Inferring chromatin factors interactions from Basset neural network

Извлечение взаимодействий факторов хроматина по нейронной сети Basset

Marc Masramon, Aleksandra Galitsyna

Abstract

Open chromatin is a crucial feature of regulation and active transcription. One of the recent advances is the creation of the machine learning method called Basset that identifies the determinants of open chromatin such as protein binding to DNA. It is based on a convolutional neural network (CNN), and its filters have been shown to correspond to various binding motifs of transcription factors (TF). We aim to discover the association between these factors according to the trained neural network and check its presence in the biological databases.

Открытый хроматин -- один из главных признаков регуляции и активной транскрипции в ядре. Недавно детерминанты открытого хроматина, такие как сайты связывания факторов транскрипции, были обнаружены с помощью метода машинного обучения (Basset), основанного на сверточной нейронной сети. Было показано, что фильтры этой сети соответствуют различным сайтам связывания факторов хроматина. Наша цель -- обнаружить ассоциации между парами этих факторов по свойствам обученной нейронной сети и проверить их по базам биологических данных о белках.

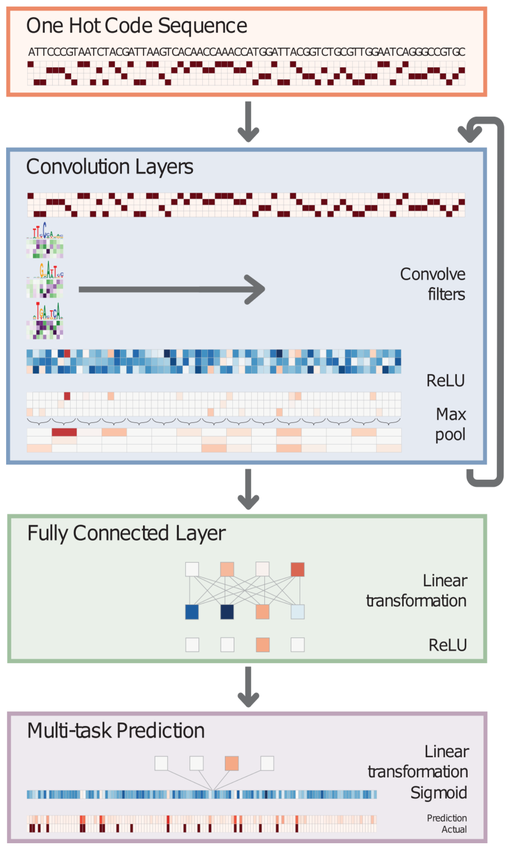


Figure 1. Basset's convolutional neural network diagram (David R. Kelley 2016).

Materials and methods

1. Inference of importance of single filters and their combinations for Basset neural network
2. Basset parameters filtering (Python 3.6)
3. Motifs search (Tomtom)
4. Networks visualisation (Gephi 9.0.2)
5. Databases of biological networks (TRRUST, RegNetworks and STRING)

I = importance of filters to NN prediction

$I^s(i)\ \forall\ i \in \{1, 300\}$ = single importance of filter i

$I^p(i, j)\ \forall\ i, j \in \{1, 300\}$ = observed pairwise importance of filters i and j

$expected(i, j) = I^s(i) * I^s(j)$

$corrected\ expected(i, j) = LR(I^s(i) * I^s(j))$

$enrichment = log(\frac{observed}{corrected\ expected})$

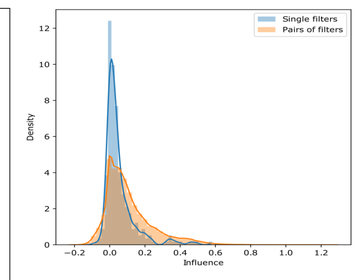


Figure 2. Distributions of the observed importance of single filters and filter combinations.

