Assessment of Transcription Factor Binding Motif and Regulon Transfer Methods

Sefa Kilic and Ivan Erill

University of Maryland Baltimore County
Department of Biological Sciences
1000 Hilltop Circle, Baltimore, Maryland 21250
{sefa1,erill}@umbc.edu

Abstract. Despite its fundamental importance in comparative genomics studies, the impact of motif transfer methods remains largely unstudied. With the recent increase in availability of transcription factor binding site data from traditional and high-throughput experiments, it has become possible to assess existing comparative genomics approaches as well as to benchmark newly developed ones. In this study, we describe three different transfer methods that define transcription factor binding motif in a target species given some regulatory activity information in a reference species. We evaluate these methods and report their performances on identifying binding sites, binding motif and regulon for a given genome.

1 Introduction

Comparative genomics has been leveraged in many studies to characterize transcriptional regulatory networks [1–4]. By analyzing the degree of conservation of functional elements across multiple genomes, comparative genomics analyses make it possible to reconstruct regulatory networks in multiple species. Thanks in large part to high-throughput experimental techniques (e.g., ChIP-seq [5]), available experimental binding site data has increased dramatically over the last few years and it has become possible for the first time to reliably assess methods used for regulatory network reconstruction.

Given a transcription factor (TF) and some reference information on its regulatory activity, the main steps of transcriptional regulatory network reconstruction are (a) transferring the available information (i.e., known binding sites and regulated genes for a TF) from reference species to define binding motif in target species, (b) searching target species with transferred motif to estimate its regulatory network (or regulon) and (c) filtering false positives via comparative analysis of putative target sites. In the transfer step, the goal is to define a putative motif based on a reference one. This motif is then used to search the genome for putative sites. The final step, called "consistency check" [6], is based on the idea that true sites are likely to be present upstream of orthologous regulated genes, while false positives should be scattered randomly and not consistently across genomes [7–14].

In this study, we focus on the first step of the comparative genomics pipeline: the transfer method. To evaluate different methods we define the goal as the ability to identify binding sites, binding motif and regulon for a TF in a genome, given the collection of experimentally validated binding sites for the same TF in a reference genome. This can be achieved using the known motif or network structure as prior information. Motif-based transfer is performed using the reference binding motif to search for putative binding sites in the target genome. The underlying assumption is that, for a given TF, the binding motif is relatively well conserved across closely related species. This method has been shown to perform well at inferring existing regulatory networks in previously uncharacterized genomes [7, 8, 15, 16]. The other source of prior information that can be used is the regulatory network itself. The putative regulon is then constructed based on orthologous transfer of the reference regulon and *de novo* motif discovery is performed on the promoter regions of putatively regulated target genes [17–23].

2 Materials and Methods

2.1 Data

Binding site data were compiled mainly from CollecTF, a database of experimentally verified TFBS in Bacteria built by our group [24]. As of October 2014, CollecTF had 4,942 experimentally verified binding sites and associated gene regulation data for 229 TFs in 134 species, from 921 publications. Additional data were incorporated from RegulonDB [25], CoryneRegNet [26], DBTBS [27] and MtbRegList [28]. The data from such databases were downloaded and merged after removal of duplicates and of data without experimental evidence. Table 1 shows the distribution of binding sites in the compiled data by database.

Table 1. Number of experimentally validated binding sites by database

Database	Number of binding sites
CollecTF	4,942
DBTBS	116
MtbRegList	202
CoryneRegNet	196
RegulonDB	2,147

Complete genome sequences and annotations for species that have binding site data were downloaded from NCBI RefSeq database. Operon predictions are based on the DOOR database [29]. For binding site search, the regions spanning from -300 bp to +50 bp relative to the corresponding gene translation start site are used. Orthologs were detected using reciprocal BLAST [30].

2.2 Direct Transfer

The most straightforward motif transfer approach is the direct transfer using the collection of known binding sites from a model species [14, 31]. Given a collection of experimentally determined sites in the reference species, a position-specific scoring matrix (PSSM) [32] is built and used to scan the promoter regions of the genome of interest to identify putative sites.

