Published in final edited form as:

Methods Mol Biol. 2012 ; 802: 323–334. doi:10.1007/978-1-61779-400-1_21.

Using ChIPMotifs for De Novo Motif Discovery of OCT4 and ZNF263 Based on ChIP-Based High-Throughput Experiments

Brian A. Kennedy,

Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

Xun Lan,

Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

Tim H.-M. Huang,

Department of Molecular Virology, Immunology & Medical Genetics, The Ohio State University, Columbus, OH, USA

Peggy J. Farnham, and

Department of Biochemistry & Molecular Biology, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA

Victor X. Jin

Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA

Abstract

DNA motifs are short sequences varying from 6 to 25 bp and can be highly variable and degenerated. One major approach for predicting transcription factor (TF) binding is using position weight matrix (PWM) to represent information content of regulatory sites; however, when used as the sole means of identifying binding sites suffers from the limited amount of training data available and a high rate of false-positive predictions. ChIPMotifs program is a de novo motif finding tool developed for ChIP-based high-throughput data, and W-ChIPMotifs is a Web application tool for ChIPMotifs. It composes various ab initio motif discovery tools such as MEME, MaMF, Weeder and optimizes the significance of the detected motifs by using bootstrap re-sampling error estimation and a Fisher test. Using these techniques, we determined a PWM for OCT4 which is similar to canonical OCT4 consensus sequence. In a separate study, we also use de novo motif discovery to suggest that ZNF263 binds to a 24-nt site that differs from the motif predicted by the zinc finger code in several positions.

Keywords

Motif; ChIP; Position weight matrix; OCT4; ZNF263

1. Introduction

During the past decade, several computational approaches have been developed to study large and complex datasets generated from high-throughput technologies such as mRNA expression profiling (1, 2), ChIP-chip (3, 4), DamID (5), DNase-chip (6), and ChIP-PET (7). The computational algorithms behind these approaches include (1) statistically driven ab initio motif discovery methods such as hidden Markov models (8), Gibbs sampling (9), expectation-maximization (MEME (10)), exhaustive enumeration (Weeder (11)), and words enumeration with a positional weight matrix updating (12); and (2) prior-compiled position weight matrices (PWMs) library-based motifs detection methods such as MATCH (13) combined with the TRANSFAC database (14) and MSCAN (15) combined with the JASPAR database (16).

All of the above-motivated methods have been proven to be useful in detecting novel motifs and deciphering the logics of transcription regulatory networks; however, there are still several major challenges facing these de novo methods. First, TF binding sites are short and easily confused among the noise of larger sequences; second, variability in TF binding sites is not well understood; and third, many consensus binding sites are derived from a small set of in vitro experiments. Some of these challenges in identifying motifs can be minimized by using ChIP-chip data to derive a consensus binding site to which a factor is bound in vivo. Also, some of the issues concerned with background (control) sequences can be eliminated using a bootstrap re-sampling of the data.

The sequences identified from ChIP-based high-throughput techniques such as ChIP-chip (4, 17), ChIP-seq (18, 19), and ChIP-PET (7) are called “peaks,” which are defined as significantly dense clusters in the sequence reads. Usually ranging from ~150 to ~1,500 bases, these peaks are currently considered to be highly reliable data sets for detecting the novel motif. Many computational tools including ours (20–24) have been recently developed to de novo find the motifs for the data generated from these techniques.

2. Methods

The flow chart in Fig. 1a demonstrated the general protocol used for de novo motif discovery. In which, sequences are ranked according to some metric external to this algorithm, and the top k sequences are selected for de novo motif detection. In the case of in vivo ChIP-based data, for which this protocol was originally developed, the criteria for selecting input sequences would be that the binding sites (sequences) were identified by a peak detection program and ranked based on a statistical measurement (a p -value or a false discovery rate). Binding sites (sequences) above an appropriate significance (such as $p < 0.05$) would be used as an input data in the following protocol.

2.1. General Protocol for De Novo Motif Discovery

1. Select the input data set of the top k sequences ordered by significance (see Note 1).
2. Process the input data set in Weeder (see Note 2).

