CircularLogo: A light weighted web application to visualize intra-motif dependencies

Zhenqing Ye[1], Tao Ma[2], Michael T. Kalmbach[1], Jean-Pierre A. Kocher[1], Liguo Wang[1,2,*]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences, Mayo Clinic, Rochester, Minnesota, United States of America

[2]Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, Minnesota, United States of America

* To whom correspondence should be addressed:

Phone: 1-507-284-8728
Fax: 1-507-284-0745
Email: Wang.Liguo@mayo.edu

Abstract

Background: The sequence logo has been widely used to represent DNA or RNA motif for more than three decades. Despite its intelligibility and intuitiveness, the sequence logo is unable to display and reveal the intra-motif dependencies and therefore is insufficient to fully characterize nucleotide motifs.

Result: We developed *CircularLogo*, a web-based interactive application that is able to not only visualize the position-specific nucleotide consensus and diversity but also unveil the intra-motif dependencies. When applying it to HNF6 binding sites and tRNA sequences, we demonstrated *CircularLogo* could display intra-motif dependencies and reveal biomolecular structure effectively. *CircularLogo* is implemented in Javascript and Python based on the Django web framework. Source code and a comprehensive user's manual are freely available at http://circularlogo.sourceforge.net. The web server can be accessed from http://bioinformaticstools.mayo.edu/circularlogo/index.html.

Conclusion: *CircularLogo* is an innovative web application that specifically designed to visualize and explore intra-motif dependencies.

Keywords:
CircularLogo, intra-motif dependency, visualization

Background

Most DNA and RNA binding proteins recognize their binding sites through specific nucleotide patterns called motifs. Motif sites bound by the same protein are not necessary the same but share a consensus sequence. Several methods have been developed to statistically model the position-specific consensus and diversity of nucleotide motifs using the position weight matrix (PWM) or position-specific scoring matrix (PSSM) [1,2]. These mathematical representations are usually visualized using sequence logos, which depict the consensus and diversity of motif residue as a stack of symbols. The height of each symbol is often proportional to its relative frequency, information content or probability [3,4].

Traditional PWM and PSSM assume statistical independences between nucleotides within the motif. However, such assumption was not completely justified, and accumulated evidence indicated the existence of intra-motif dependencies [5-8]. For examples, the analysis of wild-type and mutant *Zif268* (*Egr1*) zinc fingers using microarray binding experiments indicated that the nucleotides within transcription factor binding sites (TFBS) should not be treated independently [5]. In addition, the positional interdependence within motif was also revealed by a comprehensive experiment to examine the binding specificities of 104 distinct DNA binding proteins in mouse [8]. It has been found that taking into consideration the intra-motif dependencies will substantially improve the accuracy of *de novo* motif discovery [9]. Therefore, many new statistical methods have been developed to characterize the intra-motif dependencies, which include the generalized weight matrix model [10], sparse local inhomogeneous mixture model (Slim) [11], transcription factor flexible model based on hidden Markov models (TFFMs) [12], the binding energy model (BEM) [13], and the inhomogeneous parsimonious Markov model (PMM) [14]. However, the most commonly used visualization tools such as WebLogo [3], Seq2Logo [15], and pLogo [4] can only display individual symbol stacks and ignore the intra-dependences within the motif.

Comparing to accomplishments in developing new statistical models, much fewer efforts have been spent on developing tools to visualize the intra-motif dependencies. Currently,

tools that are capable of visualizing positional dependencies include CorreLogo and ELRM [16,17]. CorreLogo uses three-dimensional sequence logos to depict mutual information from DNA or RNA alignment via VRML and JVX output. However, three-dimensional graphs generated from CorreLogo are difficult to interpret because of the excessively complex and the distorted effect of perspective associated with the third dimension. ELRM provides Perl scripts for generating static graphs to visualize intra-motif dependences. ELRM splits up "base features" and "association features" and fails to integrate nucleotide diversities and dependencies as an entirety, in addition, ELRM is limited to its own built-in method for measuring dependence. More importantly, both CorreLogo and ELRM lack the functionality to enable users to explore and interpret the data interactively.

