# Analysis of Gene Regulatory Networks Inferred from ChIP-seq Data

Eirini Stamoulakatou(✉), Carlo Piccardi, and Marco Masseroli

Dipartimento di Elettronica, Informazione e Bioingegneria,
Politecnico di Milano, 20133 Milan, Italy
{eirini.stamoulakatou,carlo.piccardi,marco.masseroli}@springer.com

**Abstract.** Computational network biology aims to understand cell behavior through complex network analysis. The Chromatin Immuno-Precipitation sequencing (ChIP-seq) technique allows interrogating the physical binding interactions between proteins and DNA using Next-Generation Sequencing. Taking advantage of this technique, in this study we propose a computational framework to analyze gene regulatory networks built from ChIP-seq data. We focus on two different cell lines: GM12878, a normal lymphoblastoid cell line, and K562, an immortalised myelogenous leukemia cell line. In the proposed framework, we preprocessed the data, derived network relationships in the data, analyzed their network properties, and identified differences between the two cell lines through network comparison analysis. Throughout our analysis, we identified known cancer genes and other genes that may play important roles in chronic myelogenous leukemia.

**Keywords:** Biomolecular networks · Transcription factors · ChIP-seq · Next-Generation Sequencing · Cancer · Bioinformatics

## 1 Introduction

In biological sciences, network analysis is becoming one of the main tools to study complex systems. Networks used to represent the regulation of gene expression are known as Gene Regulatory Networks (GRNs) [1]. In network biology, particularly in disease/cancer research, comparisons are often performed on GRNs [2] and DNA co-methylation networks [3], obtained from the gene expression and DNA methylation profiles of healthy and disease tissues.

Here, we focus on normal and cancer GRNs that, differently from other works, we inferred from Chromatin Immuno-Precipitation sequencing (ChIP-seq) data. ChIP-seq is a Next-Generation Sequencing (NGS) technique designed to study, map and understand protein-DNA interactions on a genome-wide scale. It provides measurements of epigenetic (transcription factor and histone) regulation of genes, retaining all the advantages of the NGS technology thanks to its coverage, high resolution and cost-effectiveness. Our goal is to study the relationship between gene-related epigenetic factors and genes in a normal vs. disease case,


I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 319–331, 2019.
https://doi.org/10.1007/978-3-030-17938-0_29

possibly leading towards the discovery of novel molecular diagnostic and prognostic signatures. Particularly, we focused on two immortalized human cell lines, K562 and GM12878; they are both from blood tissue, the first one (K562) from chronic myelogenous leukemia, whereas the second one (GM12878) from normal lymphoblastoid cells.

A major contribution of this work is the study of the relation between epigenetic transcription factors and human protein-coding genes in K562 and GM12878 cell lines in the view of complex network comparison. This was possible by defining relationships between transcription factors and protein-coding genes to create gene regulatory networks. Another major aspect of this study is the creation of a computational framework with appropriate network comparison methods, according to our network characteristics, to extract differences and similarities of the compared networks. The defined comparison models are fully "data-driven", as they do not take into consideration any form of prior biological knowledge. Finally, using our analytic framework, we highlighted behaviours directly emerging from the data, drawing insights that could drive further biological investigations.

## 2 Used Data Sets

Among the numerous publicly accessible available genomic databases, we chose the following two: the ENCyclopedia Of DNA Elements (ENCODE) and GENCODE [4]; the former one as source for the NGS experimental data, the second one for the gene annotations we used. GENCODE genomic samples are organized as General Feature Format (GTF) text files, whose structure is described in [4]. Each of their lines refers to a genomic feature annotation and is made up of several tab-separated fields. The first eight fields are standard GTF fields that convey information about the feature chromosome, annotation source, feature type, start and stop genomic coordinates, score, strand, and genomic phase. The ninth field is actually a sequence of key-value pairs made up of further information about the feature.

Biosamples involved in the sequencing experiments generating our considered data came from two immortalized cell lines, namely K562 and GM12878. These two cell lines are among the most investigated ones in the ENCODE project [5], being the object of a large number of sequencing experiments from research labs all over the world, each identifying thousands of epigenetic events through the whole genome. Both cell lines belong to human blood tissue, in particular: K562 cell line consists in a chronic myeloid leukemia (CML) cell line [6], GM12878 cell line is made up of lymphoblastoid cells [7].

