BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 17 2010, pages 2195–2197
doi:10.1093/bioinformatics/btq357

Sequence analysis

Advance Access publication July 13, 2010

PriorsEditor: a tool for the creation and use of positional priors in motif discovery

Kjetil Klepper* and Finn Drabløs

Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Computational methods designed to discover transcription factor binding sites in DNA sequences often have a tendency to make a lot of false predictions. One way to improve accuracy in motif discovery is to rely on positional priors to focus the search to parts of a sequence that are considered more likely to contain functional binding sites. We present here a program called PriorsEditor that can be used to create such positional priors tracks based on a combination of several features, including phylogenetic conservation, nucleosome occupancy, histone modifications, physical properties of the DNA helix and many more.

Availability: PriorsEditor is available as a web start application and downloadable archive from http://tare.medisin.ntnu.no/priorseditor (requires Java 1.6). The web site also provides tutorials, screenshots and example protocol scripts.

Contact: kjetil.klepper@ntnu.no

Received on April 21, 2010; revised on June 17, 2010; accepted on June 30, 2010

1 INTRODUCTION

Computational discovery of transcription factor binding sites in DNA sequences is a challenging problem that has attracted a lot of research in the bioinformatics community. So far more than a hundred methods have been proposed to target this problem (Sandve and Drabløs, 2006) and the number of publications on the topic is steadily increasing.

There are two general approaches for discovering potential transcription factor binding sites with computational tools. One is to examine regulatory regions associated with a group of genes that are believed to be regulated by the same factors and search for patterns that occur in all or most of these sequences. This approach, often referred to as *de novo* motif discovery, can be used when we have no prior expectations as to what the binding motifs might look like. One concern with this approach, however, is that it might be necessary to consider rather long sequence regions to ensure that the target sites are indeed covered. Since binding motifs for transcription factors are usually short and often allow for some degeneracy, the resulting signal-to-noise ratio can be quite low, making it difficult to properly discriminate motifs from background. Another problematic issue is that DNA sequences inherently contain a lot of repeating patterns, such as tandem repeats and transposable elements, which can draw focus away from the target binding motifs when searching for similarities between sequences.

The other general motif discovery approach, called *motif scanning*, searches for sequence matches to previously defined models of binding motifs, for instance in the form of position weight matrices (PWMs; Stormo, 2000). The main drawback with motif scanning is that it tends to result in an overwhelming number of false positive predictions. According to the 'futility theorem' put forward by Wasserman and Sandelin (2004), a genome-wide scan with a typical PWM could incur in the order of 1000 false hits per functional binding site, which would make such an approach practically infeasible for accurate determination of binding sites. The problem here lies not so much in the predicted binding patterns themselves, since many of these would readily be bound by transcription factors *in vitro*. *In vivo*, however, most such binding sites would be non-functional, perhaps because the chromatin conformation around the sites precludes access to the DNA (Segal *et al*., 2006) or because the target factors require the cooperative binding of additional factors nearby to properly exert their regulatory function (Ravasi *et al*., 2010).

One way to improve accuracy in motif discovery is to try to narrow down the sequence search space as much as possible beforehand, for instance, by masking out portions of the sequences that resemble known repeats or considering only sequence regions that are conserved between related species (Duret and Bucher, 1997). Kolbe *et al.* (2004) introduced a measure they called 'Regulatory Potential' which combines phylogenetic conservation with distinctive hexamer frequency profiles to identify possible regulatory regions. This measure calculates a score for each position along the sequence, and regions receiving higher scores are deemed more likely to have a regulatory role. Regulatory Potential can be considered as an example of a 'positional prior' since each position is associated with an a priori probability of possessing some specific property. Positional priors can be used as an aid in motif discovery by assigning high prior values to regions that we consider more likely to contain functional binding sites and then focus the search on these regions. Besides conservation and oligonucleotide frequencies, other features that can be relevant for assigning prior values include: localized physical properties of the DNA double helix, distance from transcription start site or other binding sites, ChIP-chip and ChIP-seq data, and potentially tissue-specific epigenetic factors such as the presence of nucleosomes and associated histone modifications. Many of the aforementioned features have previously been applied and shown to improve the performance of motif discovery by themselves (see e.g. Bellora

*To whom correspondence should be addressed.

