assays and population eQTL studies.

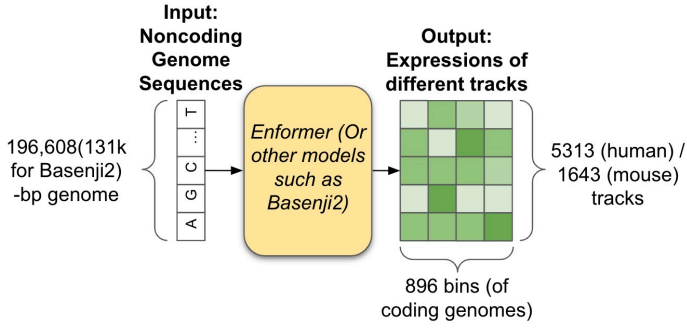


Model architecture

- First, let's see the detailed full picture:
- **Link: Enformer model architecture and comparison to Basenji2**
- The Enformer:
  (1) 7 convolutional blocks with pooling,
  (2) 11 transformer blocks, and
  (3) a cropping layer followed by final pointwise convolutions branching into 2 organism-specific network heads.

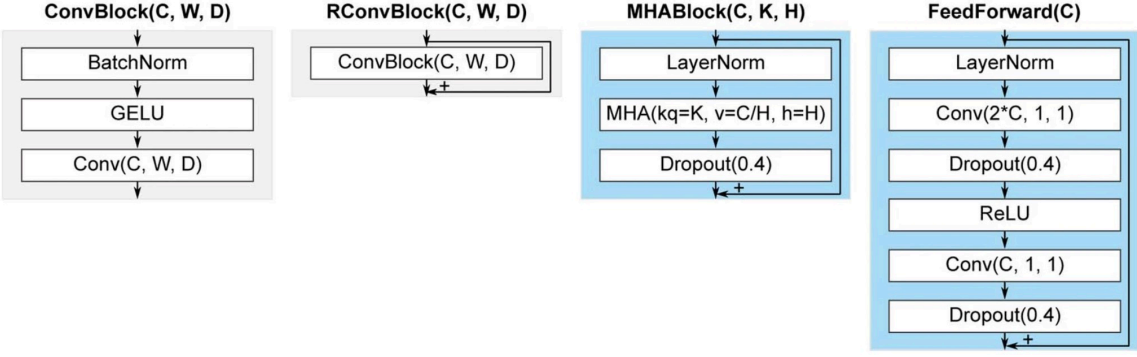
Model architecture: I/O set

- Input: one-hot-encoded DNA sequence (A = [1,0,0,0], C = [0,1,0,0], ... ,N = [0,0,0,0]) of length 196,608 bp
- Output: predicts 5,313 genomic tracks for the human genome and 1,643 tracks for the mouse genome, each of length 896 corresponding to 114,688(= 896 × 128) bp aggregated into 128-bp bins.
- Input noncoding genome sequences; Output expressions of the coding genomes.



Model architecture: Building Blocks

- ConvBlock: Convolutional with batchnorm and GELU activations.
- RConvBlock: Residual ConvBlock.
- MHABlock: multi-head (self-)attention block;
- Feedforward: used for connection.



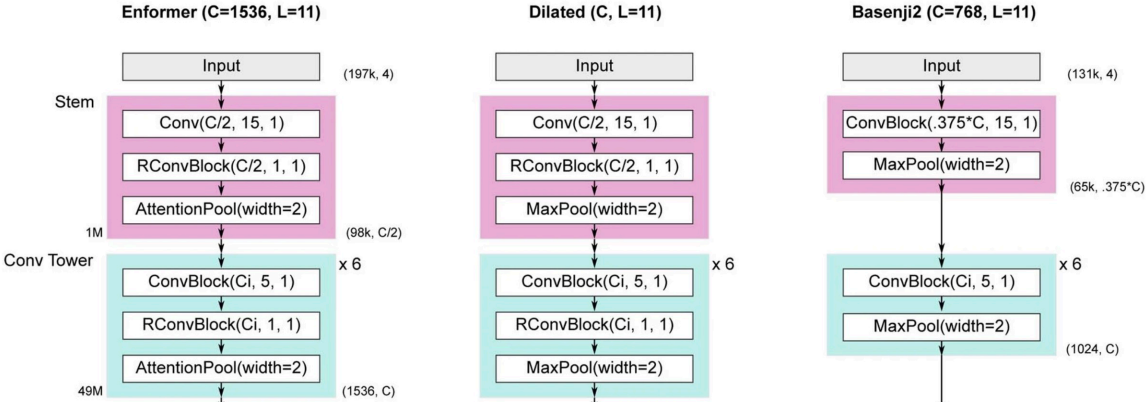
Model architecture: Constants

- Hyper-parameters:
- D, W, C, K, L, H

Constants:
D: Dilation rate
W: Convolutional filter width
C: Number of channels
K: Number of keys in multi-headed attention
H: Number of attention heads
L: Number of layers

