Targeting Transcription Factor Cascades for Precision Oncology

Anonymous Authors[1]

Abstract

Transcription factor (TF) cascades play an emerging role in various diseases, including cancer. Modulation of TF-driven activating cascades could have implications in developing personalized therapeutics in the setting of cancer care. In this study, we have generated a database of 81,488 Transcription Factor Cascades using 426 genes. We then performed a detailed exploratory data analysis (EDA) to understand the underlying patterns of the cascades. To identify and prioritize essential transcription factors in the cascades as drug targets, we calculated centrality measures using the theory of complex graphs after generating a graph representation of TF cascades. Finally, biological pathway and disease enrichment analysis of the TF-cascades helped to identify plausible biological mechanisms and disease associations from a database of 2 million data points. Collectively, our work represents the introduction of a new translational bioinformatics database with potential applications to cancer genomics and the development of precision therapeutics.

1. Introduction

1.1. Transcription Factor

Transcription factors (TF) are sequence-specific DNA binding proteins that regulate the rate of transcription of genetic information flow from the DNA to RNA. Protein families such as kinases, methylases, co-activators, histone deacetylases, histones acetyltransferases, and chromatin remodelers also regulate genes but lack a DNA-binding domain. TFs exert control over specific pathways such as immune responses and cell developmental patterning, used in the laboratory for cell differentiation, de-differentiation, and trans-differentiation. Mutations in TFs and their binding sites are one of the underlying causes of many human diseases.

1.2. Transcription Factor Cascades

Transcription factors bind to DNA sequence and regulate another transcription factor. This activated transcription factor in turn goes on to regulate a third transcription factor, thus creating cascades of gene expression. This multi-step process results in amplification of the initial signal and provides a regulatory relationship among TFs resulting in a high level of control over the expression of the target gene.

One of the common examples of TFs cascade is IRF8-KLF4 which has been reported to have implications in human diseases.

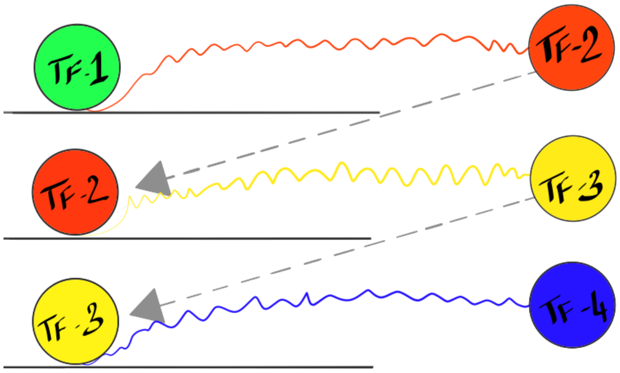


Figure 1. Visual graphic to depict a transcription cascade with L3; Transcription factor-1(TF-1) binding to the DNA and activating TF-2 which induce the expression of TF-3 which in turn regulates TF-4

1.3. Significance of TF-Cascades Dataset

Despite their role in cancer, no studies till date have identified TF cascades as therapeutic targets, perhaps due to the lack of information and appreciation for their role in cancer treatment. So, there was an urgent need to make

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML) Workshop on Computational Biology. Do not distribute.

a bio-curated knowledge base of transcription factor cascades and their associated pathways. (Karamouzis, 2011; Hagenbuchner, 2016)

2. Materials and Methods

2.1. Dataset

Data of all the protein interactions was first taken from the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database. The STRING database contains all the known interactions of approximately 5090 species. Human protein data was extracted from the STRING database, which includes TFs.

The data of TFs interaction was obtained by filtering a list of known TFs from the Human protein data. The TFs list had ENSEMBLE ID, so it needed to be converted into protein stable IDs to compare with our protein interactions. For that BioMart database was utilized, as it enables the user to convert one database ID to another ID by using state of the art algorithm to cross reference the IDs. This easy-to-use web-based tool allows extraction of data without any programming knowledge or understanding of the underlying database structure.

