BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 581–583
doi:10.1093/bioinformatics/btr704

Genome analysis

Advance Access publication December 22, 2011

CLARE: Cracking the LAnguage of Regulatory Elements

Leila Taher[1,*,†], Leelavati Narlikar[2,†] and Ivan Ovcharenko[1,*]

[1]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MA 20894, USA and [2]Chemical Engineering and Process Development Division, National Chemical Laboratory, CSIR, Pune 411008, India

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: CLARE is a computational method designed to reveal sequence encryption of tissue-specific regulatory elements. Starting with a set of regulatory elements known to be active in a particular tissue/process, it learns the sequence code of the input set and builds a predictive model from features specific to those elements. The resulting model can then be applied to user-supplied genomic regions to identify novel candidate regulatory elements. CLARE's model also provides a detailed analysis of transcription factors that most likely bind to the elements, making it an invaluable tool for understanding mechanisms of tissue-specific gene regulation.

Availability: CLARE is freely accessible at http://clare.dcode.org/.

Contact: taherl@ncbi.nlm.nih.gov; ovcharen@nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 7, 2011; revised on October 29, 2011; accepted on December 18, 2011

1 INTRODUCTION

Gene regulatory elements play a primary role in transcriptional regulation, a process required for the proper execution of various genetic programs. Proteins called transcription factors (TFs) recognize and bind specific regulatory elements on the DNA, leading to the activation or repression of a target gene. Identification of these regulatory elements is therefore important for understanding the regulatory mechanisms involved in the development and functioning of an organism. Experimental approaches such as ChIP-Seq or ChIP-chip can help identify regulatory elements on a genome-wide scale by profiling a TF of interest or a coactivator like p300 or CBP that colocalizes with active enhancers (Heintzman *et al.*, 2009; Visel *et al.*, 2009). However, these methods are restricted by the number of TFs that can be profiled, the efficacy of antibodies, experimental noise, and the number of cell types that can be tested. Furthermore, these methods are usually unable to identify regulatory regions bound by an unknown TF. As a result, computational methods for identifying TF-binding sites and regulatory elements are rapidly gaining ground [see Su *et al.* (2010) for a review]. Few tools, however, are easily accessible online and convenient for molecular biologists who are not interested in getting through the hurdles of installing and maintaining software. Here, we present CLARE (Cracking the LAnguage of Regulatory Elements), a web interface for the method that we recently developed to predict regulatory elements active in a specific tissue or biological process (Narlikar *et al.*, 2010). CLARE is written in Perl and runs on a Linux platform. Our computational method has been shown to identify mammalian heart enhancers with validation rates similar to those obtained with ChIP-Seq experiments (Blow *et al.*, 2010). CLARE is freely accessible online at http://clare.dcode.org.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Author.

2 SYSTEM OVERVIEW

Regulatory elements active in a particular tissue or biological process are likely to be bound by a common set of TFs; by either activators (in case of enhancers and promoters) or repressors (in case of silencers) present in the nucleus. Consequently, binding sites of these TFs should be statistically overrepresented in bound regulatory elements. CLARE exploits this notion in a model that describes regulatory elements based on a weighted linear combination of TF-binding sites. Figure 1 illustrates the workflow of the CLARE web server.

2.1 CLARE input

The only input required from the user is a set of sequences of regulatory elements in FASTA format. Optionally, the user can also enter a set of sequences to serve as controls and a query sequence to search for putative regulatory elements.

2.2 Modeling regulatory elements

CLARE proceeds in three main steps:

2.2.1 Creating a control set In absence of a user-supplied control set, CLARE will construct one that is length- and GC-balanced with respect to the input set of regulatory elements. This ensures that CLARE does not train purely on the GC content that is, in general, different between non-functional regions and functional regions. The server-side control set is sampled from the non-coding portion of the human genome.

2.2.2 Feature mapping Each sequence from the input and control sets undergoes a transformation into a feature vector, with features describing the occurrence of putative TF-binding sites. For this purpose, each sequence is scanned using tfSearch [(Ovcharenko *et al.*, 2005), see Supplementary Materials] with known motifs from the TRANSFAC (Matys *et al.*, 2006), JASPAR (Bryne *et al.*, 2008) and UniPROBE (Robasky and Bulyk, 2011) databases as well as the top 10 overrepresented motifs among the regulatory

elements discovered *de novo* using a Gibbs sampler [implemented in PRIORITY (Narlikar *et al.*, 2007)].

2.2.3 Model training The problem of separating known regulatory elements from background sequence is posed in the form of linear regression, with the goal of learning the weight of each feature. CLARE uses LASSO (Tibshirani, 1996) that returns an L1-regularized solution to the problem. This ensures a small number of weights have a non-zero weight, the assumption being that most motifs are not part of the biological process under consideration, and are therefore irrelevant for classification.