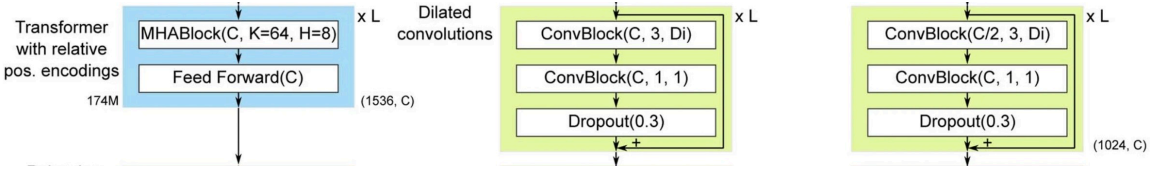
Model architecture: Conv Layers

- The convolutional blocks with pooling first reduce the spatial dimension from 196,608 bp to 1,536
- so that each sequence position vector represents 128 bp (although the convolutions do observe nucleotides in the adjacent pooled regions).



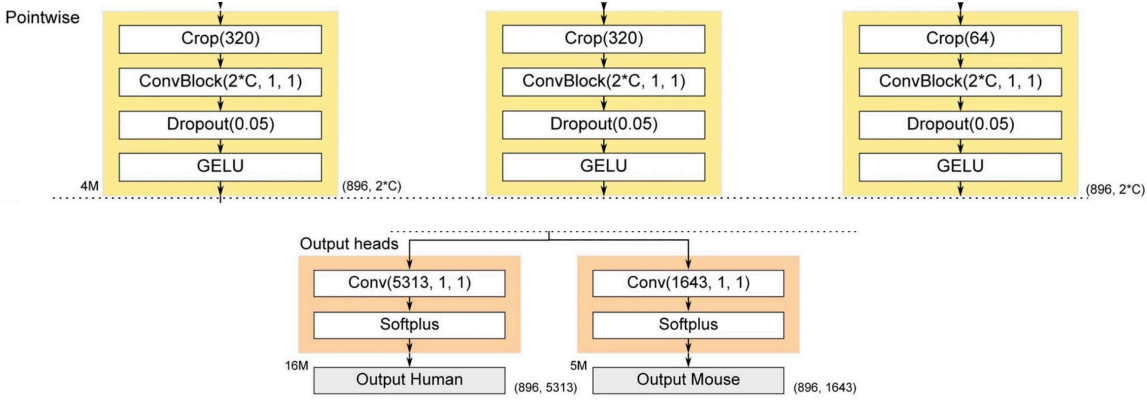
Model architecture: Transformer

- Capture long-range interactions across the sequence.
- Here is the major difference between Enformer and the Basenji2.
- They used 8 heads, value size of 192, and key/query size of 64.
- 128-bp resolution: roughly represents a well-studied length of regulatory elements that contain several motifs



Model architecture: Output Heads

- Cropping layer trims 320 positions on each side to avoid loss on the far ends:
- Far ends are disadvantaged because they can observe regulatory elements only on one side (toward the sequence center) and not the other (the region beyond the sequence boundaries).
- After pointwise part, two output heads predict organism-specific tracks.



Model architecture: Advantages

The Enformer architecture is similar to the state-of-the-art model Basenji2 (ref. 2). However, the following changes helped improve and exceed its performance:

- Enformer uses transformer blocks instead of dilated convolutions;
- attention pooling instead of max pooling,
- twice as many channels,
- 1.5 times longer input sequence (197 kb instead of 131 kb).