It is crucial to determine a threshold for PSSM search accurately. A low threshold may classify most of the true binding sites correctly while producing many false positives, whereas a high threshold is likely to miss many true positives. One approach for threshold selection is to compute score distribution of the PSSM and to specify a significance threshold [33–35]. Other commonly used approaches are to define a threshold based on PSSM scores of known binding sites [36, 37] or to choose it arbitrarily [37]. Another approach is to select a fixed amount of highest scoring sites as putative binding sites [38], essentially assuming that the size of the regulatory network is conserved to a first approximation. For all motif transfer methods in this study, the first N_T highest scoring sites are selected as putative sites, $N_T = \alpha N_R G_T / G_R$ where N_R is the number of true sites in the reference species, G_T and G_R are genome lengths for target and reference species, respectively. α is used as a scaling factor, used to increase sensitivity or specificity depending on the value of α. It should be noted that any bias introduced by this approach should be averaged out as the transfer methods are tested both ways (i.e., using species A as reference and B as target, and vice versa).

2.3 PSSM Search Followed by Motif Discovery

Another way of defining the binding motif in a target species is to perform motif discovery on pre-searched candidate sequences [39]. First, the genome of interest is scanned for putative target sites using the reference PSSM. A motif discovery algorithm (e.g., MEME [40]) is then applied to the promoters of high scoring sites. The motivation for this method is to capture motifs that are slightly different from the reference one. It also mitigates the effect of an inaccurate threshold for PSSM search. By choosing a relaxed threshold, this method relies on well established motif discovery algorithms to identify a conserved motif in the target species and disregards the regions with sites that match the reference pattern but may not align well with the true target motif. To prevent MEME from discovering motifs for other promoter elements, we replace surrounding regions of putative sites with 100 bp sequences randomly generated from genomic background, instead of using the promoters of high-scoring sites directly for motif discovery.

2.4 Network Transfer

In closely related bacteria, it has been shown that orthologous TFs tend to have conserved binding motifs [9]. Although this tendency has been observed among more distant bacteria for some TFs (e.g., ArgR/AhrC and HrcA regulating arginine metabolism [41, 42]), it does not hold for some other TFs such as the SOS response repressor LexA in Gram-negative bacteria [43] and DinR, its ortholog in Gram-positive bacteria [44]. An important limitation of the motif-based transfer methods described above is that they are expected to perform poorly if the motif is not conserved across reference and target genomes. The underlying hypothesis for network transfer is that the regulon across the reference and target genomes may be functionally conserved to some degree even if the binding motif is not.

To define the motif in target species through network transfer, the first step is to identify the set of operons that are regulated by the TF of interest. To identify target regulon, genes that are orthologous to the ones in the reference regulon are identified in the target species and their promoters, typically shared with other genes in an operon configuration, are determined. In the next step, these promoters are used for motif discovery. If genes of an operon in the reference genome are dispersed into multiple operons in the target genome, all promoters of such operons are included. The hypothesis is that this method makes motif identification possible even if the motif is not conserved at all, assuming the regulon is conserved to some extent.

2.5 Performance Assessment

To assess the performance of different transfer techniques quantitatively, we measure both (a) the distance between the true motif and the inferred motif and (b) the area under ROC curve for the inferred motif. To measure the distance, two motifs are aligned maximizing the alignment's information content. The motif distance is then computed as the sum of Euclidean distances between aligned columns of the two position specific frequency matrices (PSFMs) [45, 46]. ROC curves [36, 47–50] are computed considering experimentally validated sites in target genomes as positives and all other positions in promoter regions as negatives. To handle the problem of class imbalance on binding site prediction, we selected same number of promoters with and without true binding sites to compute ROC curves. To assess the significance of performances for all three methods, we compute the distance and area under ROC curve using a column-permuted version of the target motif as the "transferred motif".

2.6 Software

Python scripts for compilation of the binding site data are available for download at `http://github.com/sefakilic/TFBS_data`. For genome-wide PSSM search and parsing of RefSeq genome records, the Biopython library was used [51]. For visualization of data and results, the matplotlib [52] and ggplot2 [53] libraries were used. All Python and R scripts developed in this work are available for download at `http://github.com/sefakilic/cg`.