## 3 Analysis Framework

Networks provide a theoretical framework that allows a convenient conceptual representation of interrelations among a large number of elements. Furthermore,

they usually allow framing questions about the behavior of the underlying represented system, by applying well-established analyses on the network representing the considered data. Here, we focus on cell line specific gene regulatory networks, where source nodes represent genes encoding transcription factors (TFs), whereas target nodes are any genes. A link exists between a source TF encoding gene and a target gene if the encoded TF binds the target gene promoter; the links are weighted, and the weight represents the power of the binding.

We propose a network analysis framework to characterize commonalities and differences in behavior across normal GM12878 cells and cancerous K562 cells, using ChIP-seq datasets. We evaluate if some genes display extreme behaviors, and whether or not such behaviors highlight aspects of the underlying biology. The proposed framework includes the following steps: (1) High quality data extraction from NGS and genomic annotation datasets, through the GenoMetric Query Language (GMQL); (2) Transformation of the extracted metadata and genomic region data to adjacency matrixes, representing the most valuable information and the data relationships extracted; (3) Numeric characterization of each network structure through 8 topological measures; (4) Application of comparison methodologies to identify the most common and different gene connections.

### 3.1 Data Acquisition and Preprocessing

For the data acquisition and preprocessing, we chose GMQL [8] as the most suitable tool. GMQL is a high-level declarative query language, specifically designed for genomic data retrieval and processing. The GMQL portal[1] publicly provides reasonably high computational and storage capabilities and, moreover, it hosts up-to-date GENCODE and ENCODE data, among others. This last aspect allowed us to just write a GMQL query to perform the complete extraction and filtering of the genes' epigenetic status data described below, without the need to download the related data files from the GENCODE and ENCODE public repositories and write specific programs to extract the relevant data. In the following paragraphs we describe the usage of GMQL to filter and extract the highest quality epigenetic status data from ENCODE.

The goal of the defined GMQL query is to map transcription factors of the two cell lines on each gene promoter region. Thus, the first step is the selection of the transcription factors and the promoter regions. The ENCODE consortium has defined and implemented a system of 'audits', i.e., flags meant to give additional, yet essential, quality information about the provided experimental data to the research community. To extract high-quality data, we did not consider all the experiment data files labeled with at least one of the following audits: *extremely low read depth*, *extremely low read length*, or *insufficient read depth*. Furthermore, to consider only data from higly reproducible NGS ChIP-seq experiments, we selected only the called peak data files labeled as *conservative IDR threshold peaks*. Finally, in the case of more replicate data files from the same

[1] http://www.gmql.eu/.

transcription factor targeting experiments, we chose to only consider one data file for each transcription factor, the one with the largest number of called peaks. By choosing the peak set with the highest cardinality, we retain a larger amount of information, still being confident of its reasonably good quality thanks to the foregoing audit-based and reproducibility-aware filtering performed.

In our study we are exclusively interested in promoter regions of human known protein-coding genes, i.e., genomic regions around the starting position of a gene transcript. Therefore, an important aspect is to consider the right position along the human genome of each transcript of all genes of interest. The process of identifying and designating locations of individual genes and transcripts on the DNA sequence is called genome annotation. One of the most important active projects about human genome annotation is GENCODE.[2] Thus, for the promoter region extraction we chose GENCODE repository annotations, specifically the GENCODE v24 release version and the annotation type transcript; so, an annotation file for transcript isoforms was selected, reporting all the transcript start sites (TSSs) of each human gene. In order to build the promoter regions from the transcript isoforms, we used the typical −2k/+1k base interval around their first base. All the selected transcription factor binding regions are then mapped to the considered gene promoters. As a gene can have more than one promoters, we selected for every TF only the gene with the highest signal value. The dataset created by the performed GMQL mapping operation provides a matrix-like structured outcome, ideal for subsequent data analysis. In particular, we created such a dataset/matrix for each considered cell line, where the matrix rows represent transcription factors, columns represent genes, and each matrix cell contains a value that represents the maximum binding signal of a TF in a gene promoter. To create the gene regulatory network from the above data, we finally considered each TF as representing its encoding gene, thus obtaining a gene adjacency matrix for each cell line.