In this study, we developed *CircularLogo*, an interactive web application that transforms the JSON (JavaScript Object Notation) format motif representation into a circular sequence logo, which is able to display position-specific nucleotide frequencies as well as the intra-motif dependencies simultaneously. *CircularLogo* uses an open-standard, human-readable, and computer language-independent JSON data format, which is flexible to describe various properties of DNA motifs and easy to manipulate. Other commonly used motif representation formats such as MEME, TRANSFAC, and JASPAR can be easily converted into this format.

Implementation

JSON-Graph specifications of DNA motif representation

We used the JSON-Graph format to describe DNA motif with the purpose to make it intelligible and easy to manipulate. The schema of JSON-Graph format is illustrated below:


```
{
"id": "Toy motif",
"background":{"key":["A","T","C","G"],"val":[0.25,0.25,0.25,0.25]},
"pseudocounts":{"key":["A","T","C","G"],"val":[0.25,0.25,0.25,0.25]},
"nodes": [
        {"index": 0, "label": "1", "bit": 0.78, "base": ["A", "T", "C", "G"], "freq": [0.006, 0.043, 0.347, 0.604]},
        {"index": 1, "label": "2", "bit": 1.57, "base": ["T", "C", "A", "G"], "freq": [0.012, 0.017, 0.032, 0.939]},
```

{"index": 2, "label": "3", "bit": 0.61, "base": ["C", "G", "A", "T"], "freq": [0.027, 0.053, 0.388, 0.532]},
.....
],
"links": [
{"source": 0, "target": 1, "value": 2.0},
{"source": 2, "target": 4, "value": 8.0},
{"source": 2, "target": 7, "value": 6.0},
.....
]
}

The above example represents the basic data structure of JSON-Graph format file. The content within two curly braces described a DNA or RNA motif. Specifically, the "*id*" keyword specifies the name of the motif. "*background*" keyword designated nucleotides frequencies (in the order of A,C,G,T) of the genomic background. "*pseudocounts*" represent the pseudocounts of A,T,C,G that added to each position of the motif. The "*nodes*" section describes properties of motif residues, within which the "*index*" keyword specifies the order (in anticlockwise) of nucleotides in a motif, the "*label*" keyword denotes the identity of stacked nucleotides, the "*bit*" keyword is the information content calculated from nucleotide frequencies at this position, and the "*base*" keyword indicates the four nucleotide alphabets sorted incrementally by their corresponding frequencies as designated by the "*freq*" keyword. The "*links*" section describes the pairwise dependencies between nucleotide positions with the "*source*" and "*target*" keywords denoting the start and the end positions of the link, and the "value" indicates the width (the strength of dependence) of the link.

***CircularLogo* web server**

CircularLogo is a Python-Django web application, which uses NGINX (https://www.nginx.com/) as the web server and uWSGI (https://pypi.python.org/pypi/uWSGI) as the web server gateway interface for multiple concurrent client requests. It is hosted on Amazon Elastic Compute Cloud (Amazon EC2).

Measure intra-motif dependencies using χ^2 statistic

With FASTA sequences as input, we implemented two methods to calculate dependence between two nucleotide positions: mutual information and χ^2 statistic. The χ^2 test is widely used to test the independence of two categorical variables, and the χ^2 statistic score (Q) is the natural measurement of dependency of two events by quantifying the co-incidence. Assuming a DNA motif is l nucleotides long and is built from N sequences, for a given two positions i and j within the motif ($1 \leq i \leq l$, $1 \leq j \leq l$, $i \neq j$), the observed di-nucleotide frequencies are denoted as O_{ij}, and the expected di-nucleotide frequencies are represented as E_{ij}. E_{ij} can be specified in prior or calculated by shuffling the N sequences. The χ^2 statistic score is calculated as:

$$Q = \sum_{k=1}^{m} \frac{\left(O_{ij}^{k} - E_{ij}^{k}\right)^2}{E_{ij}^{k}}, Q \sim x^2\,(m-1), m = 16,\ O_{ij} \in [AA, AT, AC, AG, \dots]$$

Measure intra-motif dependencies using mutual information

The other built-in approach to measure dependence is the mutual information (*MI*), which is a widely used metric to measure the mutual dependence between two variables. The mutual information of two discrete random variables X and Y can is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

In our situation, x and y represent nucleotide base at position i and j, respectively. $p(x)$ and $p(y)$ denote the frequencies of x and y at position i and j, respectively. $p(x, y)$ defines the frequency of dinucleotide xy.