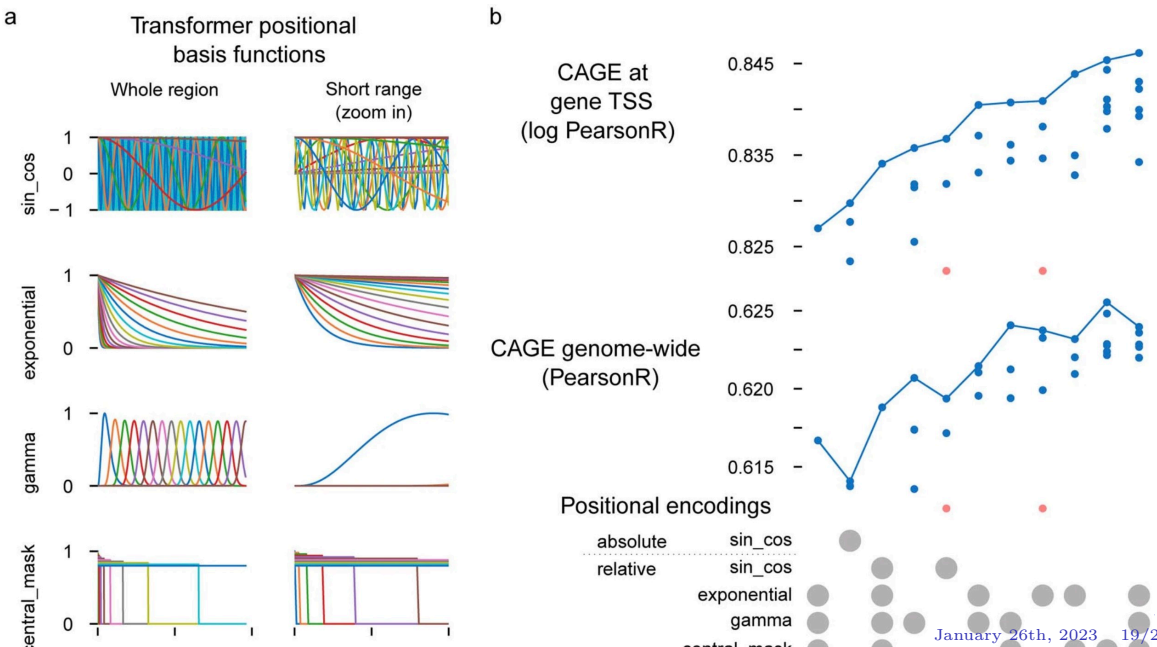
Model architecture: Relative position encoding

- Adding relative positional encodings $\mathbf{R}_{ij}$ to the $\mathbf{q}_i\mathbf{k}_j^T$ attention term.
- $\mathbf{R}_{ij} = \mathbf{q}_i\mathbf{r}_{ij}^T + \mathbf{u}\mathbf{k}_j^T + \mathbf{v}\mathbf{r}_{ij}^T$, where $\mathbf{r}_{ij} = \mathbf{w}^R\mathbf{f}(ij)$ is a linear function of different relative basis functions $\mathbf{f}(ij)$, and $\mathbf{u}$ and $\mathbf{v}$ are the position-agnostic embeddings used to evaluate the preference for specific keys ($\mathbf{u}$) or relative distances ($\mathbf{v}$).
- 3 different relative basis functions: $\mathbf{f}^{exponential}, \mathbf{f}^{centralmask}, \mathbf{f}^{gamma}$.
- For each basis function: symmetric $\mathbf{f}(|x|)$ and asymmetric $\text{sign}(x)\mathbf{f}(|x|)$ version to introduce directionality. The same number of relative positional basis functions as the value size of MHA (192). The 192 basis functions are equally divided among the basis function classes and the symmetric versus asymmetric versions thereof.
- With 3 basis function classes, each basis function class provides 64 positional features (32 symmetric and 32 asymmetric).

Model architecture: relative basis functions

- $f_i^{exponential}(r) = e^{-\log(2)\frac{r}{r_{1/2,i}}}$. Here $r_{1/2,i}$ is placed linearly in the log-space between 3 and sequence length. (Might be a typo here)
- $f_i^{centralmask}(r) = 1\{r \leq 2^i\}$
- $f_i^{gamma}(r) = \Gamma(r|\alpha = \frac{\mu_i}{\sigma^2}, \beta = \frac{\mu_i^2}{\sigma^2})$. The Γ-distribution. μ_i is placed linearly from (sequence length / number of features) to sequence length and σ = sequence length / (2 $\times$ number of features).

Model architecture: Positional encodings plot and comparison



Model training and evaluation: Cross-species sets

- Cross-species training/validation/test sets were constructed using the following procedure to partition homologous sequences into the same set.
- First, divided both the human and mouse genomes into 1 Mb regions. We constructed a **bipartite graph**, in which the vertices represent these regions.
- Next, place edges between 2 regions if they have **>100 kb of aligning sequence** (in the hg38-mm10 syntenic net format alignment downloaded from the UCSC Genome Browser42)
- Finally, partition **connected components** in the bipartite graph randomly into training, validation, and test sets.
- The dataset contains 34,021 training, 2,213 validation, and 1,937 test sequences for the human genome, and 29,295 training, 2,209 validation, and 2,017 test sequences for the mouse genome.