### 3.2 Gene Regulatory Network Analysis

A primary aspect in gene regulatory networks is to capture the interactions between molecular entities from high-throughput data. The GRNs that we constructed are weighted directed networks, where nodes represent genes and links between nodes exist solely if the regulatory element, a transcription factor encoded by a source gene, binds a target gene promoter.

The problem of detecting significant dissimilarities in paired biological networks is different from popular graph theory problems, like graph isomorphism or subgraph matching, for which various graph matching and graph similarity algorithms exist and have been also applied on biological networks [9,10]. Several approaches to compare gene regulatory networks constructed from healthy and disease samples have been developed [11,12]. The majority of them focuses on the comparison of the entire networks, using statistics that describe network global properties [13]; but these statistics are not sensitive enough to detect smaller,

[2] https://www.gencodegenes.org/.

yet important, differences. On the other hand, there are numerous alignment-based methods that compare networks using the properties of the individual nodes, e.g., local similarity [14]. The aim of these methods is to identify matching nodes, and use these nodes to identify exact subnetwork matches. These approaches are computational intensive, as exact graph matching is NP-hard. In addition, alignment-free comparison methods exist, which have been used to identify evolutionary relationships [15]. These methods are based on the fact that differences in network structure is essential, as structural properties of biological networks can bring important biological insights, such as determining the relationships between protein functions from protein interaction network topology. To achieve network structure comparison, they count the occurrences of subgraph shapes in the local neighbourhoods of all nodes in a network [16].

Our created networks have a peculiar structure, mainly due to the fact that ChIP-seq experiment data exist only for a limited set of TFs; thus, in our GRNs the number of source nodes (TFs) is much lower (about 100) than the number of the target nodes (human protein-coding genes, about 19,000). This makes difficult to directly apply reliably the methodologies mentioned above. On the other hand, motifs and modules have long been identified as important components of biological networks [3]; thus, we focused on looking for strongly connected components (SCCs) in each considered network, and on evaluating the one-step ego-nets in each SCC. So, we avoid comparing the entire networks, and concentrate on their most informative nodes. The one-step ego-net of a node/gene $g$ is the (sub)network consisting of all the nodes within one edge distance from $g$, also including all the edges between those nodes. For directed graphs, as in our case, a node $g$ ego-net contains the $g$ "out" neighborhood, i.e., in our case the genes where $g$ points to and their connections. To analyse the ego-nets of each SCC of the two networks under comparison, we applied standard approaches such as pairwise (on matching nodes) metrics to quantify similarity based on network properties, discover specific features, and detect anomalous nodes/genes.

The state-of-the-art offers a well-established set of graph metrics for complex networks. The most important metrics for a detailed analysis of a weighted directed network have been previously described in [2]; they are used in the current study and here summarized. The *degree* of a node is the total number of edges incident to it. Thus, the average value of the network degrees, measured over all network nodes, is called the *average degree* of the network, as we handle directed graph we computed in and out degrees. For the *total weight*, we sum the weights of all the edges of the graph. The *diameter* of a network is the maximal distance between any pair of nodes in the network. The *modularity* measures to what extent the network is structured in communities. It takes values between 0 and 1; a higher modularity means a stronger division between well-delimited communities, i.e., subnetwotks with large internal edge density but weakly connected each other, while a lower modularity means that no such subnetwork exist. The metric that quantifies the degree correlation, i.e. to what extent nodes with large degree are connected to nodes with large degree, is called *assortativity*. The network's heterogeneity can be measured by the *degree distri-*

*bution entropy*. As *principal eigenvalue* we denote the largest eigenvalue of the weighted adjacency matrix of the network. For each node/gene, the *connectivity* is defined as the sum of the connection strengths with the other nodes/genes of the network.