2.2. Exploratory Data Analysis

We used a data mining approach called Exploratory Data Analysis (EDA) to analyse the database and summarise their main characteristics as it is difficult to look at each feature of a dataset and determine important characteristics.. EDA uses visual methods for finding the underlying patterns in the data thus, enabling us to see what the data tells us. This is a preliminary analysis commonly used before modeling tasks.

Several graphical and non-graphical methods were used for analysis. We confined our exploration to univariate analysis, as we are mainly concerned with categorical variables.

2.2.1. Univariate Analysis

Univariate analysis is a type of quantitative analysis, where the data being analysed consists of only one variable. It's a simple analysis model and more descriptive in understanding the data. We used univariate analysis to describe the data and gather better insights from the data. Multiple bar charts and box plots were created to describe the data.

2.2.2. Cramér's V (c)

We used Cramér's V to understand the intercorrelation of two discrete variables. Cramér's V is an association measure for nominal random variables. The coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating perfect association. The empirical estimators used for Cramér's V have been proved to be biased, even for large samples.

2.3. Graph Analytics

In statistics, graph theory is the study of graphs, which are mathematical structures used to model pairwise relations between objects. A graph is a combination of vertices (nodes) and edges. G = (V, E) where V represents the set of all vertices and E represents the set of all edges of the graph.

We created a connected graph out of the TF interactions dataset which we created. This was done to better visualise and understand the high scoring TFs, which are the nodes in the graph.We can then study the most influential TFs for drug discovery and targeted therapy.

2.4. Enrichment Analytics

We performed gene set enrichment analysis for functional exploration of the Transcription Factor (TF) cascades and to identify the associated disease phenotypes. We retrieved the functional profile of the cascades in order to better understand the underlying biological pathways. (Kuleshov MV, 2016)

3. Results

3.1. EDA Summary

We created a comprehensive website compiling all the analysis results and published it site at:

`https://sonishsivarajkumar.github.io/TFCascades/`

There were 81,488 unique TFs cascades created having a maximum length of 62.

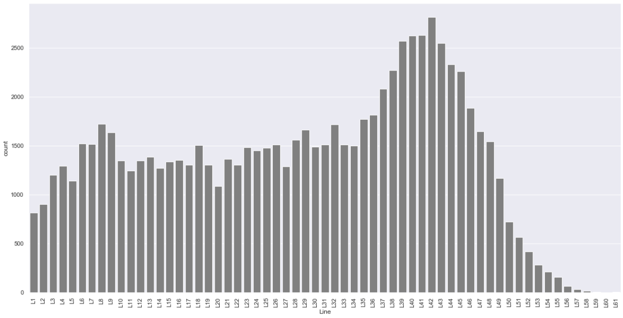
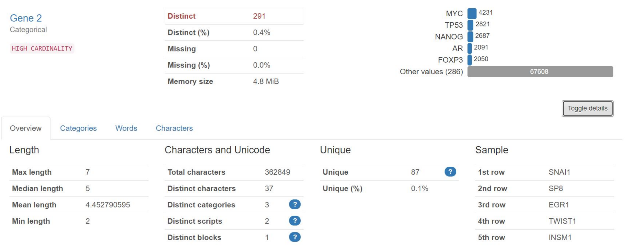


Figure 2. Distribution of cascades across different cascade levels. For instance, Line 1 contains 800 with 2 activational transcription factors.

Univariate Analysis results showed the most occurring genes in each column, which can serve as the initial analysis for all the following studies in the dataset.