Enhancer prioritization

This work obtained a set of enhancer–gene pairs.

- Each enhancer–gene pair contains a label denoting whether a significant expression change was induced after CRISPRi treatment.
- Three different scores: gradient $\times$ input, attention, and in silico mutagenesis (ISM).
- For each enhancer–gene pair, determine the major TSS of the gene by taking the highest predicted CAGE value in K562 using Enformer.
- Extracted the DNA sequence centered at the main TSS and computed 3 different enhancer–gene scores.

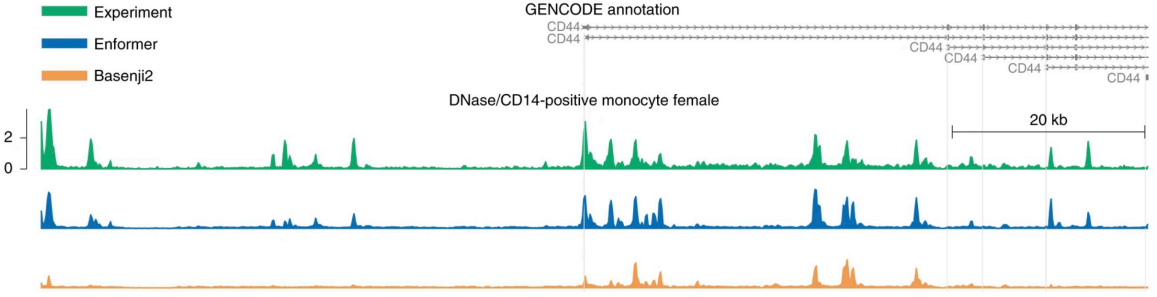
Other related predictions

There are many other related predictions this work made, including classification, benchmark testing, etc.

- GTEx SLDP: predicted the effect of a genetic variant on various annotations;
- Fine-mapped GTEx classification: to study specific eQTLs without needing to consider LD, studied statistical fine-mapping of GTEx v8 using the SuSiE method.
- Benchmarking variant effect predictions on saturation mutagenesis data.

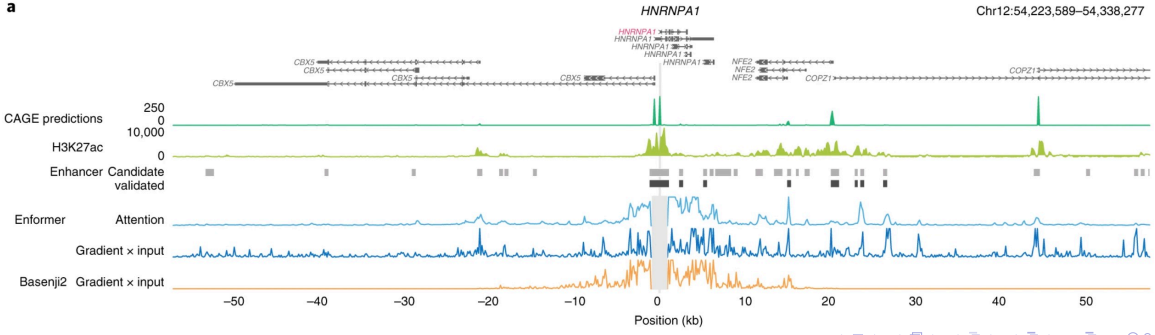
Enformer improves gene expression prediction

- Enformer substantially outperformed the previous best model, Basenji2, for predicting RNA expression as measured by Cap Analysis Gene Expression10 (CAGE) at the TSS of human protein-coding genes,
- With the mean correlation increasing from 0.81 to 0.85 (Fig. 1b, left).
- This performance increase is twice as large as the performance increase between Basenji1 and Basenji2
- closes one-third of the gap to experimental-level accuracy, estimated at 0.94.