Our proposed analysis method compares not the networks themselves, but instead the ensemble of all gene neighborhoods (ego-nets) in each SCC of the networks, through a pairwise approach. This idea of using the content of subgraphs to build a comparison method between networks arises from the fact that modules are important biological network components.

The statistical comparison measures we used in our method are the following:

- The *cosine similarity* ($CS$), a measure of similarity between two vectors: it expresses the cosine of the angle between them, not from the perspective of magnitude, but from that of orientation. The resulting similarity between the two vectors ranges from $-1$, meaning exactly opposite directions, to 1, meaning exactly the same direction, with 0 usually indicating independence. This measure is applied in our context by building a vector with elements consisting of each metric of interest measured on the graph.
- The *Jaccard index* ($J$), a statistic used for comparing the similarity of sample sets. It measures similarity between finite sample sets, defined as the size of the intersection divided by the size of the union of the sample sets. The Jaccard index always gives a value between 0, which means no similarity, and 1, for identical sets. In our study we used the Jaccard similarity pairwise for each matching node of the SCCs of the two compared networks. For each node we built a set of its ego-net edges, with the edges being represented as an object (source node, target node) since the networks are directed. This measure gave us the percentage of similarity between the matched genes based on their interactions with the other genes, and ranges from 0 to 100.
- The *fidelity metric* $\phi$, another network similarity measure, computed following the approach proposed in [17]. It is a statistical formula that generates a single value to summarize the similarity between two sets of properties/topological features (which characterize two entities of the same nature).

Additionally, in our network comparison analysis we included the identification of patterns for neighborhoods (ego-nets) of the normal and cancer networks, and the report of deviations, if any, as proposed in [18]. The detection of outliers is intimately related with the pattern discovery: only if the majority of nodes closely obey to a pattern, we can then confidently consider as outliers the few nodes that deviate. In order to detect the patterns and the outliers of the SCCs of the normal and cancer networks, we selected and grouped the topological features of the ego-nets into pairs, where we expect to find patterns of normal behavior and point out anomalies that deviate from the patterns. All methods presented here were implemented using Python programming language and its pyGMQL [28] and Networkx [29] packages.

## 4 Results

Here, we present and discuss the results obtained for our considered normal and cancer cell line networks, using two distinct network analysis approaches: *single-network analysis* and *differential network analysis*, which answer different questions. In our context, the single-network analysis aims at identifying both the key genes (i.e., hub genes) and the similarities in the binding behavior of the TFs present in a given data set. Conversely, the differential network analysis aims to uncover similarities and differences in the TFs of the two data sets. More specifically, using feature vectors with the aforementioned statistical measures, we evaluated the similarity of the TFs present in both data sets, and also we identified common behavior trends and outlier nodes for the two cell lines.

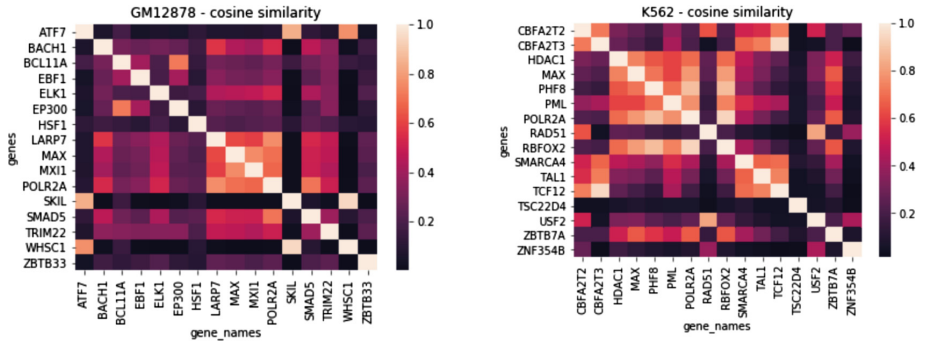
### 4.1 Single Network Analysis

The two weighted directed networks constructed, one for the normal GM12878 and one for the cancer K562 cell line, were individually analyzed. Both resulted having a single giant strongly connected component (SCC), with only TF encoding genes (source nodes), and a single out-component, i.e., a set whose nodes are reachable with a directed path from the SCC, with about 90% of the network nodes, including a few TF encoding genes not in the SCC. Table 1 reports the topological feature values measured for the two networks and their SCCs.

