Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock

Yuichi Kumakir,†, Maki Ukai-Tadenuma*, Ken-ichiro D. Uno†, Junko Nishio†, Koh-hei Masumoto*, Mamoru Nagano†, Takashi Komori†, Yasufumi Shigeyoshi†, John B. Hogenesch†, and Hirotoki Ueda*‡

*Laboratory for Systems Biology and †Functional Genomics Unit, Center for Developmental Biology, RIKEN, 2-2-3 Minatojima-Minamimachi, Chuo-ku, Kobe 650-0047, Japan; ‡Department of Anatomy and Neurobiology, Kinki University School of Medicine, 377-2 Ohno-Higashi, Osaka-Sayama, Osaka 589-8511, Japan; and †Institute for Translational Medicine and Therapeutics and the Department of Pharmacology, University of Pennsylvania School of Medicine, 810 Biomedical Research Building II/III, 421 Curie Boulevard, Philadelphia, PA 19104-6160

Edited by Joseph S. Takahashi, Northwestern University, Evanston, IL, and approved August 15, 2008 (received for review March 16, 2008)

Mammalian circadian clocks consist of regulatory loops mediated by Clock/Bmal1-binding elements, DBP/E4BP4 binding elements, and RevErba/ROR binding elements. As a step toward system-level understanding of the dynamic transcriptional regulation of the oscillator, we constructed and used a mammalian promoter/enhancer database (http://promoter.cdb.riken.jp/) with computational models of the Clock/Bmal1-binding elements, DBP/E4BP4 binding elements, and RevErba/ROR binding elements to predict new targets of the clock and subsequently validated these targets at the level of the cell and organism. We further demonstrated the predictive nature of these models by generating and testing synthetic regulatory elements that do not occur in nature and showed that these elements produced high-amplitude circadian gene regulation. Biochemical experiments to characterize these synthetic elements revealed the importance of the affinity balance between transactivators and transrepressors in generating high-amplitude circadian transcriptional output. These results highlight the power of comparative genomics approaches for system-level identification and knowledge-based design of dynamic regulatory circuits.

Results and Discussion

The rapidly expanding number of sequenced mammalian genomes (1–3), annotated and cloned full-length cDNAs (4–6), transcriptional starts sites (TSSs) (7–9) and transcription factor binding sites (TFBSs) (10–12) has provided new opportunities to unravel the control of dynamic transcriptional programs. Comparative genomics approaches applying these resources have been used to identify target genes of specific biological pathways. These efforts used consensus sequence searches (13, 14), positional weight matrices (15), hidden Markov models (HMMs) (16), and specifically tailored algorithms (17, 18) to define candidate response elements and target genes in raw genomic sequence. Additionally, post hoc analysis employing evolutionary conservation (15, 16, 18) together with positional information of TSSs (15) and/or translational start sites (16) has helped to further define candidate elements and genes and greatly expanded our knowledge of transcriptional output regulation.

The mammalian circadian clock is an ideal system to apply these tools as it consists of integrated transcriptional regulatory loops that direct output through at least three types of transcriptional regulatory elements, the Clock/Bmal1-binding elements (E-box) (CACGTG) (19–21), DBP/E4BP4 binding elements (D-box) (TTATGT[C]/AAA) (21–23), and RevErba/ROR binding elements (RRE) (A[T]A[A/T]N[AG]GGTCA) (15, 21, 24, 25). Several groups, including our own, have shown that approximately 5–10% of mammalian genes display circadian expression in central and peripheral clock tissues (26). However, for the most part, the transcriptional regulation of these thousands of clock-controlled genes has remained uncharacterized. We and others have used comparative genomics approaches to analyze E-box (21, 27, 28), D-box (21), and RRE (15, 21), highlighting the importance of both their core consensus and flanking sequences (15, 21, 27, 28) in circadian gene control. In this study, we further extend comparative genomics approaches toward a system-level understanding of the dynamic transcriptional regulations of the mammalian circadian clock.