3 Results and Discussion

We measured the performance of the transfer methods used in the literature by applying them to all pairs of species with at least 10 binding sites for a particular TF, yielding 411 pairs of species for motif and network transfer. Most of such pairs belong to either Fur or LexA (Fur: 154, LexA: 134, CcpA: 20, PhoP: 12, CodY: 12, OmpR: 6, CRP: 4, RpoN: 4, FNR: 2, PurR: 2, DtxR: 2, ArgR: 2, PvdS: 2, CsgD: 2 species pairs). We set the scaling factor $\alpha = 1.15$ for the direct transfer and $\alpha = 2.3$ for the PSSM search with motif discovery to achieve high sensitivity. For the motif discovery, we used MEME [40] with the following command line settings `-zoops -revcomp -dna nmotifs 5`. The minimum and maximum motif widths were set as 50% and 150% of the reference motif width, respectively.

The transfer methods described above are evaluated and their performances are reported as a function of TF protein distance. The protein distance between TFs in two species is defined as the percentage of residues that are not identical in the pairwise alignment.

Figures 1 and 2 show the relative performance of transfer methods using the Euclidean distance (normalized by motif alignment length) and the area under ROC curve of the inferred motif, respectively. As it can be observed in both figures, direct transfer and motif discovery on pre-searched promoters perform very similarly. As expected, these two methods perform well when the motif is conserved across the reference and target species. However, as the protein distance increases, the average performance of these two methods, which rely on the assumption of motif conservation, decreases dramatically. As the reference-target protein distance increases further, these methods do not perform significantly better than the permuted version of the target motif.

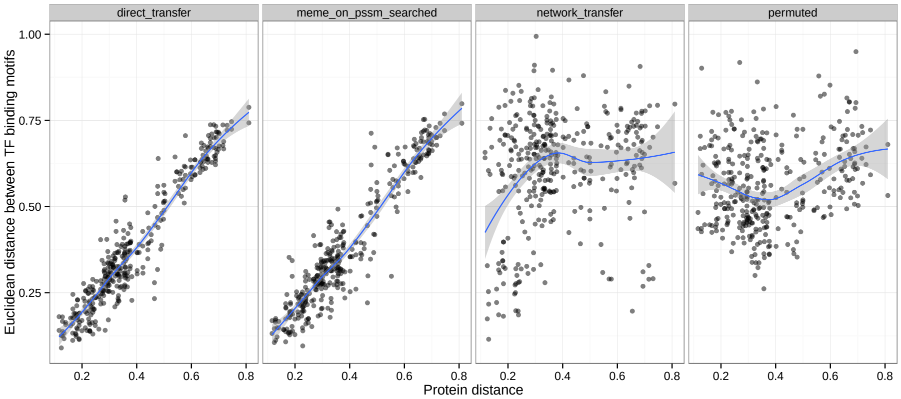

Fig. 1. The Euclidean distance between the true target motif and inferred motif.

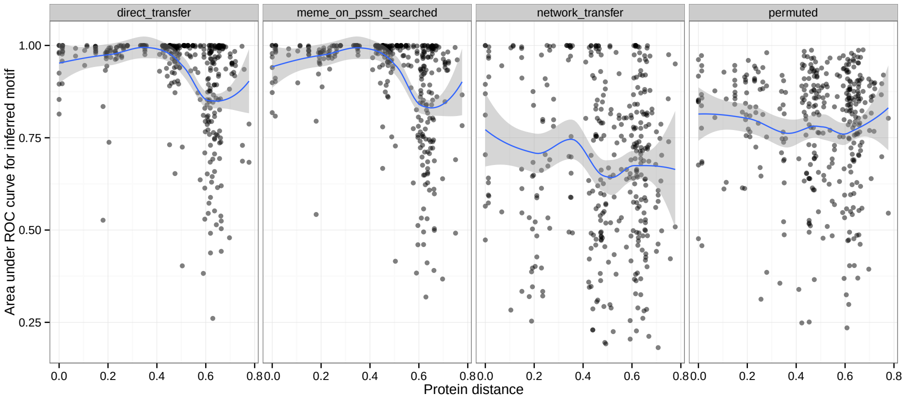

Fig. 2. Area under ROC curves for motifs inferred via different transfer methods.