Gene 2
Categorical
HIGH CARDINALITY

Distinct	291
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	4.8 MiB

MYC	4231
TP53	2821
NANOG	2687
AR	2091
FOXP3	2050
Other values (286)	67608

Length

Max length	7
Median length	5
Mean length	4.452790595
Min length	2

Characters and Unicode

Total characters	362849
Distinct characters	37
Distinct categories	3
Distinct scripts	2
Distinct blocks	1

Unique

Unique	87
Unique (%)	0.1%

Sample

1st row	SNAI1
2nd row	SP8
3rd row	EGR1
4th row	TWIST1
5th row	INSM1

Figure 3. Exploratory Data Analyis report for Gene Level 2 in https://sonishsivarajkumar.github.io/TFCascades/

3.2. Graph Analytics Summary

Graph analysis was performed on the TF cascades, to find the relationships and interactions between the cascades. A connected graph was generated by connecting all the TFs as nodes. Edges indicate the cascade between two TFs. The graph had 426 nodes and 866 edges, which corresponds to 426 unique genes in humans. Average degree per node was analyzed and there is a linear growth initially, but later it shows no growth and becomes constant which means that the network complexity increases with the increase in number of nodes but after a threshold value the average node degree shows no growth

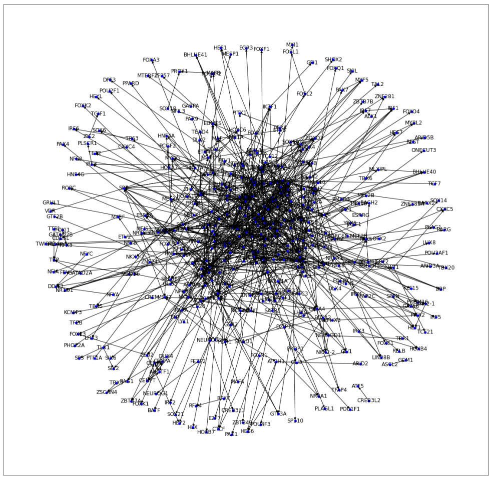


Figure 4. Cascade network representated as a graph; nodes are blue dots (n=426) and edge are black lines (n=866)

The next step was to find the most important nodes in the network. To find the high scoring nodes we calculated different mathematical indicators from the theory of complex graphs:

- **Betweenness centrality**: This indicator can detect individuals who influence the transfer of information.The betweenness centrality of a node v is given by the expression:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

- **Closeness Centrality**: This indicator makes it possible to detect the individuals who have a significant power on the transfer of information. Closeness was defined by Bavelas (1950) as the reciprocal of the farness,that is:

$$C(x) = \frac{1}{\sum_y d(y,x)} \tag{2}$$

  where $d(y,x)$ is the distance between vertices x and y, and N is the number of nodes.

- **Eigenvector centrality**: The individuals having a high spectral centralized are the individuals who have the most relation in the network, they are central and have influence in a general way on the network.For a given graph $G := (V, E)$ with $|V|$ vertices let A= $(a_{v,t})$ be the adjacency matrix, i.e. $(a_{v,t})$=1 if vertex v is linked to vertex t , and $(a_{v,t})$=0 otherwise. The relative centrality score of vertex v can be defined as:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t \tag{3}$$

STAT3 gene came out to be most associated in the network, with an eigenvector centrality value of 0.366 and *MYC* the next, with a value of 0.304, followed by *TP53* and *NANOG* with values 0.261 and 0.240 respectively.[Table 1]*KCNIP3* was found to be least influential, with an eigenvector centrality value of -6.20E-18, followed by *KCNIP3*, *PROP1*, *RFX7*, *GABPA*.[Table 2]

3.3. Enrichment Analysis Results

Enrichment Pathway analysis was performed on all cascades We compared the TF cascades dataset with the Kyoto Encyclopedia of Genes and Genomes(KEGG) human pathway database and generated the final TF cascades reference database, containing 2 million associated pathways.Table 3 shows a sample of the enriched data.