Prediction of Direct Clock Targets Through Utilization of the Mammalian Promoter/Enhancer Database. To generate a resource that facilitates identification of clock-controlled genes, we constructed a mammalian promoter/enhancer database (http://promoter.cdb.riken.jp/) by integrating information sources such as conserved non-coding regions, TSSs and TFBSs [supporting information (SI) Fig. S1 and SI Appendix]. Although excellent and similar databases exist such as DBTSS (8), CisView (29) and ECRbase (30), none were tailored to specifically identify clock gene targets and having local control of the database facilitated manipulation of the underlying data (see also SI Appendix). We then developed a comparative genomics strategy employing this database and profile HMMs using the HMMER software package (31). Profile HMMs are powerful tools to extract the statistical properties of input sequences by representing multiple sequences as a transition probability matrix marching from one position to the neighboring position. HMMs were built and calibrated on known functional clock-controlled elements experimentally verified in our previous (15, 21) or current studies (Fig. S2 and Table S1), consisting of 12 E-boxes, 10 D-boxes and 15 RREs (Table S2). Profile HMM searches to identify new clock-controlled elements from conserved non-coding regions between human and mouse identified 1,108 E-boxes, 2,314 D-boxes, and 3,288 RREs candidate elements (see the circadian section of the mammalian promoter/enhancer database: http://promoter.cdb.riken.jp/circadian.html for element lists). To set appropriate reporting thresholds, we used the match scores of known functional clock-controlled elements (Material and Methods). Predicted clock-controlled elements exhibited an un-biased distribution of chromosomal position spread over whole mouse genome (Fig. 1 and http://promoter.cdb.riken.jp/circadian.html).


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freedly available online through the PNAS open access option.

‡Present address: Laboratory for Mammalian Epigenetic Studies, Center for Developmental Biology, RIKEN, 2-2-3 Minatojima-Minamimachi, Chuo-ku, Kobe 650-0047, Japan.

**To whom correspondence should be addressed. E-mail: uedah-tky@umin.ac.jp.

This article contains supporting information online at www.pnas.org/cgi/content/full/0802636105/DCSupplemental.

© 2008 by The National Academy of Sciences of the USA
genes with D-boxes, and 34 genes with RREs exhibited significant circadian average expression of transcripts harboring each element exhibit circadian indicated. E-box,'' ''Random D-box,'' and ''Random RRE,'' see also 3,288 RREs). The E-box displays a biased distribution of distance from TSSs, while the D-box and RRE show unbiased, near random distributions (''Random elements of conserved non-coding regions (1,108 E-boxes, 2,314 D-boxes, and from transcriptional starts sites (TSSs) for predicted conserved clock-controlled regions (blue, ''Non-coding region'') and (3) searches in the entire genome elements within conserved non-coding regions (red, “Conserved element”), (2) searches for mouse elements within conserved or non-conserved non-coding regions (blue, “Non-coding region”) and (3) searches in the entire genome relaxing both element conservation and search space (orange, “Whole genome”). FDRs in “Conserved elements” search are plotted against the average match score of human and mouse elements. (C) The distributions of distance from transcriptional starts sites (TSSs) for predicted conserved clock-controlled elements of conserved non-coding regions (1,108 E-boxes, 2,314 D-boxes, and 3,288 RREs). The E-box displays a biased distribution of distance from TSSs, while the D-box and RRE show unbiased, near random distributions (“Random E-box,” “Random D-box,” and “Random RRE,” see also SI Appendix). (D) The average expression of transcripts harboring each element exhibit circadian rhythms in the liver. The average expression of 36 genes with E-boxes, 29 genes with D-boxes, and 34 genes with RREs exhibited significant circadian oscillations (P = 0.01 for E-box, P = 0.0005 for D-box, and P = 0.001 for RRE). Data were normalized so that the average signal intensity over 12-point time courses is 1.0. Estimated peak times of circadian oscillation were also indicated.

However, the match score on its own does not give a good estimate of accuracy and false discovery from HMM searches. To better estimate the prediction for each HMM, we searched each against randomized genome sequence to generate a background distribution of false positive occurrences (see SI Appendix for detail). We found that the value of the false discovery rate (FDR) is inversely proportional to the match score of the HMM, which is a representation of the statistical significance of the candidate element (Fig. 1B). Importantly, we found the accuracy of the HMM-based prediction as measured by the FDR is dependent on the initial search conditions. HMM searches in conserved non-coding regions (the original condition) had the lowest FDR, while higher rates were observed in conserved or non-conserved non-coding regions, or in searches of raw genome sequence (Fig. 1B). These results demonstrate the value of using human/mouse conservation and a confined search space for generation of the most accurate response element predictions.