Although network transfer is often capable of inferring non-conserved motifs for large protein distances, the permutation analysis suggests that these data points are not statistically significant. In fact, it can be concluded that the network transfer method does not perform well for any level of reference-target distance overall. Our analysis demonstrating the poor performance of this method is consistent with previous studies reporting high plasticity in transcriptional regulatory networks and therefore weakening the assumption of functional conservation made in network transfer [54–56]. In this context, our further analysis revealed that the poor performance is due to stringent nature of the reciprocal BLAST for ortholog detection. This results in too few promoters with true sites for MEME to be able to detect the signal. Figure 3 shows the precision ($TP/(TP+FP)$) and recall ($TP/(TP+FN)$) rates for each transferred regulon *before* the motif discovery step. Here, a promoter is considered as a true positive (TP) if it contains a true site from the target motif and selected as a target promoter to be searched by MEME. False positives (FP) are promoters with no true sites but in the collection MEME searches and false negatives (FN) are promoters with sites but not in the collection that MEME searches for a motif. Low precision and recall rates suggest that, in most cases, most of the operons in the inferred regulon are not members of the true target regulon. As a result, MEME is not able to recover the true binding motif.

4 Conclusion

In this paper, we report the first benchmarking of methods for transfer of regulatory information across bacterial genomes. We performed experiments using TF binding motif data compiled from CollecTF and other publicly available databases. Our analysis suggests that the traditional approach of direct transfer via PSSM search performs best, especially when the reference and target species

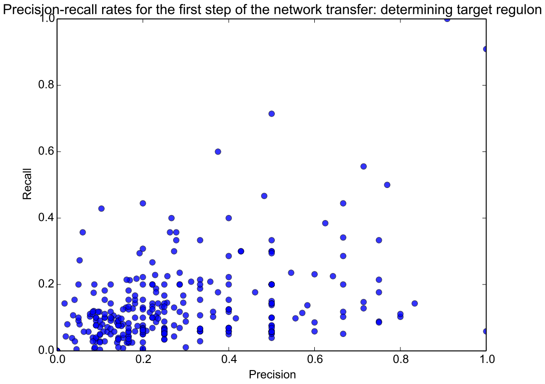

Fig. 3. Precision and recall curves for the first step (i.e., regulon transfer) of the network transfer.

are closely related. We also tested a network transfer approach which is based on the assumption of regulatory network conservation. In accordance with the recent studies suggesting extensive rewiring of regulatory networks, we found that the network transfer approach does not perform well because of small overlap between networks and a large amount of noise in the transfer process that overcomes the power of motif discovery method.

Our further analysis indicates that another reason for poor network transfer is the strictness of the reciprocal BLAST-based regulon transfer method. One direction for future work is to modify the network transfer method to use functional similarity (e.g., clusters of orthologous groups, COGs [57]) for regulon transfer rather than direct orthology. With this approach, instead of considering genes that are orthologous to those in the reference network, target genes that have same or similar function to reference regulon are also considered for motif discovery. A second future direction is to investigate whether combining the information from the extended network transfer with relaxed PSSM searches can enhance the performance of direct transfer as the similarity between reference and target motifs decays.

Bibliography

[1] MS Gelfand, EV Koonin, and AA Mironov. Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Research*, 28(3):695–705, 2000.

[2] D. A. Ravcheev, A. A. Best, N. V. Sernova, M. D. Kazanov, P. S. Novichkov, and D. A. Rodionov. Genomic reconstruction of transcriptional regulatory networks in lactic acid bacteria. *BMC Genomics*, 14:94, 2013.

[3] Antonio CA Meireles-Filho and Alexander Stark. Comparative genomics of gene regulation—conservation and divergence of cis-regulatory information. *Current opinion in genetics & development*, 19(6):565–570, 2009.

[4] Manolis Kellis, Nick Patterson, Bruce Birren, Bonnie Berger, and Eric S Lander. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology*, 11(2-3):319–355, 2004.

[5] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS computational biology*, 9(11):e1003326, 2013.

[6] D. A. Rodionov. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.*, 107(8):3467–3497, Aug 2007.

[7] K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides, and G. D. Stormo. A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, 11(4):566–584, Apr 2001.

[8] A. A. Mironov, E. V. Koonin, M. A. Roytberg, and M. S. Gelfand. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, 27(14):2981–2989, Jul 1999.

[9] K. S. Makarova, A. A. Mironov, and M. S. Gelfand. Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol.*, 2(4):RESEARCH0013, 2001.

[10] D. A. Rodionov, A. A. Mironov, A. B. Rakhmaninova, and M. S. Gelfand. Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. *Mol. Microbiol.*, 38(4):673–683, Nov 2000.

[11] E. M. Panina, A. A. Mironov, and M. S. Gelfand. Comparative analysis of FUR regulons in gamma-proteobacteria. *Nucleic Acids Res.*, 29(24):5195–5206, Dec 2001.

