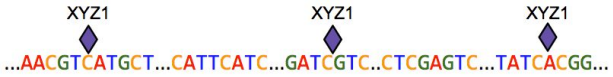
Memory Matching Networks for Genomic Sequence Classification

Jack Lanchantin, Ritambhara Singh, and Yanjun Qi
University of Virginia, Department of Computer Science

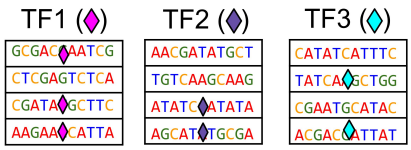
ICLR 2017

1. Genomic Classification Task

Transcription Factors (TFs) are proteins which bind to certain locations of DNA, and in turn regulate gene expression.



Predicting and understanding the **Transcription Factor Binding Sites** (TFBSs), or subsequences where TFs bind is important to biologists.



2. Motifs (Templates) and PWMs

The binding of a TF is triggered by local sequential patterns within TFBSs, known as "**motifs**".

Researchers first predicted TFBSs by constructing motifs using **position weight matrices** (PWMs) which best represented the positive binding sites, and then compared the test sequence to the PWMs. However, constructing good PWMs is difficult.

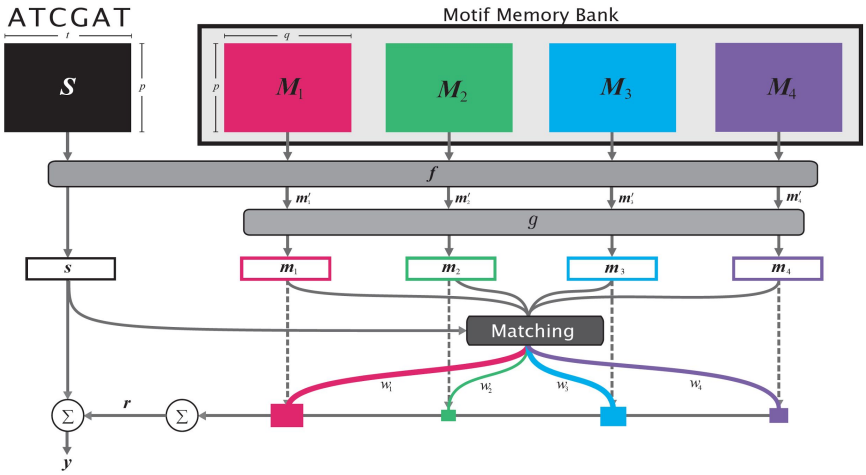
3. Memory Matching Network: Automatically Learning Motif Templates

Matching Networks

In the Matching Network (MN) model by Vinyals et al. (2016), they train a differentiable nearest neighbor model to find the closest matching image of a new unseen image from a support set of images.

Memory Matching Networks

In Memory Matching Networks (MMN), we replace the fixed support set with a memory template support set. MMNs are for a general classification setting (i.e. not one-shot) and we seek to instead learn the support set, which remains constant for every new test classification.



$$
\begin{aligned}
S &= \text{LookupTable}_S(\mathbf{x}) \\
M_i &= \text{LookupTable}_{M_i}(1, ..., q) \\
\mathbf{s} &= f(S) = \text{attLSTM}(S) \\
\mathbf{m}'_i &= f(M_i) = \text{attLSTM}(M_i) \\
\mathbf{m}_{1..\ell} &= g(\mathbf{m}'_1, ..., \mathbf{m}'_\ell) = \text{LSTM}(\mathbf{m}'_1, ..., \mathbf{m}'_\ell) \\
w_i &= \frac{\exp(K[\mathbf{s}, \mathbf{m}_i])}{\sum_j \exp(K[\mathbf{s}, \mathbf{m}_j])} \\
\mathbf{r} &= \sum_i w_i \mathbf{m}_i \\
\mathbf{y} &= W(\mathbf{s} + \mathbf{r})
\end{aligned}
$$

$$
\begin{aligned}
\text{attLSTM}(Z) &= \sum_n \alpha_n \mathbf{h}_n \\
\mathbf{h}_{1...N} &= \text{LSTM}(\mathbf{z}_1 ... \mathbf{z}_N) \\
\alpha_n &= \frac{\exp(\mathbf{h}_n^\top \mathbf{u})}{\sum_j \exp(\mathbf{h}_j^\top \mathbf{u})}
\end{aligned}
$$

4. Results

We use the 61 leukemia cell TF datasets from Alipanahi et al. (2015) which had a training set of at least 10,000 sequences. Each dataset has 1,000 testing sequences with an even positive/negative split. We train a separate model for each TF dataset. The reported AUC statistics are across all 61 datasets. We also show the attention weights of a specific positive "ATF1" sample.