Curiously, in both networks the most important (hub) nodes, identified using the page-rank algorithm [19], were mitochondrial genes. The TFs with largest degree were identified using the reverse page-rank algorithm (applying page-rank to the networks obtained by reversing the directions of all links). For the cancer network they were ATF7, RBFOX2, ATF1, NFIC, NRF1, PKNOX1, RFX1, VEZF1 and L3MBTL2, whereas for the normal network they were IKZF1, ELF1, FOXK2, PKNOX1, ZNF143 and BHLHE40. IKZF1 is a leukemia tumor suppressor associated with chromatin remodeling, with also increasing evidence that IKZF1 loss also affects signaling pathways that modulate therapy response [20]. Also ELF1 is a key transcription factor in the regulation of genes involved

**Table 1.** Topological feature values for the two networks and their SCCs.

| Features | K562 | GM12878 | K562-SCC | GM12878-SCC |
|---|---|---|---|---|
| Nodes | 18,732 | 18,732 | 230 | 111 |
| Isolated nodes | 2,312 | 4,305 | - | - |
| Source nodes | 238 | 115 | 230 | 111 |
| Edges | 923,025 | 481,704 | 20,320 | 5,556 |
| Average degree in/out | 56.261 | 33.384 | 88.343 | 55.051 |
| Assortativity | −0.054 | −0.043 | −0.021 | −0.011 |
| Diameter | 4 | 4 | 4 | 3 |
| Modularity | 0.29 | 0.34 | 0.27 | 0.33 |




(a) GM12878 (b) K562

**Fig. 1.** Heatmaps showing the cosine similarity between TFs in the two cell lines.

in hematopoiesis [21]. PKNOX1 is a Hox co-factor, whose function alteration is directly linked to hematopoiesis and leukemia. ZNF143 is also involved in leukemia development [22].

Using the cosine similarity function pairwise, we identified the TFs with similar behavior in each network, i.e., that bind the same genes with similar strength. Figure 1 reports the cosine similarity heatmaps we created for some of such TFs; values closer to 1 show greater similarity. The heatmaps clearly show some clusters with high similarity in each network. For the GM12878 one, LARP7, MAX, MXI1 and POLR2A were the TFs with greatest similarity; conversely, ATF7, SKIL and WHSC1 had totally different bindings with respect to the other TFs. In the K562 cancer network, HDAC1, MAX, PHF8, PML, RBFOX2 and POLR2A created a cluster of similarity, and SMARCA4, TAL1 and TCF12 another one. The first cluster TFs resulted enriched in the *Homo sapiens transcriptional misregulation in cancer* KEGG pathway. TSC22D4 and ZNF354B resulted the TFs with the greatest dissimilarities to the others.

**Table 2.** Topological features for ego-nets of normal and cancer cell line SCCs.

| Features | GM12878 | | K562 | |
|---|---|---|---|---|
| | Average | Range (min; max) | Average | Range (min; max) |
| Nodes | 33 | (2; 67) | 34 | (4; 67) |
| Edges | 798 | (1; 2,153) | 855 | (8; 2,287) |
| Average degree in/out | 18 | (0.5; 32) | 19.5 | (2; 34) |
| Total weight | 215,000 | (21; 532,000) | 273,000 | (316; 618,000) |
| Density | 0.633 | (0.471; 1.500) | 0.600 | (0.511; 1.150) |
| Degree entropy | 3.121 | (0.630; 4.101) | 3.330 | (1.307; 4.105) |
| Assortativity | −0.212 | (−0.500; −0.011) | −0.188 | (−0.370; −0.060) |
| Principal eigenvalue | 9,869 | (0; 14,899) | 29,290 | (69; 39,138) |
| Connectivity | 15,872 | (1,002; 126,000) | 18,389 | (1,578; 164,000) |