3. Process the input data set in MEME.
4. Process the input data set in MaMF.
5. The union of the output of these three programs is the set of candidate motifs, of size i .
6. Construct position weight matrices for each of the i candidate motifs.
7. Perform bootstrap re-sampling by randomizing each of the k sequences for 100 times, and generate a total of $100 \times k$ sequences. These sequences have same nucleotides' identities with original sequences but in different orders (see Note 3).
8. Scan these randomized sequences for each candidate motif (using the PWMs derived from step 6) starting at a minimal core score of 0.5 and a minimal PWM score of 0.5. This score is the sum of the weight for the nucleotide in the sequence being scored at a position i in the PWM, for each such i in the PWM, $1, \dots, n$ where n is the length of the PWM (25).
9. Retrieve core scores and PWM scores at the Top X % percentile (one-tailed p -value is less than $X/100$).
10. Filter these i candidate motifs to those which meet any additional experimental constraints, if any (see Note 4).
11. Apply the Fisher test to measure the significance of the motifs using nonenrichment (or control) data (see Note 5).
12. Discard nonsignificant motifs, i.e., motifs with a significance of $p > 0.001$, to obtain a significant set of m putative motifs.
13. Feed this set of motifs and their PWMs to STAMP (26) for phylogenetic hierarchical clustering and comparison with TRANSFAC (14) and JASPAR (16) known motifs.
14. STAMP will output the final set of n motifs with significant similarity to known motifs (see Note 6).

2.2. Introduction to W-ChIPMotifs

The flow chart in Fig. 1b illustrates the workflow of our Web-based implementation of this algorithm, W-ChIPMotifs. Usage of W-ChIPMotifs web service is simple and does not require any knowledge of the underlying software (<http://motif.bmi.ohio-state.edu/>)

¹It is important to use a large enough number of sequences to get statistically significant results from de novo motif discovery. Use at least ten different sequences; however, there are also technical concerns: MEME performs best with less than 2,000 input sequences.

²The W-ChIPMotifs currently include three ab initio motif programs: MEME, MaMF, and Weeder. We will plan to add more programs in the next version of program.

³In step 7 of Subheading 2.1, these randomized sequences no longer correspond to binding sites, but have the same nucleotide frequencies as the original binding sites and are therefore used as a negative control set for motif finding.

⁴In step 10 of Subheading 2.1, for many experiments there will be no such additional constraints. See Subheading 2.7, step 11 for an example.

⁵It is very important to use Bonferroni correction to adjust the p -value by multiplying by the number of samples being input in order to reduce inaccuracy from small sample sizes.

⁶Common transcription factors with poorly specified positional weight matrices may show up as matches from STAMP with poor but possibly acceptable p -values. Experience and background knowledge are important in interpreting these results.

ChIPMotifs). There are three required inputs from the user: the DNA sequence data, contact information, and a transcription factor name. DNA sequences are required to be in the FASTA format. They can be uploaded either by selecting an existing file or by directly copying the data into the form. Results will be emailed to the address given in the contact information. The transcription factor name is used as a label in the results. Also, control data can be specified as an optional input, which is used to infer the statistical significance for detected motifs. In case of no control data input from users, we will use default control data sets where we randomly selected 5,000 promoter sequences per run from all human or mouse promoter sequences depending on the user selected species.

2.3. W-ChIPMotifs Workflow

1. Select the input data set of the top k sequences ordered by significance (see Note 1).
2. Provide these sequences in a FASTA format and contact information.
3. Optionally provide control data. If no control data is submitted, a default control data set is used composing 5,000 randomly selected promoter sequences from all promoter region sequences in the target species.
4. Process the input data set in Weeder (see Note 2).
5. Process the input data set in MEME.
6. Process the input data set in MaMF.
7. The union of the output of these two programs is the set of candidate motifs, of size i .
8. Construct position weight matrices for each of the i candidate motifs.
9. Perform bootstrap re-sampling by randomizing each of the user input's sequences for 100 times (see Note 3).
10. These randomized sequences are used for scanning the identified motifs (represented with PWMs, from step 8) at a minimal core score of 0.5 and a minimal PWM score of 0.5.
11. Retrieve core and PWM scores at the top 0.1, 0.5, and 1% percentiles.
12. Apply the Fisher test to measure the significance of each motif.
13. We also apply the Bonferroni correction by adjusting the p -value multiplying by the number of samples being input. If the adjusted p -value ended up greater than 1.0, it would be rounded down to 1.0 (see Note 5).
14. Discard nonsignificant motifs, i.e., motifs with a significance of $p > 0.001$, to obtain a significant set of m putative motifs.
15. Feed this set of motifs and their PWMs to STAMP for phylogenetic hierarchical clustering and comparison with TRANSFAC and JASPAR known motifs.