HNF6 motif analysis

The ChIP-exo data of HNF6 was published by our group previously and was deposited in Array Express with accession number E-MTAB-2060 (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2060/). ChIP-exo data was analyzed using MACE[18]. 5549 border pairs that were 25 nucleotides long were extracted from MACE output because 25mer is the most abundant out of all identified

border pairs[18]. For each peak border, we extended 20 nucleotides to both upstream and downstream with the aim to estimate background dependencies. The 5549 65-nucleotide sequences were available from https://sourceforge.net/projects/circularlogo/files/test/.

tRNA sequence analysis

The sequences of tRNA were downloaded from RFAM database [22]. In particular, 1114 sequences of RFAM 'seed' alignment RF00005 were downloaded from https://correlogo.ncifcrf.gov/ccrnp/trnafull.html. After excluding those sequences with gaps in the alignment, 291 sequences were used as the final dataset to generate circular logo of tRNA (https://sourceforge.net/projects/circularlogo/files/test/). Mutual information was used as the metric to measure intra-motif dependencies. Links with scores less than 20 were filtered out.

Results

Circular nucleotide motif

Unlike the traditional sequence logo that displays motif residues on the two-dimensional Cartesian coordinate system (with the horizontal x-axis denoting the position of residues and the vertical y-axis denoting the information contents, frequencies, or probabilities of residues), *CircularLogo* visualizes motifs using the polar coordinate system, which facilitates the display of pairwise intra-motif dependencies with linked ribbons. Since traditional PWM or PSSM motif representations do not preserve intra-motif dependency, we proposed to use the JSON-Graph as the main input format to *CircularLogo*. When the input file is in JSON-Graph format, all nucleotide frequencies and dependencies are pre-calculated, and *CircularLogo* simply transforms this file into the graphical representation. *CircularLogo* also accepts the FASTA format motif representation as input. In this scenario, *CircularLogo* will transform these FASTA sequences into a JSON-Graph representation by calculating the intra-motif dependency with the built-in χ^2 statistic or mutual information metric, and determined the height of the stacked nucleotides in the same way as webLogo does [3].

With *CircularLogo,* users can interactively adjust a variety of parameters to explore intra-motif dependencies and fine-tune the appearance of the final output. For example, any nucleotide in the genome has a certain level of dependencies with its immediate neighbors. We considered such dependencies as the background noise since they are not likely to be biologically meaningful. Due to genome heterogeneities, the different set of DNA motifs might have different levels of background dependency. Therefore, instead of using a fixed, hard threshold, *CircularLogo* provided users a slider bar to interactively filter out weak background links.

Nucleotide dependencies within HNF6 motif

HNF6 (also known as ONECUT1) is a transcription factor regulating gene expression in a variety of cellular processes. The protein-DNA binding boundaries of HNF6 in mouse genome were previously defined by our group using ChIP-exo[18]. 5549 binding sites that were 25 nucleotides long were used to explore the intra-motif dependencies. For each binding site, we also extended 20 nucleotides to up- and downstream with the aim to estimate background dependency level. Pair-wise dependencies between all 65 positions were displayed in **Fig. 1A**. As we expected, dependencies between nucleotides within the core HNF6 motif (i.e. nucleotides within 29^{th} and 36^{th} position) were much higher than those of flanking regions (**Fig. 1B**). **Fig. 1C** indicated background links relating to node 5 (i.e. the 5^{th} position of input DNA sequence). **Fig. 1D** indicated dependencies related to node 33 within the HNF6 core motif after spurious links were removed.

Nucleotide dependencies within tRNAs

The transfer RNA (tRNA) is usually 76 to 90 nucleotides long and is involved in the translation of message RNA into the amino acid sequence. Its typical cloverleaf secondary structure is composed of D-loop, anticodon loop, variable loop and TψC loop, as well as four base-paired stems between these loops (**Fig. 2A**). The nucleotide bases within these stems are not necessarily conserved, but base-pairing is required for structural stability. Thus we expect higher positional dependence between nucleotides within stems than those within loops. To visualize the nucleotide dependencies of tRNA, we used *CircularLogo* with the mutual information as a measurement of dependence.