This reference database helps researchers to identify the TF cascades associated with each pathway.Complete analysis, including their p-value scores, are generated along with the data.The challenge of data interpretation and hypothesis

Table 1. Top 10 most associated TFs and their centrality measures

ID	Degree	betweenness_centrality	clust_coefficient	closeness_centrality	eigenvector_centrality
STAT3	40	0.138	0.0756	0.353	0.366
MYC	41	0.169	0.0354	0.354	0.304
TP53	39	0.146	0.026	0.348	0.261
NANOG	25	0.067	0.097	0.328	0.240
POU5F1	21	0.048	0.157	0.316	0.218
FOXO3	16	0.040	0.125	0.329	0.174
SOX2	16	0.0116	0.2	0.298	0.173
SNAI1	18	0.069	0.078	0.327	0.165
KLF5	12	0.017	0.212	0.308	0.154
FOXM1	15	0.046	0.105	0.315	0.153

Table 2. Top 10 least associated TFs and their centrality measures

ID	Degree	betweenness_centrality	clust_coefficient	closeness_centrality	eigenvector_centrality
KCNIP3	1	0	0	2.392E-3	-6.20E-18
PROP1	1	0	0	2.392E-3	-5.35E-18
RFX7	1	0	0	2.392E-3	-4.74E-18
GABPA	1	0	0	2.392E-3	-4.63E-18
NEUROG1	1	0	0	2.392E-3	-4.18E-18
ASCL2	1	0	0	2.392E-3	-3.69E-18
RFX3	1	0	0	2.392E-3	-3.52E-18
PAX1	1	0	0	2.392E-3	-2.10E-18
TTF1	1	0	0	5.468E-3	-1.14E-18
IRX3	1	0	0	3.189E-3	-8.80E-19

Table 3. A glimpse of enrichment analysis results

Term name	P-value	Z-score	Combined score	Adjusted p-value	cascade
acute myeloid leukemia	5.193E-3	391.117	2057.429	9.716E-3	[’STAT3’, ’NANOG’]
adipocytokine signaling pathway	7.086E-3	284.686	1409.030	9.716E-3	[’STAT3’, ’NANOG’]
small cell lung cancer	8.482E-3	237.071	1130.783	1.009E-2	[’STAT3’, ’TP53’]
prostate cancer	8.581E-3	234.271	1114.690	1.009E-2	[’STAT3’, ’TP53’]
cell cycle	1.003E-2	193.155	882.437	1.115E-2	[’STAT3’, ’TP53’]

generation is thus addressed by the integration of pathway annotations and biological processes.

4. Discussions and Future Scope

Cancer treatments have been advancing in the last few years with targeted agents against tumor cells.But the possibilities of transcriptomic level targets for cancer therapy is less explored. Transcription factor cascades and their pathways can be intersected with many diseases to investigate the possibilities of targeting the cascades for transcriptomic level treatment and discovering new drugs.

Gene regulatory activity of distinct Transcription factors and their cascades can be modified by development of drugs and is an exciting field with tremendous potential. The understanding of TF cascades and allied pathways provides promising targets for novel treatment strategies. With the advancements in artificial intelligence and big data, the data can be utilised for development of efficient transcription factor-targeted pharmaceuticals.

References

Claudia Villicaña, e. The basal transcription machinery as a target for cancer therapy. *Cancer Cell International*, 14, 2014.

Hagenbuchner, e. Targeting transcription factors by small compounds–current strategies and future implications. *Biochemical pharmacology*, 107(1-13), 2016.

Karamouzis, e. Transcription factor networks as targets for therapeutic intervention of cancer: the breast cancer paradigm. *Molecular medicine (Cambridge, Mass.)*, 17 (11-12):1133–6, 2011.

Kuleshov MV, e. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 44(W1):W90–7, 2016.

Peters LA, e. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nature genetics*, 49(10):1437–1449, 2017.

Shameer K, e. Interpreting functional effects of coding variants: challenges in proteome-scale prediction, annotation and assessment. *Briefings in bioinformatics*, 17(5): 841–62, 2016.