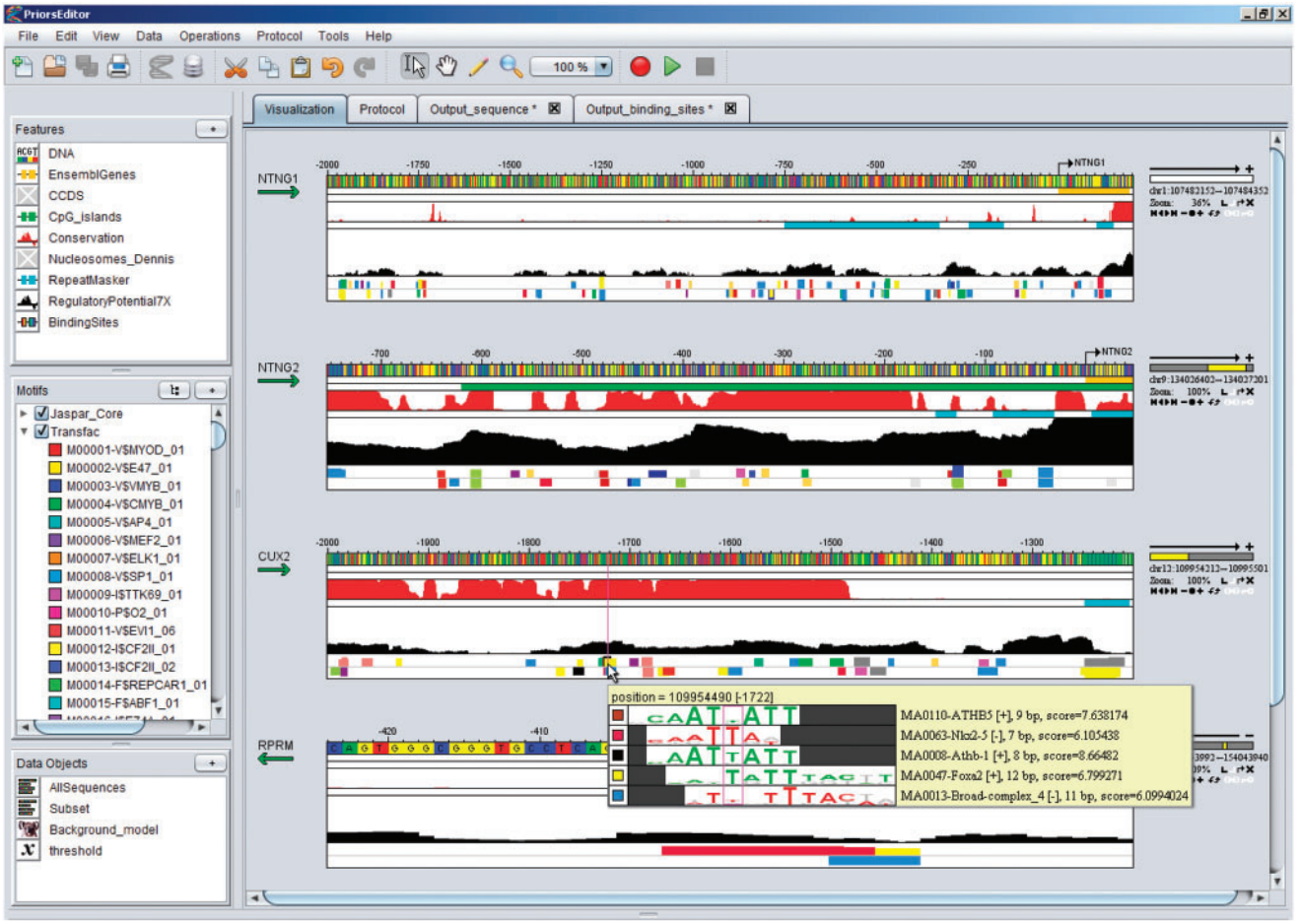

Fig. 1. The top left panel in this screenshot shows examples of some of the features that can be used as a basis to create positional priors. These features are visualized as data tracks in the main panel for a selected set of sequences. The bottom-most track contains predicted matches to TRANSFAC and JASPAR motifs in regions with non-zero RegulatoryPotential7X scores.