[12] S. A. Leyn, X. Li, Q. Zheng, P. S. Novichkov, S. Reed, M. F. Romine, J. K. Fredrickson, C. Yang, A. L. Osterman, and D. A. Rodionov. Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in Shewanella oneidensis. *J. Biol. Chem.*, 286(41):35782–35794, Oct 2011.

[13] D. A. Rodionov, I. Dubchak, A. Arkin, E. Alm, and M. S. Gelfand. Reconstruction of regulatory and metabolic pathways in metal-reducing deltaproteobacteria. *Genome Biol.*, 5(11):R90, 2004.

[14] A. E. Kazakov, D. A. Rodionov, E. Alm, A. P. Arkin, I. Dubchak, and M. S. Gelfand. Comparative genomics of regulation of fatty acid and branched-chain amino acid utilization in proteobacteria. *J. Bacteriol.*, 191(1):52–64, Jan 2009.

[15] D. A. Rodionov, X. Li, I. A. Rodionova, C. Yang, L. Sorci, E. Dervyn, D. Martynowski, H. Zhang, M. S. Gelfand, and A. L. Osterman. Transcriptional regulation of NAD metabolism in bacteria: genomic reconstruction of NiaR (YrxA) regulon. *Nucleic Acids Res.*, 36(6):2032–2046, Apr 2008.

[16] I. Erill, M. Jara, N. Salvador, M. Escribano, S. Campoy, and J. Barbe. Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. *Nucleic Acids Res.*, 32(22):6617–6626, 2004.

[17] D. A. Shelton, L. Stegman, R. Hardison, W. Miller, J. H. Bock, J. L. Slightom, M. Goodman, and D. L. Gumucio. Phylogenetic footprinting of hypersensitive site 3 of the beta-globin locus control region. *Blood*, 89(9):3457–3469, May 1997.

[18] E. M. Panina, A. A. Mironov, and M. S. Gelfand. Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 100(17):9912–9917, Aug 2003.

[19] L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, 29(3):774–782, Feb 2001.

[20] T. Wang and G. D. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380, Dec 2003.

[21] S. A. Leyn, M. D. Kazanov, N. V. Sernova, E. O. Ermakova, P. S. Novichkov, and D. A. Rodionov. Genomic reconstruction of the transcriptional regulatory network in Bacillus subtilis. *J. Bacteriol.*, 195(11):2463–2473, Jun 2013.

[22] S. Zhang, M. Xu, S. Li, and Z. Su. Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes. *Nucleic Acids Res.*, 37(10):e72, Jun 2009.

[23] Shaoqiang Zhang, Shan Li, Phuc Pham, and Zhengchang Su. Simultaneous prediction of transcription factor binding sites in a group of prokaryotic genomes. *BMC Bioinformatics*, 11(1):397, 2010.

[24] Sefa Kılıç, Elliot R. White, Dinara M. Sagitova, Joseph P. Cornish, and Ivan Erill. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. *Nucleic Acids Research*, 2013.

[25] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair S. García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, Gerardo

Salgado-Osorio, Shirley Alquicira-Hernández, Kevin Alquicira-Hernández, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Araceli M. Huerta, César Bonavides-Martínez, Yalbi I. Balderas-Martínez, Lucia Pannier, Maricela Olvera, Aurora Labastida, Verónica Jiménez-Jacinto, Leticia Vega-Alvarado, Victor del Moral-Chávez, Alfredo Hernández-Alvarez, Enrique Morett, and Julio Collado-Vides. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(D1):D203–D213, 2013.

[26] Josch Pauling, Richard Röttger, Andreas Tauch, Vasco Azevedo, and Jan Baumbach. CoryneRegNet 6.0—updated database content, new analysis methods and novel features focusing on community demands. *Nucleic Acids Research*, 40(D1):D610–D614, 2012.

[27] Nicolas Sierro, Yuko Makita, Michiel de Hoon, and Kenta Nakai. DBTBS: a database of transcriptional regulation in bacillus subtilis containing upstream intergenic conservation information. *Nucleic acids research*, 36(suppl 1):D93–D96, 2008.

[28] Pierre-Étienne Jacques, Alain L. Gervais, Mathieu Cantin, Jean-Franc¸ois Lucier, Guillaume Dallaire, Geneviève Drouin, Luc Gaudreau, Jean Goulet, and Ryszard Brzezinski. MtbRegList, a database dedicated to the analysis of transcriptional regulation in mycobacterium tuberculosis. *Bioinformatics*, 21(10):2563–2565, 2005.