16. STAMP will output the final set of n motifs with significant similarity to known motifs.
17. The results from W-ChIPMotifs are composed of two files. The first file contains detected motifs with their SeqLOGOs, PWMs, core and PWM scores, p -values, and Bonferroni correction p -value at different percentile levels. The second file contains matched similar motifs from the STAMP tool. These files are in PDF format.

2.4. W-ChIPMotifs Implementation

W-ChIPMotifs is written in Perl, and uses a Web interface developed with PHP. Multiple scripts are used to produce output from the included motif discovery programs, parse this output, and apply statistical techniques. The sequence logos for the motifs are generated using the WEBLOGO tool. The open-source HTMLDOC program is used to convert these logos to PDF format (<http://www.htmldoc.org/>). A tree in Newick format is created with the DRAWTREE tool (see Note 7). The PHPMailer package is used for sending results to the user from the W-ChIPMotifs email account.

2.5. Case Studies for De Novo Motif Discovery of OCT4 and ZNF263

We present two case studies in the application of these techniques (see the sample data at <http://motif.bmi.ohio-state.edu/BookChIPMotifs>). The study in OCT4 illustrates how in vivo ChIP sequence data can be used to computationally predict motifs ab initio. The ZNF263 research shows that computationally predicted motifs may differ from in vitro predicted motifs while still having high predictive capability, i.e., they can be used to identify sites on the genome which correlate with the genome wide in vivo experimental results.

2.6. In Vivo OCT4 Motif Discovery

Recent ChIP-chip studies have revealed that many in vivo binding sites have a weak match to the consensus sequence for the transcription factor being analyzed. Possible explanations for these observations include (a) the consensus site was derived from in vitro analyses and does not represent the preferred in vivo binding site and/or (b) the factor is recruited to a weak binding site via interaction with a protein that binds nearby. To investigate case (b), we performed the following analysis. Using OCT4 ChIP-chip data derived from genomic tiling arrays and the ChIPMotifs approach, we developed a refined OCT4 PWM. We then used the in vivo derived PWM and a ChIPModules approach to identify transcription factors co-localizing with OCT4 in a testicular germ cell tumor (Ntera2 cells). We found that the consensus binding site for SRY, a transcription factor critical for testis development, co-localizes with the OCT4 PWM. To further characterize the relationship between OCT4 and SRY binding sites, we used ChIP-chip analysis of human promoter microarrays, and found that 49% of the top ~1,000 OCT4 target promoters were also bound by SRY. This analysis represents the first identification of SRY target promoters. Our studies not only validate the ChIP-Motifs and ChIPModules combinatorial approach but also identify a possible new regulatory partner of OCT4.

⁷“Newick format” is a common textual representation of a tree graph.