After filtering out weak links, we observed four apparent clusters of connected links corresponding to the four stems (**Fig. 2B**). More interestingly, the nucleotides with three loops (D-loop, Anticodon loop, and TψC loop) exhibited much higher sequence conservation than that of nucleotides located in stems, suggesting the loops are the functional domains of tRNA. For example, D-loop is the recognition site of aminoacyl-tRNA synthetase, an enzyme involved in aminoacylation of the tRNA molecule [19,20], and TψC loop is the recognition site of the ribosome.

Discussion

Measuring intra-motif dependencies is not a trivial task; new statistical models and experimental approaches are still under active development. To not limit *CircularLogo* to several pre-defined approaches, we proposed to use the plain text, JSON-Graph format file to describing DNA/RNA motifs. So that it is straightforward to make a customized JSON-Graph file with positional dependencies pre-calculated by user-preferred methods.

When the input data to *CircularLogo* is in FASTA format, we provided two approaches (χ^2 statistic score and mutual information) to measure the positional dependencies. Although commonly used, both methods are biased or even unable to quantify dependencies between sites that are highly conserved (e.g. invariable sites) [6,21]. To address this problem, users need to provide as many sequences as possible to capture the low-abundance variants at those highly conserved sites. This is practically not difficult thanks to many genome-wide, high-throughput technologies. For example, people usually identified tens of thousands of potential TFBS using ChIP-seq or other similar technologies.

Conclusions

Visualization is critical for efficient data exploration and effective communication in scientific research. *CircularLogo* is an innovative tool offering the panorama of DNA or RNA motifs taking into consideration the intra-site dependencies. We demonstrated the utility and practicality of this tool with examples that *CircularLogo* is able to depict

complex dependencies within motifs and reveal biomolecular structure (such as stem structures in tRNA) effectively.

List of abbreviations

BEM, the Binding Energy Model
JSON, JavaScript Object Notation
MEME, Multiple Em for Motif Elicitation
MACE, Model-based Analysis of ChIP-Exo
MI, Mutual Information
PSSM, Position-Specific Scoring Matrix
PMM, the inhomogeneous Parsimonious Markov Model
PWM, Position Weight Matrix
TFBS, Transcription Factor Binding Sites
TFFMs, Transcription Factor Flexible Model

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The datasets analyzed during the current study are available at:
https://sourceforge.net/projects/circularlogo/files/test/repository

Competing interests

The authors declare that they have no competing interests

Funding

This works is partly supported by the Mayo Clinic Center for Individualized Medicine.

Authors' contributions

LW and JPK conceived the study. ZY and TM implemented *CircularLogo* software and performed the analysis. MK built *CircularLogo* web server. LW, ZY and JPK wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

Reference

1. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16:16–23.

2. Boeva V. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Front Genet. 2016;7:24.

3. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004;14:1188–90.

4. O'Shea JP, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. Nat. Methods. 2013.

5. Bulyk ML, Johnson PLF, Church GM. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. 2002;30:1255–61.

6. Eggeling R, Gohr A, Keilwagen J, Mohr M, Posch S, Smith AD, et al. On the value of intra-motif dependencies of human insulator protein CTCF. PLoS ONE. 2014;9:e85629.

7. Man TK, Stormo GD. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res. 2001;29:2471–8.

8. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009;324:1720–3.

9. Grau J, Posch S, Grosse I, Keilwagen J. A general approach for discriminative de novo motif discovery from high-throughput data. Nucleic Acids Res. 2013;41:e197.

10. Zhou Q, Liu JS. Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics. 2004;20:909–16.

11. Keilwagen J, Grau J. Varying levels of complexity in transcription factor binding

motifs. Nucleic Acids Res. 2015;43:e119.

12. Mathelier A, Wasserman WW. The Next Generation of Transcription Factor Binding Site Prediction. PLoS Comput. Biol. Public Library of Science; 2013;9:e1003214.