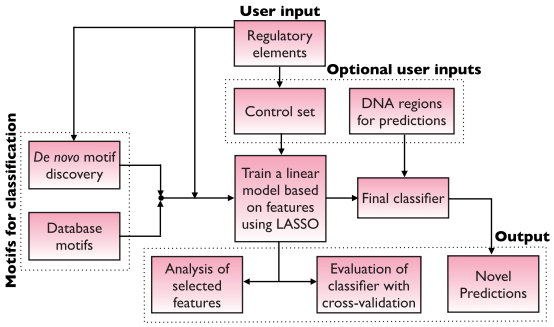

Fig. 1. The flowchart depicts the organization of CLARE.

2.3 CLARE output

After the job is completed, results are reported in a web-based table. CLARE provides three primary outputs:

2.3.1 Relevant features CLARE returns a bar graph displaying the weights of the features that are finally selected. A large positive (negative) weight implies that the respective feature is positively (negatively) correlated with the input set and negatively (positively) correlated with the control set.

2.3.2 Classifier performance The classifier performance is assessed by 5-fold cross-validation. The average receiver operating characteristic (ROC) curve and the area under it are reported.

2.3.3 Predictions CLARE returns a score for each of the candidate regions supplied by the user, indicating the likelihood of the sequence being a regulatory element.

A basic protocol for the utilization of CLARE is given in the Supplementary Materials. A typical CLARE job for input sets with up to a few hundred sequences is completed in <30 min. However, the run-time of CLARE is highly dependent on the number and size of input sets, and can run into hours for sets on the order 1000 sequences (Supplementary Materials). This is primarily due to the *de novo* motif discovery and LASSO.

3 CASE STUDY: ENHANCERS REGULATING FOREBRAIN EXPRESSION

As a test case, we used CLARE to predict sequences with forebrain enhancer activity in the human genome. The model was trained with 49 human sequences that were validated for forebrain enhancer activity in transgenic mouse assays (Visel *et al.*, 2009). The classifier achieved an area under the ROC curve of 0.83, suggesting it can effectively distinguish forebrain enhancers from background sequences. The training sequences were extracted from a larger dataset comprising 2453 p300 ChIP-Seq predictions in the genome of developing mouse embryos. After mapping the mouse peaks to the human genome, we selected a 2 Mb region in the human genome with the largest amount of p300 ChIP-Seq peaks for enhancer prediction. This region, on chromosome 5, encompasses the gene ZNF608, which is associated with Cornelia de Lange Syndrome (Liu *et al.*, 2009), characterized by slow growth before and after birth, intellectual disability and skeletal abnormalities. Thirteen-percent of the windows evaluated in this locus scored positively, comprising regions overlapping with all ChIP-Seq peaks (Supplementary Materials). These predictions also show significant overlap with regions conserved in chicken and mouse, with fold-enrichments of 2.4 and 1.5, respectively ($P < 0.05$), suggesting that they are functional and may therefore have clinical relevance (Loots and Ovcharenko, 2007). The datasets and results are available online at: http://clare.dcode.org/index.php?id=example.

4 CONCLUSIONS

CLARE is a novel web server, capable of predicting regulatory elements and functional TF-binding sites starting from a collection of known regulatory sequences. Detailed advantages of CLARE are listed in the Supplementary Materials. Briefly, this tool can aid in prioritizing genomic regions and TF-binding sites for further experimental validation. Ultimately, this should improve the functional annotation of the non-coding portion of the genome, and our understanding of the mechanisms controlling gene regulation in eukaryotes.

Funding: This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of Interest: none declared.

REFERENCES

Blow,M.J. *et al.* (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, **42**, 806–810.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

Bussemaker,H.J. *et al.* (2000) Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl Acad. Sci. USA*, **97**, 10096–10100.

Chang,L.W. *et al.* (2007) PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis. *Nucleic Acids Res.*, **35**, W238–W244.

Corcoran,D.L. *et al.* (2005) Footer: a quantitative comparative genomics method for efficient recognition of cis-regulatory elements. *Genome Res.*, **15**, 840–847.

Heintzman,N.D. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.

Liu,J. *et al.* (2009) Transcriptional dysregulation in NIPBL and cohesin mutant human cells. *PLoS Biol.*, **7**, e1000119.

Loots,G. and Ovcharenko,I. (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics*, **23**, 122–124.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Narlikar,L. *et al.* (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.

Narlikar,L. *et al*. (2010) Genome-wide discovery of human heart enhancers. *Genome Res*., **20**, 381–392.

Ovcharenko,I. *et al*. (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.*, **15**, 184–194.

Robasky,K. and Bulyk,M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*., **39**, D124–D128.

Su,J. *et al*. (2010) Assessing computational methods of cis-regulatory module prediction. *PLoS Comput. Biol*., **6**, e1001020.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat Methodol*, **58**, 267–288.

Visel,A. *et al*. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.