2.7. Methods for OCT4 Data

1. Input a set of 154 in vivo OCT4 binding sequences into the Weeder and MEME programs (see Note 1).
2. Using these programs, we identified ten candidate motifs, each having a length of 8–12 bp.
3. We then constructed ten positional weight matrices for each candidate motif.
4. We randomized the sequences of each of the 154 OCT4 binding sequences 100 times to generate a set of 15,400 randomized sequences (see Note 8).
5. We then scanned these randomized sequences for each candidate motif (using the PWMs derived from Weeder and MEME) starting at a minimal core score of 0.5 and a minimal PWM score of 0.5.
6. We retrieved core scores and PWM scores at the Top 0.1% percentile (one-tailed p -value is less than 0.001).
7. We retrieved core scores and PWM scores at the Top 0.5% percentile (one-tailed p -value is less than 0.005, see Note 9).
8. We retrieved core scores and PWM scores at the Top 1% percentile (one-tailed p -value is less than 0.01).
9. Using these scores, we tested the 154 OCT4 binding regions (Dataset 1) and 499 regions that were not bound by OCT4 (defined as Dataset 2).
10. A Fisher test was applied and the p -value was used to define the significance measure for this data (see Note 5).
11. We filtered the set by keeping only those motifs that were found in the OCT4 binding sites, but not in the control Dataset 2, which were considered to be over-represented motifs.
12. These motifs have a confidence level at the Top 0.1% percentile and a Fisher test p -value less than 0.001. Thus, a p -value of 0.00026 for the OCT4H_PWM at the top 0.1% percentile with a core score of 0.88 and PWM score of 0.85 (see Note 10) is considered to be significant, nonsignificant motifs are discarded (Fig. 2).

2.8. The Results for OCT4 Data

The motif NATGCAAANN which resembles the OCT4 consensus site of ATGCAAAT (Fig. 2a) was identified. We found that a 0.88 match to the core sequences (S_c) and a 0.85 match to the PWM (S_p) clearly distinguishes the OCT4 dataset from the control set (with a p -value of 0.00026) and demonstrates high specificity (eliminating 60% of the fragments in

⁸In step 4 of Subheading 2.7, these randomized sequences no longer correspond to binding sites, but have the same nucleotide frequencies as the original binding sites and are therefore used as a negative control set for motif finding.

⁹In steps 6–8 of Subheading 2.7, allowing too many changes from the consensus motif results in the identification of OCT4 binding sites in the great majority of both datasets, whereas requiring a complete match to the consensus eliminates the majority of the true binding sites.

¹⁰We compute any possible six consecutive nucleotides for the OCT4H_PWM and define the one with a maximum value as a core and the corresponding value as core score, while a sum of the OCT4H_PWM is considered as PWM score.

the negative control set) and high sensitivity (capturing ~70% of the binding sites). However, when using 0.88 (S_c) and 0.85 (S_p) criteria, 28.6% of the experimentally determined Oct4 binding regions still lack a match to the OCT4H_PWM.

2.9. In Vivo Motif Discovery for ZNF263

Recent in vitro studies (27) have shown that approximately half of a set of 104 mouse DNA-binding proteins recognized multiple different sequence motifs. Half of all human transcription factors use C2H2 zinc finger domains to specify site-specific DNA binding and yet very little is known about their role in gene regulation. Based on in vitro studies, a zinc finger code has been developed that predicts a binding motif for a particular zinc finger factor (ZNF). However, very few studies have performed genome-wide analyses of ZNF binding patterns, and thus, it is not clear if the binding code developed in vitro will be useful for identifying target genes of a particular ZNF. We performed genome-wide ChIP-seq for ZNF263, a C2H2 ZNF that contains nine finger domains, a KRAB repression domain, and a SCAN domain and identified more than 5,000 binding sites in K562 cells (28). Although ZNFs containing a KRAB domain are thought to function mainly as transcriptional repressors, many of the ZNF263 target genes are expressed at high levels. To address the biological role of ZNF263, we identified genes whose expression was altered by treatment of cells with ZNF263-specific small interfering RNAs. Our results suggest that ZNF263 can have both positive and negative effects on transcriptional regulation of its target genes.

2.10. Methods for ZNF263 Data

1. We identified a set of 1,473 binding sites in common in the two ChIP-seq experiments at the top 0.1% level to derive an in vivo binding motif for ZNF263 (see Note 1).
2. A set of ~24,000 human promoter sequences of 500 bp in length for each promoter from 1,000 bp upstream to the 5' transcription start site were selected as a negative control data set.
3. Process the input data set in Weeder.
4. Process the input data set in MEME.
5. Process the input data set in MaMF.
6. The union of the output of these two programs is the set of candidate motifs.
7. Construct position weight matrices for each of the candidate motifs.
8. Perform bootstrap re-sampling by randomizing each of 1,473 sequences for 100 times.
9. Scanned these randomized sequences for each candidate motif (using the PWMs derived from step 11) starting at a minimal core score of 0.5 and a minimal PWM score of 0.5.
10. Retrieve core scores and PWM scores at the Top 0.1% percentile (one-tailed p -value is less than 0.001).