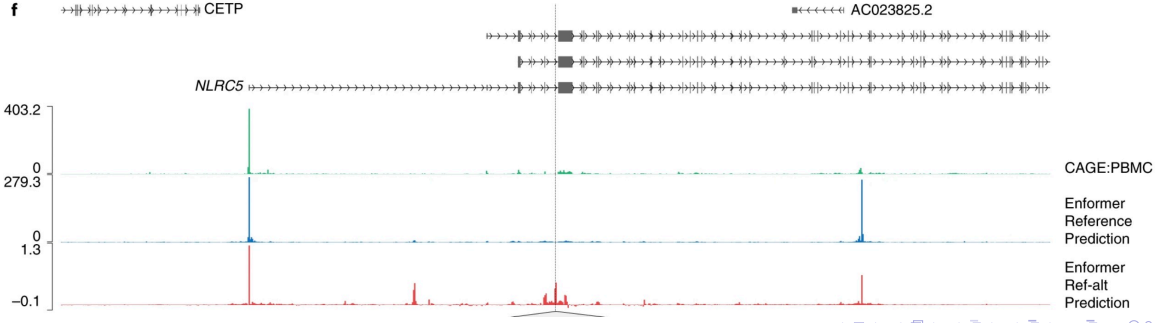
Enformer attends to cell-type-specific enhancers

- Contribution score of distal enhancers: 100kb > 20kb;
- Enhancer-gene pair distance classification performance: Enformer significantly outperforms Basenji2, and is even better than ABC in some cases, of which contribution score depends on experimental data.
- Cell type-specific contribution scores are higher than cell type-agnostic contribution scores, suggesting that the model uses different enhancer sequences in different cell types



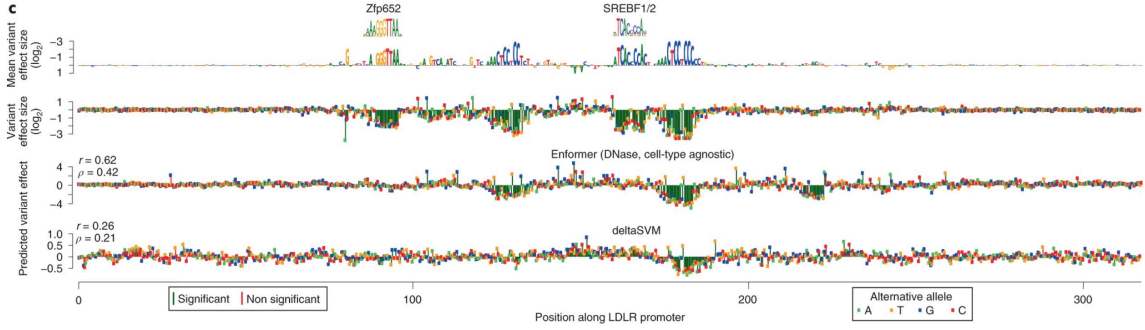
Enformer improves variant effect prediction on eQTL data

- Improved prediction performance: Compared to Basenji2, Enformer has better accuracy in the direction of expression change of finely mapped eQTLs.
- Enformer's variation effect predictions for DNase hypersensitivity are more consistent with GTEx's SLDP than the DeepSEA method used in ExPecto1.
- More accurate identification of causal variants: Using predicted difference vectors to represent variants and train random forest classifiers, Enformer provides more accurate classifiers for most organizations.



Enformer improves MPRA mutation effect prediction

- Lasso regression featuring Enformer predictions has the best average correlation.
- It outperforms the winning team from CAGI5 and captured four transcription factor binding sites of the LDLR locus, which is more comprehensive than deltaSVM and can reflect cell type predictions.



Conclusion

- This work modified Transformer to process longer DNA sequences, thereby effectively simulating the effect of enhancers on long-distance gene expression.
- The Enformer more accurately predicts the effect of variants on gene expression than previous models.
- Enformer represents a major step forward in understanding the complexity of genome sequences, enabling the study of the relationship between disease and biological variation in enhancers.
- Enformer proposes and adopts a new set of position encoding methods to improve the performance of transformer.
- In the future, we envision improving the sensitivity of the Enformer model to genetic variation and promoting the development of diagnostic tools for genetic diseases by analyzing and training more functional genomic data sets.

References

[1] Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. 50, 1171–1179 (2018).
[2] Kelley, D. R. Cross-species regulatory sequence activity prediction. PLoS Comput. Biol. 16, e1008050 (2020).
[3] Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 28, 739–750 (2018).
[4] Agarwal, V. and Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. 31, 107663 (2020).
[5] Gasperini, M., Tome, J. M. and Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet. 21, 292–310 (2020).
[6] Avsec, Ž., Agarwal, V., Visentin, D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 18, 1196–1203 (2021). https://doi.org/10.1038/s41592-021-01252-x