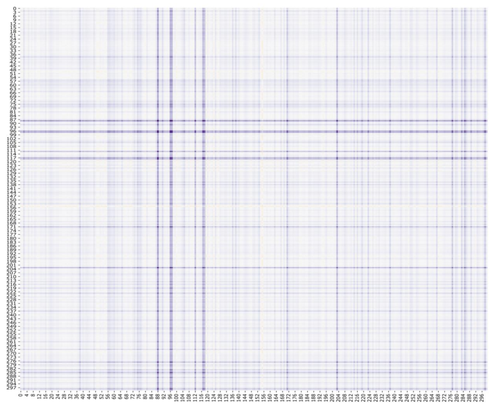


Figure 3. Heatmap of the observed importance of filter pairs.

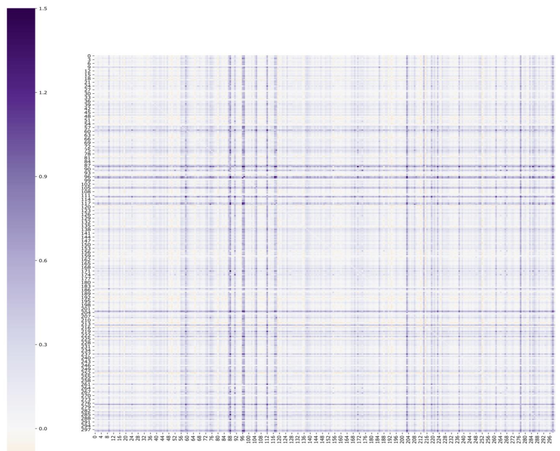


Figure 4. Heatmap of expected importance of single filters under the assumption of filters independence.

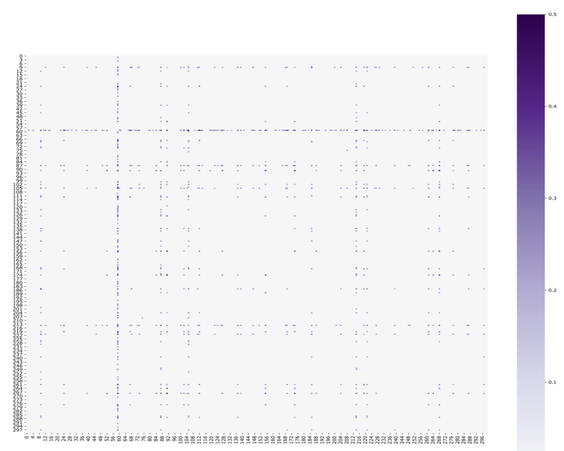


Figure 5. Heatmap of enrichment of filters passing the threshold.