11. Filter these candidate motifs to those which are over-represented in the input set compared to the negative control set.
12. Apply the Fisher test to measure the significance measure for the motifs (see Note 5).
13. Discard nonsignificant motifs, i.e., motifs with a significance of $p > 0.001$, to obtain a significant set of m putative motifs.
14. Feed this set of motifs and their PWMs to STAMP for phylogenetic hierarchical clustering and comparison with TRANSFAC and JASPAR known motifs.
15. STAMP will output the final set of n motifs with significant similarity to known motifs.
16. A de novo ZNF263 motif (Fig. 3a) is then determined.
17. For those ZNF263 binding sites without a good match to the first identified novel ZNF263 motif, ChIPMotifs were further run on these sites, and other known or novel motifs were then determined.
18. To obtain a motif predicted for ZNF263 by the zinc finger code, we used a prediction program ZIFIBI that predicts binding sites for zinc finger domains (see Note 11).
19. We merged the individual triplet predictions to obtain a predicted WebLogo for fingers 2–9 (Fig. 3b).
20. To search a set of genomic regions for the predicted motif, we adapted the WebLogo to create a nucleotide string; the sequence NNGGANGANGGGANNANGGA was used as the predicted motif bound by fingers 2–9.
21. Because there is a gap between fingers 5 and 6, we also made individual motifs for fingers 2–5 and 6–9; the sequence NGGGANNANGGA was used as the motif bound by fingers 2–5, and the sequence NNGGANGANGGA was used as the motif bound by fingers 6–9.

2.11. The Results for ZNF263 Data

We used in vivo derived ZNF263 PWM to scan a set of 5,273 sites identified from the Top 0.5% level from two biological replicates in K562 cells (28). We found that 75% of the 5,273 sites contained a good match (Core/position weight matrix 0.80/0.75) to this motif. We next examined the distribution of this motif in the two largest categories of ZNF263 binding site locations, promoters, and introns. We found that 86% of the 5' transcription start site category and 73% of the intragenic category contained this site. Therefore, it seems that ZNF263 is recruited to the intragenic sites using the same motif as used in the core promoter regions. Our results suggest that ZNF263 binds to a 24-nt site, Fig. 3a, that differs

¹¹In step 18 of Subheading 2.10, this program predicted motifs for fingers 2–3–4, 3–4–5, 6–7–8, and 7–8–9.

from the motif predicted by the zinc finger code in several positions. Interestingly, many of the ZNF263 binding sites are located within the transcribed region of the target gene.

References

1. Lockhart D, Dong H, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol.* 1996; 14:1675–1680. [PubMed: 9634850]
2. Schena M, Shalon D, Davis RW, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 1995; 270:467–470. [PubMed: 7569999]
3. Iyer VR, Horak CE, Scafe CS, et al. Genomic binding sites of the yeast cell-cycle transcription factor SBF and MBF. *Nature.* 2001; 409:533–538. [PubMed: 11206552]
4. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. *Science.* 2000; 290:2306–2309. [PubMed: 11125145]
5. Steensel B, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol.* 2000; 18:424–428. [PubMed: 10748524]
6. Crawford GE, Davis S, Scacheri PC, et al. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods.* 2006; 3:503–509. [PubMed: 16791207]
7. Loh YH, Wu Q, Chew JL, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genet.* 2006; 38:431–440. [PubMed: 16518401]
8. Pedersen JT, Moulton J. Genetic algorithms for protein structure prediction. *Curr Opin Struct Biol.* 1996; 6:227–231. [PubMed: 8728656]
9. Lawrence C, Altschul S, Boguski M, et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science.* 1993; 262:208–214. [PubMed: 8211139]
10. Bailey TL, Elkan C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol.* 1995; 3:21–29. [PubMed: 7584439]
11. Pavesi G, Mereghetti P, Mauri G, et al. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 2004; 32:W199–203. [PubMed: 15215380]
12. Liu J, Stormo GD. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics.* 2008; 24:1850–1857. [PubMed: 18586699]
13. Kel AE, Gossling E, Reuter I, et al. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 2003; 31:3576–3579. [PubMed: 12824369]
14. Wingender E, Chen X, Hehl R, et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 2000; 28:316–319. [PubMed: 10592259]
15. Alkema WB, Johansson O, Lagergren J, et al. MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.* 2004; 32:W195–198. [PubMed: 15215379]
16. Sandelin A, Alkema W, Engstrom P, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004; 32:D91–94. [PubMed: 14681366]
17. Weinmann AS, Yan PS, Oberley MJ, et al. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Gene Dev.* 2002; 16:235–244. [PubMed: 11799066]
18. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell.* 2007; 129:823–837. [PubMed: 17512414]
19. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods.* 2007; 4:651–657. [PubMed: 17558387]
20. Ettwiller L, Paten B, Ramialison M, et al. Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Methods.* 2007; 4:563–565. [PubMed: 17589518]
21. Gordon DB, Nekudova L, McCallum, et al. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics.* 2005; 21:3164–3165. [PubMed: 15905282]
22. Hong P, Liu XS, Zhou Q, et al. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics.* 2005; 21:2636–2643. [PubMed: 15817698]