13. Zhao Y, Ruan S, Pandey M, Stormo GD. Improved models for transcription factor binding site identification using nonindependent interactions. Genetics. 2012;191:781–90.

14. Eggeling R, Roos T, Myllymäki P, Grosse I. Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. BMC bioinformatics. 2015;16:375.

15. Thomsen MCF, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 2012;40:W281–7.

16. Bindewald E, Schneider TD, Shapiro BA. CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. Nucleic Acids Res. 2006;34:W405–11.

17. Yang C, Chang C-H. Exploring comprehensive within-motif dependence of transcription factor binding in Escherichia coli. Sci Rep. 2015;5:17021.

18. Wang L, Chen J, Wang C, Uusküla-Reimand L, Chen K, Medina-Rivera A, et al. MACE: model based analysis of ChIP-exo. Nucleic Acids Res. 2014.

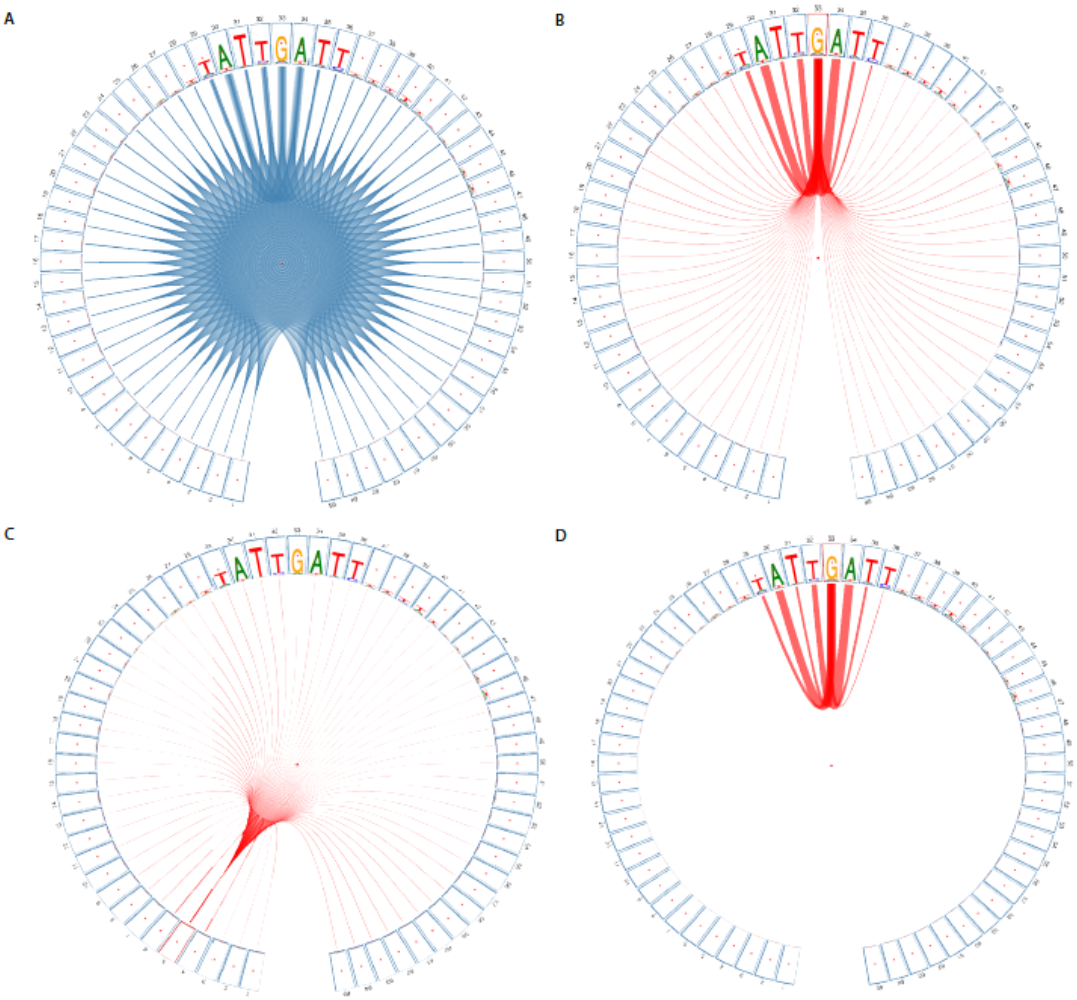
19. Smith D, Yarus M. Transfer RNA structure and coding specificity. I. Evidence that a D-arm mutation reduces tRNA dissociation from the ribosome. Journal of molecular biology. 1989;206:489–501.