Several intriguing features resulted from this analysis. Interestingly, like cAMP-responsive elements (CREs) (16), putative E-box displayed a biased distance distribution from TSSs, while putative D-box and RRE had random localization distributions in the genome and were not more likely to be near TSSs (Fig. 1C). This result is consistent with and extends an earlier report that described the preferential localization of CpG containing transcription factor binding sites including the E-box and CRE to proximal promoter regions of housekeeping genes (32). In addition to housekeeping genes, we see circadian E-box sequences present in many genes with specific functions such as enzymes and signaling molecules.

Gratifyingly, we noted a significant enrichment of previously identified clock-controlled genes (15, 33) in our predicted clock-controlled elements (the 100 most significant sequences for each HMM search, E-box, D-box, and RRE, respectively). After removing the 21 clock-controlled genes used for HMM generation and training, we found an additional 19 putative clock-controlled genes (out of the 6,195 genes common in our mammalian promoter/enhancer database and U74 mouse microarray) versus the expected 10.67 that would have arisen from chance, a significant enrichment of clock-controlled genes (P < 0.01, see also SI Appendix).

The presence of statistically significant clock-controlled elements in the promoters of these genes suggests that their message levels may oscillate. To examine this possibility, we selected the 100 most significant putative clock-controlled elements for each model (E-box, D-box, and RRE) and determined their gene expression levels from previously obtained liver data (15). The average expression of 36 E-box containing genes, 29 D-box genes, and 34 RRE genes exhibited significant circadian oscillations (P = 0.01 for E-box, P = 0.0005 for D-box, and P = 0.001 for RRE) with a surprisingly consistent peak time of expression (E-box = 8.8, D-box = 10.8, and RRE = 13.6, Fig. 1D). Taken in sum, these data show that our HMM models and strategy identifies elements and genes that oscillate with a circadian period with peak phases of expression that are consistent with the previously reported literature.

**In Vitro Validation of Putative Clock Controlled Genes.** To provide further validation of these predictions, we used an *in vitro* system of the autonomous circadian clock to empirically test candidate elements in circadian transcriptional output assays. In brief, we used a cell culture system (15, 34) that allowed the monitoring of circadian transcriptional dynamics using a destabilized luciferase reporter (dLuc) driven by known or putative clock-controlled response elements (Fig. 2A). We selected the ten most significant sequences for each HMM search, E-box, D-box, and RRE respectively, and located within 1kb of the TSS (Table S3). In sum, these data show that our HMM models and strategy identifies elements and genes that oscillate with a circadian period with peak phases of expression that are consistent with the previously reported literature.
Investigating the Contribution of Flanking Sequences. Using these synthetic elements, we next attempted to explore the contribution of E-box flanking regions to identify critical residues that modulate amplitude and rhythmicity. We clustered their nucleotide sequences, and, interestingly, found two patterns of high-amplitude E-box flanking sequences adjacent to the core CACGTG element (Fig. S4). However, these positions do not absolutely dictate high-amplitude rhythmicity, as some elements that meet these rules exhibit lower-amplitude oscillations, possibly because they exhibit much higher GC content. In either case, these experimental results also imply that amplitude information is encoded in specific residues adjacent to the core consensus element and further strengthen the previous reports by other groups on the importance of flanking sequence of E-box (27, 28, 35, 36). Interestingly, the identified patterns in this study partly overlap with the computational models based on the evolutionarily conserved E-box structure from insects to mammals (27).