### 4.2 Differential Network Analysis

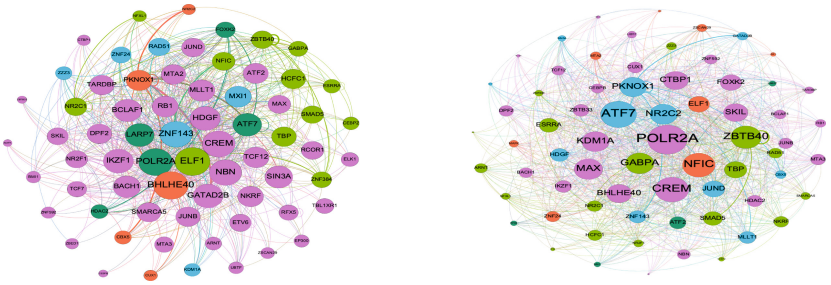
For the network comparison analysis we focused on the single SCC in each of the two networks, considering only the TFs whose data were available for both cell lines, i.e., 68 TFs. The average, minimum and maximum values of the ego-net features extracted for such TFs are reported in Table 2; no relevant differences in the global features were found between the two cell lines.

The obtained global results led us to apply the comparison methods at local level in order to highlight differences, if existing. As a first approximation, we simply checked which were the most different TFs between the normal and cancer cell lines. Using the Jaccard, cosine and fidelity similarity measures, we computed pairwise similarity scores for every pair of TFs. Despite the global topological features showed relatively similar values in both cell lines, at local level we discovered interesting dissimilarities (data not shown). To further explore the topological differences among the two cell lines, we characterized the structure of the ego-net extracted for each TF using the same 9 standard measures for network topology as in Table 2. These measures capture important characteristics of a network structure, which in part determines its functionality. In particular, we sought to detect the structural heterogeneity among TFs. For each ego-net of a TF, we created a feature vector with these feature values, which we used for pairwise cosine and fidelity similarity between each pair of TF/ego-nets. The cosine similarity, however, proved to be not a good metric, as all results were close to 1 (identical). In addition, we applied Jaccard similarity using as input the TF/ego-nets edges, this metric demonstrated to be a good method. Most different TFs found, according Jaccard similarity, are in Table 3.

All these TFs, except of BACH1, appeared to have greater activity in K562 than in GM12878 cell line (data not shown). Interestingly, CTBP1 appeared to bind strongly in the cancer cell line, but it had only a bond in the normal cell line data. An explanation of this behavior may be that, according to KEGG, CTBP1 is a leukemia cancer gene. In the same context, ZBTB33 and CEBPB are

**Table 3.** Similarity values, from three different statistical measures of the most different TFs in the compared SCCs according to Jaccard similarity.

| Transcription factors | *Jaccard* (%) | $\phi$ | $CS$ |
|---|---|---|---|
| CTBP1 | 0.0 | 0.56 | 0.99 |
| ZBTB33 | 0.22 | 0.58 | 0.99 |
| CEBPB | 0.37 | 0.57 | 0.99 |
| NR2C2 | 0.41 | 0.55 | 0.99 |
| KDM1A | 0.52 | 0.52 | 0.97 |
| BACH1 | 0.80 | 0.17 | 0.99 |
| BCLAF1 | 1.19 | 0.34 | 0.96 |




(a) GM12878 (b) K562

**Fig. 2.** Graphical representation of the compared SCCs. Colors denote the community [26]; node size is according to node degree.

responsible for cancer-driven myelopoiesis, which promotes cancer progression [23,24]. KDM1A plays an important role in hematopoiesis and was identified as a dependency factor in leukemia stem cell populations [25]. BCLAF1 is in the 6q23.3 cytogenetic location, a genomic region that has been reported to exhibit a high frequency of deletions in tumors such as lymphomas and leukemias. The relation of NR2C2 and BACH1 functions to cancer progression remains unclear.

We also performed pathway enrichment analysis of the communities we identified in the SCCs (Fig. 2) using the Louvain algorithm [26]. In the two largest communities in K562, which include 70% of the SCC nodes, the enriched KEGG pathways were the *Homo sapiens p53 signaling pathway* and *chronic myeloid leukemia* pathway. In the largest community of the GM12878 SCC, it was the *MAPK signaling pathway*; according to [27], the activation of this pathway is essential for the antileukemic effects of dasatinib, a target therapy used to treat certain cases of chronic myeloid leukemia.