20. Hardt WD, Schlegl J, Erdmann VA, Hartmann RK. Role of the D arm and the anticodon arm in tRNA recognition by eubacterial and eukaryotic RNase P enzymes. Biochemistry. 1993;32:13046–53.

21. Paninski L. Estimation of entropy and mutual information. Neural computation. 2003.

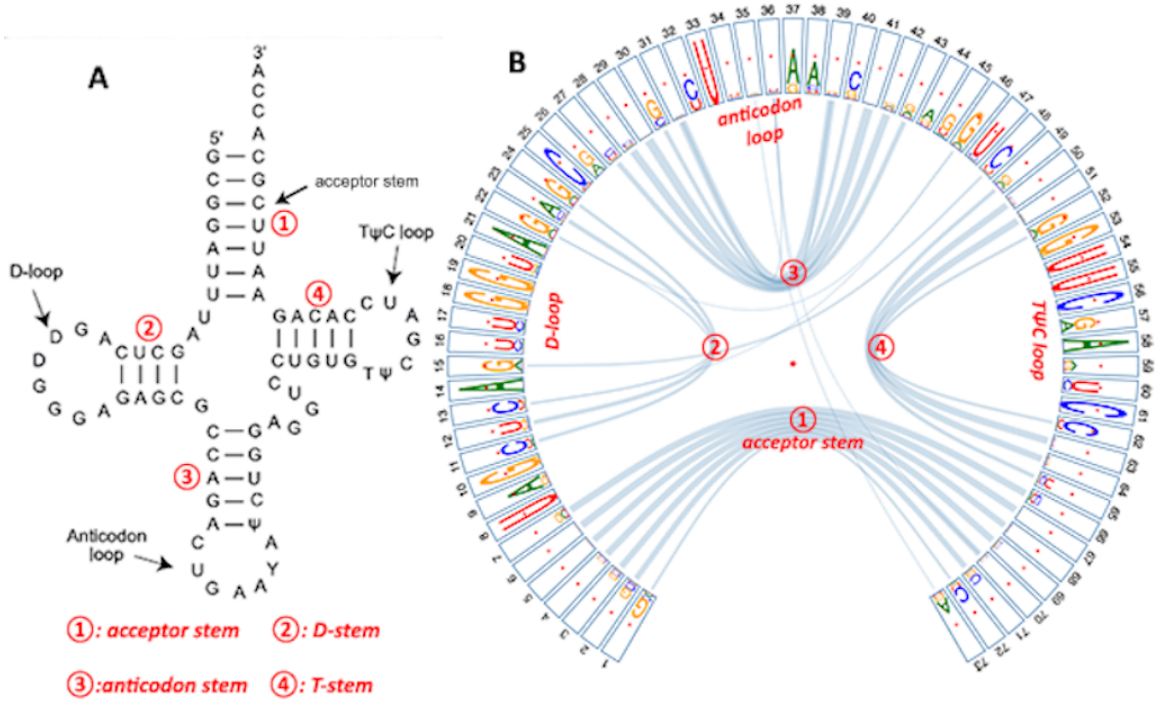
22. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003;31:439–41.

Figure 1



(A) Motif generated from *CircularLogo* describing the pairwise dependencies between 65 nucleotides (20 upstream nucleotides + 25 HNF6 binding sites defined from ChIP-exo data + 20 downstream nucleotides). (B) All links related to node 33. (C) All links related on node 5, representing background level dependencies. (D) Links related to node 33 after removing spurious, background links.

Figure 2



(A) The typical cloverleaf secondary structure of Phe-tRNA in yeast. (B) tRNA motif represented with the circular motif logo, the width of links indicates the strength of dependency (measured by mutual information). The labels ①, ②, ③, ④ indicate acceptor stem, D-stem, anticodon stem, and T-stem, respectively.