High-Amplitude Oscillations Require Appropriate Affinity Balance Between Activators and Repressors. To explore the properties of these elements that result in high amplitude oscillations, we took an amplified molecular modeling and experimental approach. First, we assumed concentrations of activators and repressors were within similar ranges (see also SI Appendix Discussion in more general cases). We further hypothesized that flanking region DNA sequence impacted DNA-binding affinity of clock gene regulators and therefore altered amplitude. We further hypothesized that tightly binding sequences would have higher amplitudes of circadian oscillation. To test this notion, we analyzed the DNA-binding affinity of activators and repressors.
to these response elements using competitive binding assays (Fig. 4A and Fig. S5A). For the D-box and RRE elements, “high-scoring” elements showed approximately the same DNA-binding affinity for their activators and repressors, while “low-scoring” elements of D-box and RRE showed relatively weak affinity, confirming this hypothesis. Surprisingly, in the case of E-box, “low-scoring” sequences had a higher affinity for the Arntl/Clock activator complex (4.8 times higher than positive control; Fig. S5A) than “high-scoring” sequences or the positive control, whereas the “low-scoring” E-box sequence showed approximately the same affinity to the Bhlhb2 repressor.

To assist in interpreting these results, we used in silico modeling (Fig. 4B and C) and treated affinity of activators and repressors as parameters and amplitude as the output of the model. Interestingly, this analysis showed that a high affinity activator complex coupled with a normal affinity repressor complex capitulated lower amplitude rhythms (Fig. 4B), suggesting that the enhanced retention of an activator alone promotes its saturation on a promoter and consequently dampens amplitude in competition-based models. Further, in silico analysis showed not only affinity strength of the activator and repressor (Fig. S5B) but also the appropriate affinity balance of activators and repressors is necessary for high-amplitude circadian oscillations (Fig. 4C; See also SI Appendix Discussions for in silico modeling).

Supporting this notion, a new clock gene, clockwork orange (cwo, a fly ortholog of mammalian Bhlhb2 and Bhlhb3) was recently reported to directly suppress gene expression of several clock genes through E-box elements (37–39). Quantitative and qualitative impairment of cwo revealed an important role of this transcriptional repressor for high-amplitude oscillation of the Drosophila circadian clock. The findings in this report, along with the studies of the cwo gene, collectively show that a competitive balance between direct activator(s) and direct repressor(s) for the E-box element is important for driving high-amplitude oscillations of circadian output genes. In addition to this in vivo biological evidence in the fly, we listed the evolutionary conserved “low-scoring” E-boxes in the mammalian genome (predicted low-amplitude) as candidates for unbalanced affinities. Interestingly, this list includes E-boxes on core clock genes (Cry2, Bhlhb3, Nr1d1, and Ror), and some low-amplitude clock-controlled genes such as Il2, and to promote follow-up, is available at http://promoter.cdb.riken.jp/circadian.html.

What could explain the discrepancy between E-box motifs and
other circadian response elements? We hypothesized that these differences might be encoded at the protein sequence level of the DNA-binding domains of activators and repressors. Interestingly, the DNA binding domains of transactivators and transrepressors of the RRE and D-box are more similar to each other (65% identity and 91% homology for RRE regulatory elements and 44% identity and 69% homology for D-box regulators) than those of E-box transactivators and transrepressor (22–30% identity and 55–57% homology) (Table S4). Based on these findings, we speculate that the DNA-binding domains for transactivators and transrepressors of the RRE and D-box have evolved similar affinities. In contrast, the evolutionarily and structurally divergent regulators of E-boxes, 108 bHLH proteins including several families of activators and repressors, as well as the unrelated period and cryptochrome gene families, may have required the co-evolution of specific DNA-binding domains and E-box sequences with specific flanking regions to generate higher amplitude rhythmicity.

Conclusion

In summary, we have applied a comparative genomics strategy to the understanding of a dynamic transcriptional regulatory system, the mammalian circadian clock. Our informatics strategy employs a model-based search with excellent statistical properties, the evolutionary conservation of putative transcriptional regulatory elements across mouse and human non-coding regions, and statistical dissection of regulatory elements by each prediction. Experimental validation of this strategy in vitro and in vivo using real-time monitoring of transcriptional activity and quantitative PCR assay has led to the identification of dozens of novel clock-controlled genes and the elements that likely dictate their rhythmicity. High-scoring conserved E-boxes (mean HMM-score = 16.15) had a 40% rate of validation, while low-scoring conserved E-boxes (mean HMM-score = 2.5143) had a 7.1% probability of generating high-amplitude rhythmicity in reporter assays. Linear interpolation from these two numbers generates an estimate of approximately 347 novel conserved E-boxes that likely confer circadian rhythmicity (see also SI Appendix). Moreover, to demonstrate their predictive nature, we have taken these in silico models and designed synthetic elements that exhibit high-amplitude transcriptional rhythmicity as well as the best canonical regulatory elements. Furthermore, experimental measurement and in silico analysis of affinity of regulators to synthetic elements revealed the importance of the appropriate affinity balance between activators and repressors for rhythmicity, for E-box sequences, lower affinity DNA element generates higher amplitude rhythms. The experimental, analytical, and synthetic approaches discussed here are especially timely as genomics tools are increasingly uncovering the complexity and flexibility of transcriptional regulatory circuits. We predict the general themes and resources reported here will enhance understanding of the biology mediated by complex and dynamic transcriptional regulation including the mammalian circadian clock.