Finally, using the approach of [18] we tried to identify TFs with significant anomalous behavior in the two SCCs. Using the number of nodes and edges, we were able to detect if the ego-nets of the TFs had a star or connected (complete) shape, i.e., minimal or maximal density. Upper diagrams in Fig. 3 show that, in both cell lines, all TFs created almost complete ego-nets, except CTBP1 that bound only one TF gene. The total weight and the number of edges detected TFs with considerable higher total edge weight compared to the number of edges in their ego-net. As shown in Fig. 3 (lower diagrams), PKNOX1, ZBTB40, NR2C1/2, FOXK2 and BCLAF1 bound with stronger connections other TF genes in both cell lines. Interesting result from this analysis is that the number of nodes and the number of edges of the ego-nets as well as the number of edges and the total weight follow power-law, as we can observe from the linear function in log-log scale.

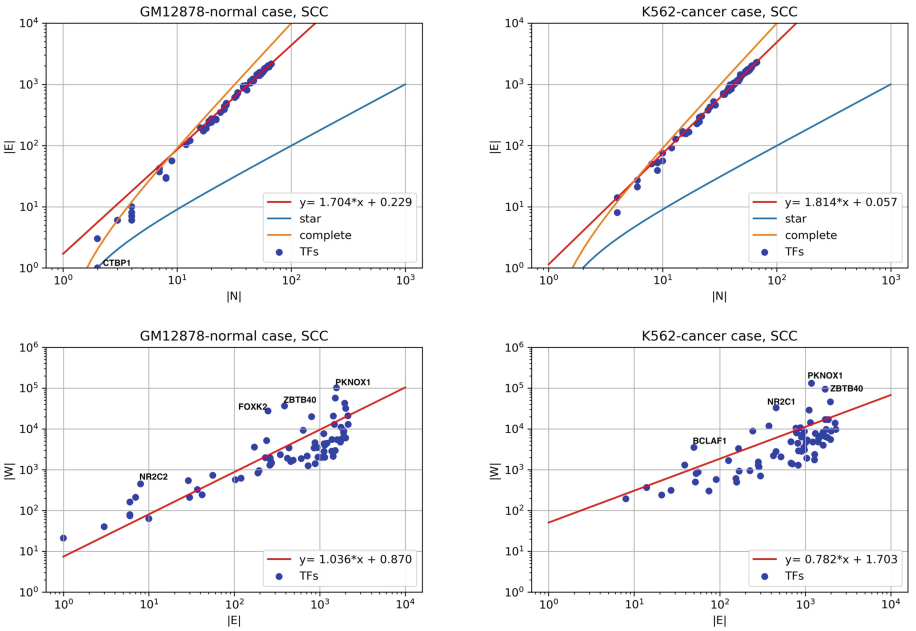



**Fig. 3.** (a) Ego-net edge count ($|E|$) vs. node count ($|N|$). Red line: linear function fit on median values; blue line: (N-1) function, star graphs, whereas orange line is the N(N-1) function, complete graphs, where n is the number of nodes. (b) Total weight ($|W|$) vs. total count ($|E|$) of edges in the ego-nets for all nodes. (Color figure online)

## 5 Conclusions

In this manuscript we have shown how to build gene regulatory networks from ChIP-seq data, and how to evaluate them individually or comparatively when built from a normal and a cancer cell line. Through our analysis, we explored the characteristics of the two compared cell lines and identified differences in their transcription factor functions. As a future work, we will explore further the biological meaning of our results trying to evaluate them using gene expression data and we will extend our analysis to more cell lines.

**Acknowledgments.** This work has been supported by the ERC Advanced Grant 693174 "Data Driven Genomic Computing" (GeCo).