23. Jin VX, O'Geen H, Iyengar S, et al. Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Res.* 2007; 17:807–817. [PubMed: 17567999]
24. Jin VX, Apostolos J, Nagisetty NS, et al. W-ChIPMotifs: a web application tool for de novo motif discovery from ChIP-based high-throughput data. *Bioinformatics.* 2009; 25:3191–3193. [PubMed: 19797408]
25. Jin VX, Leu YW, Liyanarachchi S, et al. Identifying estrogen receptor alpha target genes using integrated computational genomics and chromatin immunoprecipitation microarray. *Nucleic Acids Res.* 2004; 32:6627–6635. [PubMed: 15608294]
26. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* 2007; 35:W253–258. [PubMed: 17478497]
27. Badis G, Berger MF, Philippakis AA, et al. Diversity and complexity in DNA recognition by transcription factors. *Science.* 2009; 324:1720–1723. [PubMed: 19443739]
28. Fietze S, Lan X, Jin VX, et al. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263 (ZNF263). *J Biol Chem.* 2010; 285:1393–1403. [PubMed: 19887448]

Fig. 1.

Ab initio motif discovery workflows. **(a)** The general ChIPMotifs workflow from initial data selection to the final set of motif predictions. **(b)** The workflow of W-ChIPMotifs, in more detail, with the specific input file formats and types of data being processed at each stage. The overall workflow can be summarized as follows: (1) select input sequences; (2) run Weeder, MEME, and MaMF on the input; (3) use bootstrap re-sampling and the Fischer exact test to filter the output by quality; (4) use STAMP to predict a phylogenetic hierarchy of the results and identify matches to existing motifs.

Fig. 2.

The OCT4 motif and position frequency matrix. The motif nATGCAAAnn, **(b)** which resembles the OCT4 consensus site, **(a)** of ATGCAAAT was identified. Importantly, our CHIPMotifs analysis provided not only a consensus site, but also a position frequency matrix, **(c)** (OCT4H_PWM) for in vivo OCT4 binding in the above table.

Fig. 3. Comparison of in vivo and in vitro predicted ZNF263 motifs. **(a)** A WebLogo representing the 24 nt experimentally in vivo derived ZNF263 binding site is shown. **(b)** A WebLogo representing the ZNF263 binding site in vitro predicted using the zinc finger code is shown. ZNFs bind in the C-terminal to N-terminal orientation; therefore, the first 12 nt in the motif are those predicted to be bound by fingers 9–6, and the second 12 nt in the motif are those predicted to be bound by fingers 5–2. For searching of the ZNF263 binding sites for the predicted motif, the sequence nnGGAnGANGGAnGGGAnnAnGGA was used as the motif bound by fingers 2–9; the sequence nGGGAnnAnGGA was used as the motif bound by fingers 2–5, and the sequence nnGGAnGANGGA was used as the motif bound by fingers 6–9.