Materials and Methods

Detailed information on the construction of the mammalian promoter/ enhancer database, determination of distance from TSSs for natural and randomly positioned elements, calculation of FDR for putative elements, animals, genome sequences, oligonucleotide sequences, plasmid constructions, quantitative PCR, rhythmicity analysis of real-time bioluminescence data, amplitude analysis of real-time bioluminescence data, rhythmicity analysis of quantitative PCR data, over representation analysis of clock-controlled genes, estimation of the number of high-amplitude E-boxes, microarray expression data analysis of genes with predicted clock-controlled elements, affinity analysis of competitive DNA binding data, and in silico analysis of affinity to amplitude mechanism are available in SI Appendix.

Real-Time Circadian Reporter Assays. Real-time circadian assays were performed as previously described (40) with the following modifications. NIH 3T3 cells (American Type Culture Collection) were grown in DMEM (Invitrogen) supplemented with 10% FBS (JRH Biosciences) and antibiotics (25 units ml⁻¹ penicillin, 25 μg ml⁻¹ streptomycin; Invitrogen). Cells were plated at 5 × 10⁴ cells per well in 24-well plates 24 h before transfection. Cells were transfected with 0.32 μg of plasmids in total (0.13 μg reporter plasmid and 0.19 μg empty plasmid) per well using FuGENE6 (Roche Applied Science) according to the manufacturer’s instructions. After 72 h, medium in each well was replaced with 500 μl of culture medium (DMEM/10% FBS) supplemented with 10 mM Hepes (pH 7.2), 0.1 mM luciferin (Promega), antibiotics and 0.01 μM forskolin (nacalai tesque). Bioluminescence was measured with photomultiplier tube (PMT) detector assemblies (Hamamatsu Photonics). The modules and cultures were maintained in a darkroom at 30 °C and interfaced with computers for continuous data acquisition until 96 h after forskolin stimulation. Photons were counted 2 min at 24-min intervals.

Construction, Search, and Design of Putative cis-Acting Elements. A HMM is a statistical model in which the target system is assumed to be a Markov process with unknown parameters. A HMM describes a probability distribution over input training sequences, i.e., probabilities of the state transition and emission. The extracted model can be used to find the probability of query sequence that is a product of all transition and emission probabilities at various training sequences. Nucleotide sequences for known clock-controlled elements, 12 E-boxes (18 bp), 10 D-boxes (24 bp), and 15 RREs (23 bp), experimentally verified in previous (21) and current studies (Table S1 and Fig. S2), were used as a training dataset to construct HMMs. We also attempted to construct an HMM for the E-box, but were unable (i.e., positive controls exhibited poor scores) because of the small number of experimentally validated E-box (only three: Per2, Bhlhb3, and Cry1) and the relatively short core sequence of known functional clock-controlled elements. We trained 2048 HMMs of all possible clock-controlled elements, 12 E-boxes (18 bp), 10 D-boxes (24 bp), and 15 RREs (23 bp), experimentally verified in previous (21) and these were sufficient to produce circadian transcriptional activity in circadian reporter assays. These sites were aligned without gaps according to the direction of consensus sequences (TTATGTT/C)AA for the D-box; ref. 22), [AT][TA][AT][T]/[A][GG]GCTA for the RRE; ref. 24]. Because the consensus sequence for E-box is palindromic (CACGTG; ref. 19), we generated all possible alignments by changing sequence directions (forward and reverse) and selected one alignment as described below. These alignments were used to build HMMs using hmmt program in the HMMER 1.8.4 software package (31) with default parameters (using sim annealing, starting kt for sim annealing run as 5.0, and multiplier for sim annealing as 0.95). We used the older version 1.8.4 package (the current version is 2.3.2) in this study because the version 2 series was optimized for analysis of protein sequences. Following construction, models were used to search genomic regions for putative clock-controlled elements using the hmms program with default parameters (by using threshold matches score to report as 0) except use ‘-c’ option only for bidirectionally search. The average score was used in the search for the conserved elements between human and mouse. To select only one alignment for each E-box, we constructed 2048 HMMs of all possible alignments, and calculated match scores of 12 known E-box sequences in the original HMMER search and chose the alignment that generated the highest average match score for further work.