[29] F. Mao, P. Dam, J. Chou, V. Olman, and Y. Xu. DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, 37(Database issue):D459–463, Jan 2009.

[30] DP Wall, HB Fraser, and AE Hirsh. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711, 2003.

[31] D. A. Rodionov, P. S. Novichkov, E. D. Stavrovskaya, I. A. Rodionova, X. Li, M. D. Kazanov, D. A. Ravcheev, A. V. Gerasimova, A. E. Kazakov, G. Y. Kovaleva, E. A. Permina, O. N. Laikova, R. Overbeek, M. F. Romine, J. K. Fredrickson, A. P. Arkin, I. Dubchak, A. L. Osterman, and M. S. Gelfand. Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the Shewanella genus. *BMC Genomics*, 12 Suppl 1:S3, 2011.

[32] Gary D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[33] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7:389, 2006.

[34] Rodger Staden. Methods for calculating the probabilities of finding patterns in sequences. *Computer applications in the biosciences: CABIOS*, 5(2):89–96, 1989.

[35] Sven Rahmann, Tobias Müller, and Martin Vingron. On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology*, 2(1), 2003.

[36] J. P. Cornish, F. Matthews, J. R. Thomas, and I. Erill. Inference of self-regulated transcriptional networks by comparative genomics. *Evol. Bioinform. Online*, 8:449–461, 2012.

[37] Alexey E Kazakov, Dmitry A Rodionov, Morgan N Price, Adam P Arkin, Inna Dubchak, and Pavel S Novichkov. Transcription factor family-based reconstruction of singleton regulons and study of the crp/fnr, arsr, and gntr families in desulfovibrionales genomes. *Journal of bacteriology*, 195(1):29–38, 2013.

[38] Chih Lee, Chun-Hsi Huang, et al. Lasagna-search: an integrated web tool for transcription factor binding site search and visualization. *Biotechniques*, 54(3):141–153, 2013.

[39] N. Habib, I. Wapinski, H. Margalit, A. Regev, and N. Friedman. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. *Mol. Syst. Biol.*, 8:619, 2012.

[40] Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(suppl 2):W369–W373, 2006.

[41] W. K. Maas. The arginine repressor of Escherichia coli. *Microbiol. Rev.*, 58(4):631–640, Dec 1994.

[42] U. Klingel, C. M. Miller, A. K. North, P. G. Stockley, and S. Baumberg. A binding site for activation by the Bacillus subtilis AhrC protein, a repressor/activator of arginine metabolism. *Mol. Gen. Genet.*, 248(3):329–340, Aug 1995.

[43] Joseph T Wade, Nikos B Reppas, George M Church, and Kevin Struhl. Genomic analysis of lexa binding reveals the permissive nature of the escherichia coli genome and identifies unconventional target sites. *Genes & development*, 19(21):2619–2630, 2005.

[44] K. W. Winterling, D. Chafin, J. J. Hayes, J. Sun, A. S. Levine, R. E. Yasbin, and R. Woodgate. The Bacillus subtilis DinR binding site: redefinition of the consensus sequence. *J. Bacteriol.*, 180(8):2201–2211, Apr 1998.

[45] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biol*, 8(2):R24, 2007.

[46] In-Geol Choi, Jaimyoung Kwon, and Sung-Hou Kim. Local feature frequency profile: a method to measure structural similarity in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3797–3802, 2004.

[47] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[48] Anthony Mathelier and Wyeth W Wasserman. The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9):e1003214, 2013.

[49] Mohammad Talebzadeh and Fatemeh Zare-Mirakabad. Transcription factor binding sites prediction based on modified nucleosomes. *PloS one*, 9(2):e89226, 2014.

[50] Kyoung-Jae Won, Bing Ren, and Wei Wang. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, 11(1):R7, 2010.

[51] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[52] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[53] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[54] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput Biol*, 3(9):e175, 09 2007.

[55] M. Madan Babu, Sarah A. Teichmann, and L. Aravind. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of Molecular Biology*, 358(2):614 – 633, 2006.

[56] Irma Lozada-Chávez, Sarath Chandra Janga, and Julio Collado-Vides. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Research*, 34(12):3434–3445, 2006.

[57] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.