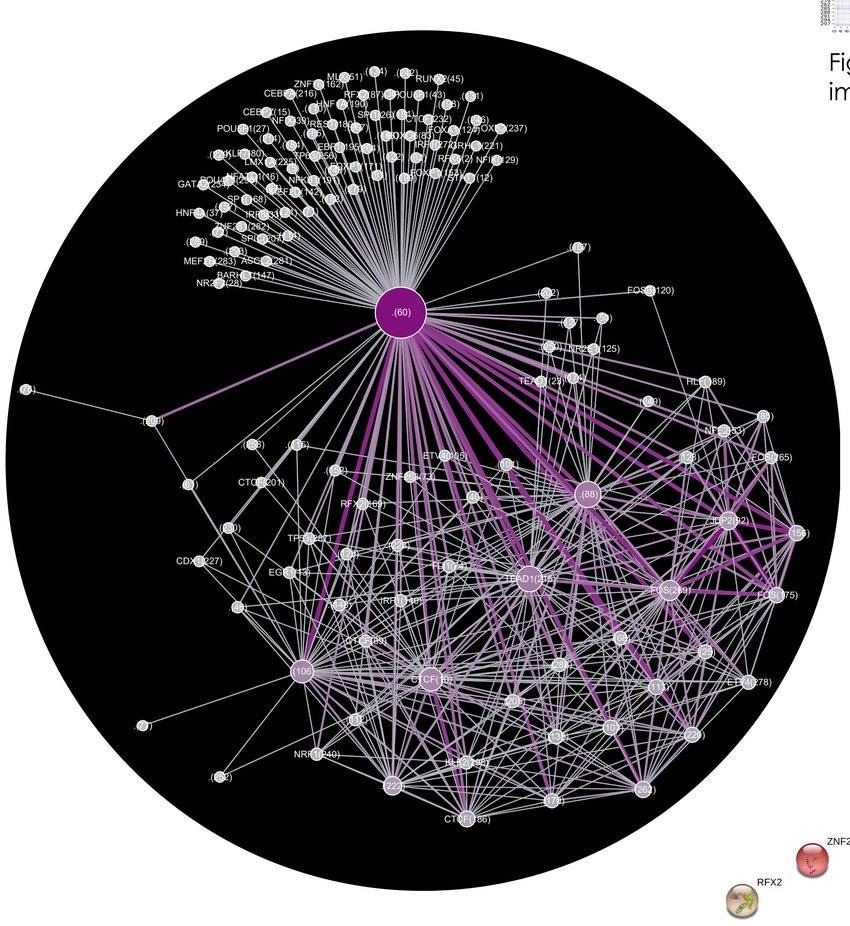
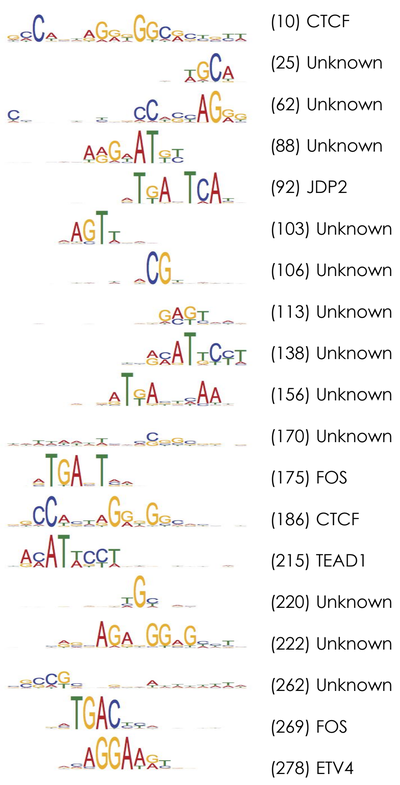


Figure 8. Retrieved network of pairwise importance of filters with corresponding motifs (the legend below).



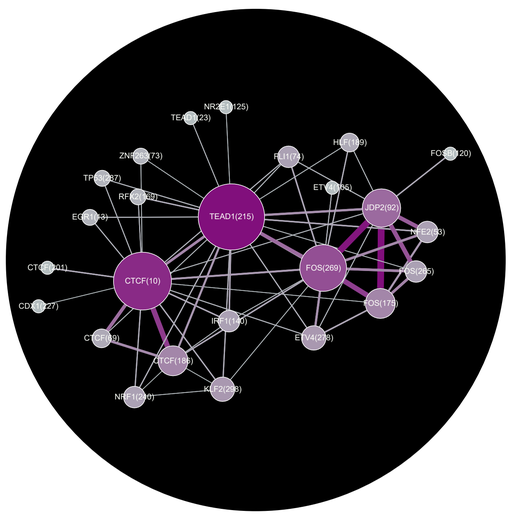


Figure 9. The same network with removed filters having no hits in the CIS-BP database.