To design the “high-scoring” and “low-scoring” sequence of clock-controlled elements, bidirectional HMMER searches were performed against all possible sequences of the same lengths as training dataset (18 bp for E-box, 24 bp for D-box, and 23 bp for RRE) that contain ordinary consensus sequence at the center (CACGTG for E-box; ref. 19; TTATGTTA for D-box; ref. 22; [AT][TA][AT][T]/[A][GG]GCTA for RRE; ref. 24), then filtered out those that naturally exist in either the human or mouse genome. The sequence of the highest and lowest score was selected as the “high-scoring” and “low-scoring” sequences, respectively. All HMMER searches, except the directional search in the selection of E-box alignments, were performed bidirectionally. The higher score was adopted if match scores were obtained for both directions at the same position. The training data are available in Table S2. The HMMs are publicly available on the circadian section of the mammalian promoter/enhancer database: http://promoter.cdb.riken.jp/circadian.html.

Competitive DNA Binding Assays. In vitro transcription/translation of Flagged mouse protein from pMU2-Amtl, pMU2-Clock, pMU2-Bhlhb2, pMU2-Dbp, pMU2-Nfil3, pMU2-Nrd1, and pMU2-Rora were performed with TNT T7 Quick Coupled Transcription/Translation System (Promega) according to the manufacturer’s specification. The synthesized and radiolabeled protein was mixed in equal volume. The complementary oligonucleotides of three tandem repeats sequence of designed and control cis-acting elements, which were labeled with biotin on 5’-end or non-labeled (for competitor)
(Hokkaido System Science), were annealed to generate probes. Competitive DNA binding assays were performed with NoShift Transcription Factor Assay Kit (Novagen) according to the manufacturer’s specifications with the following modifications. Ten pmol biotinylated annealed oligonucleotides were incubated with competitor oligonucleotides (final concentration were 0, 1, 3, 10, 30, and 100 pmol) and 5 μl of in vitro transcribed/translated reticulocyte lysates in the binding mixture. After the samples bound to streptavidin-coated microarray plate, the wells were washed, and Anti-Flag M2 Monoclonal Antibody-Peroxidise Conjugate (SIGMA) was applied into the each well. The lysates in the binding mixture. After the samples bound to streptavidin-coated

ACKNOWLEDGMENTS. We thank Hajime Tei and Yoshiyuki Sakaki for Arnvl and Clock expression vectors, Yutaka Suzuki, Sumio Sugano, and Seiichi Hashimoto for advice on the mammalian promoter/enhancer database, Takao Kondo for high-throughput monitoring systems, Hideki Ukai, Ryohtaku Kito, and Kazuhito Yagita for real-time circadian reporter assay, Rikuhito G. Yamada, Tetsuya J. Kobayashi, and Tekeya Kasukawa for statistical analysis, and Michael Royle for critical reading. This work was supported by RIKEN Research Collaborations with Industry Program (H.R.U.), and in part, by RIKEN Center for Developmental Biology (CDB) internal grant (H.R.U.), New Energy and Industrial Technology Development Organization (NEDO) Scientific Project grant (H.R.U.) and a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (Genome Network Project to H.R.U.). J.B.H. is supported by grants from the National Institute of Mental Health and National Institute of Neurological Disorders and Stroke of the National Institutes of Health (P50MH074924–01, Joseph S. Takahashi, P. I, NIH/1R01NS054794–01A2, J.B.H. P. I, NINDS).