## References

1. Rodríguez-Caso, C., et al.: On the basic computational structure of gene regulatory networks. Mol. Biosyst. **5**(12), 1617–1629 (2009)
2. Zhang, B., et al.: A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. **4**, Article no. 17 (2005)

3. Zhu, X., et al.: Getting connected: analysis and principles of biological networks. Genes Dev. **21**(9), 1010–1024 (2007)
4. Harrow, J., et al.: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. **22**(9), 1760–1774 (2012)
5. Wang, J., et al.: Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. **22**(9), 1798–1812 (2012)
6. Lozzio, C., Lozzio, B.B.: Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood **45**(3), 321–334 (1975)
7. ENCODE Cell Types (2018). https://genome.ucsc.edu/encode/cellTypes.html. Accessed 28 Dec 2018
8. Masseroli, M., et al.: Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. Bioinformatics (2018, in press)
9. Przulj, N., et al.: Biological network comparison using graphlet degree distribution. Bioinformatics **23**(2), 177–183 (2007)
10. Yang, Q., Sze, S.: Path matching and graph matching in biological networks. J. Comput. Biol. **14**(1), 56–67 (2007)
11. Choi, J.K., et al.: Differential co-expression analysis using microarray data and its application to human cancer. Bioinformatics **21**(24), 4348–4355 (2005)
12. Fuller, T.F., et al.: Weighted gene co-expression network analysis strategies applied to mouse weight. Mamm. Genome **18**(6–7), 463–472 (2007)
13. Ratmann, O., et al.: From evidence to inference: probing the evolution of protein interaction networks. HFSP J. **3**(5), 290–306 (2009)
14. Phan, H.T., et al.: PINALOG: a novel approach to align protein interaction networks-implications for complex detection and function prediction. Bioinformatics **28**(9), 1239–1245 (2012)
15. Liu, X., et al.: New powerful statistics for alignment-free sequence comparison under a pattern transfer model. Theor. Biol. **284**(1), 106–116 (2011)
16. Waqar, A., et al.: Alignment-free protein interaction network comparison. Bioinformatics **30**(17), 430–437 (2014)
17. Topirceanu, A., et al.: Statistical fidelity: a tool to quantify the similarity between multi-variable entities with application in complex networks. Int. J. Comput. Math. **94**(9), 1787–1805 (2016)
18. Akoglu, L., McGlohon, M., Faloutsos, C.: oddball: Spotting anomalies in weighted graphs. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS (LNAI), vol. 6119, pp. 410–421. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13672-6_40
19. Page, L., et al.: The PageRank citation ranking: bringing order to the web. In: 7th International World Wide Web Conference, pp. 161–172 (1999)
20. Marke, R., et al.: The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. Haematologica **103**, 565–574 (2018)
21. Larsen, S., et al.: The hematopoietic regulator, ELF-1, enhances the transcriptional response to Interferon-$\beta$ of the OAS1 anti-viral gene. Sci. Rep. **5**, 17497 (2015)
22. Ngondo-Mbongo, R.P., et al.: Modulation of gene expression via overlapping binding sites exerted by ZNF143, Notch1 and THAP11. Nucleic Acids Res. **41**(7), 4000–4014 (2013)
23. Koh, D.I., et al.: KAISO, a critical regulator of p53-mediated transcription of CDKN1A and apoptotic genes. Proc. Natl. Acad. Sci. USA **111**(42), 15078–15083 (2014)
24. Hirai, H., et al.: Non-steady-state hematopoiesis regulated by the C/EBP$\beta$ transcription factor. Cancer Sci. **106**(7), 797–802 (2015)

25. McGrath, J.P., et al.: Pharmacological inhibition of the histone lysine demethylase KDM1A suppresses the growth of multiple acute myeloid leukemia subtypes. Cancer Res. **76**(7), 1975–1988 (2016)
26. Lambiotte, R., et al.: Fast unfolding of communities in large networks. J. Stat. Mech. **2008**, P10008 (2008)
27. Dumka, D., et al.: Activation of the p38 Map kinase pathway is essential for the antileukemic effects of dasatinib. Leuk. Lymphoma **50**(12), 2017–2029 (2009)
28. Nanni, L., et al.: Exploring genomic datasets: from batch to interactive and back. In: Proceedings ExploreDB, pp. 1–6 (2018)
29. Hagberg, A.A., et al.: Exploring network structure, dynamics, and function using NetworkX. In: Proceedings 7th Python in Science Conference, pp. 11–15 (2008)