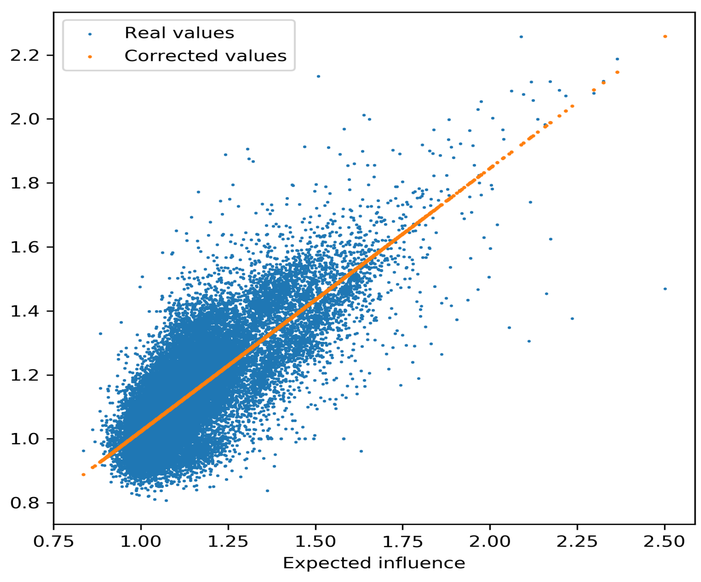


Figure 6. Retrieval of corrected expected importance using linear regression.

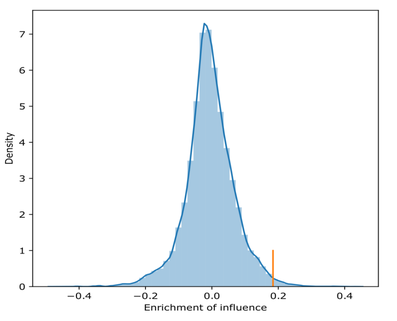


Figure 7. Distribution of the enrichment and the threshold (top 1%).

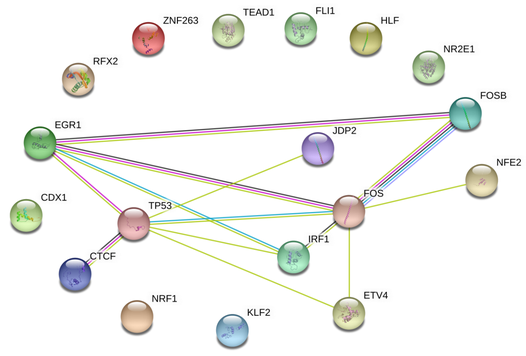


Figure 10. Network of protein-protein interactions from the STRING database (for selected factors from Fig. 8).

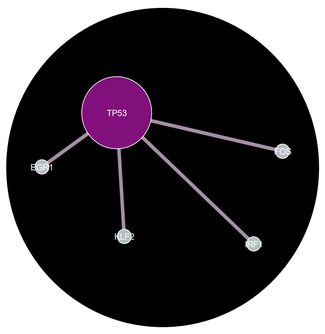


Figure 11. The TRRUST network of regulatory relationships of factors constructed by text mining of Pubmed articles (for selected factors from Fig. 8).

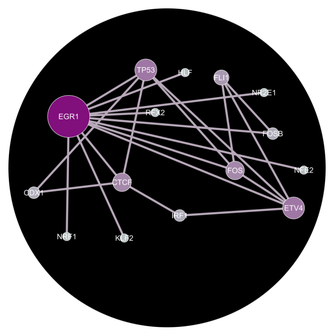


Figure 12. Network of regulatory relationships of factors from the RegNetwork curated database of predicted and validated experimentally relations (for selected factors at Fig. 8).

Results

1. The importance of BASSET filters and their combinations was retrieved, the filters were interpreted as DNA motifs (fig. 2-7).
2. The pairs of filters that demonstrate enriched importance over the background were visualised as networks(fig. 8,9).
3. Networks were compared with the TRRUST, RegNetworks and STRING databases (fig. 10-12).

Conclusions

1. Most pairwise important filters do not correspond to known TF motifs.
2. Annotated filters do not follow the pattern of interactions of factors in biological databases.

Further plans

1. Comparison of filters motif similarity with their pairwise importance
2. Testing of alternative background normalization methods.
3. Comparison of pairwise influences of motifs to co-occurrences of respective binding motifs.

Laboratory of Comparative and Functional Genomics
Лаборатория Сравнительной и Функциональной Геномики