et al., 2007; Segal *et al.*, 2006; Whitington *et al.*, 2009), and it has also been demonstrated that further gain can be achieved by integrating information about multiple features (see e.g. Ernst *et al.*, 2010; Lähdesmäki *et al.*, 2008).

We present here a program called PriorsEditor, which allows users to easily construct positional priors tracks by combining various types of information and utilize these priors to potentially improve the motif discovery process (Fig. 1).

2 SOFTWARE DESCRIPTION

The first step in constructing a priors track with PriorsEditor is to specify the genomic coordinates for a set of sequences one wishes to analyze. Next, data for various features can be imported to annotate these genomic segments. PriorsEditor supports three types of feature data. The first type, *numeric data*, associates a numeric value with each position in the sequence and can be used to represent features such as phylogenetic conservation scores, DNA melting temperatures and nucleosome-positioning preferences. Numeric data tracks are also used to hold the final positional priors. The second feature type, *region data*, can be used to refer to continuous stretches of the DNA sequence that share some unifying properties which distinguish them from the surrounding sequence. Different regions are allowed to overlap, and regions can also be assigned values for various attributes, including type designations, score values and strand orientations. Features best represented as regions include genes, exons, repeat regions, CpG-islands and transcription factor binding sites. The last feature type, *DNA sequence data*, represents the DNA sequence itself in single-letter code. DNA sequence data can be passed on to motif discovery programs for further analysis, and it can also be used to estimate various physical properties of the DNA double helix, such as GC content, bendability and duplex-free energy. Additional feature data can be obtained from web servers such as the UCSC Genome Browser (Rhead *et al.*, 2010) or be loaded from local files.

Once the data for the desired features have been loaded, the data tracks can be manipulated, compared and combined to create a priors track using a selection of available operations. These include operations to extend regions by a number of bases upstream and/or downstream, merge overlapping regions or regions within close proximity, filter out regions, normalize data tracks, smooth numeric data with sliding window functions, interpolate sparsely sampled

data, weight numeric data tracks by a constant value or position-wise by another track, combine several numeric tracks into one using either the sum or the minimum or maximum value of all the tracks at each position and several more. It is also possible to specify conditions for the operations so that they are only applied to positions or regions that satisfy the condition. For example, to design a priors track that will focus the search toward conserved regions within close proximity of other binding sites, one could start off with a phylogenetic conservation track, then load a track containing previously verified binding sites from the ORegAnno database (Griffith *et al*., 2008), extend these sites by a number of bases on either side and lower the prior values outside these extended sites.

After a priors track has been constructed, there are several ways to make use of this new data. The most straightforward way is to provide it as input to a motif discovery program that supports such additional information, for instance, PRIORITY (Narlikar *et al*., 2006) or MEME version 4.2+ (Bailey *et al*., 2010). Unfortunately, not many motif discovery programs are able to incorporate priors directly, so an alternative is to mask sequence regions that have low priors by replacing the original base letters with Xs or Ns since most motif discovery tools will simply ignore positions containing unknown bases when searching for motifs. Apart from being used to narrow down the sequence search space, priors information can also be applied to post-process results after motif discovery has been carried out, for instance, by filtering out predicted binding sites that lie in areas with low priors or adjusting the prediction scores of these sites based on the priors they overlap.

Positional priors tracks and masked sequences can be exported for use with external tools, but it is also possible to perform motif discovery from within PriorsEditor itself by using operations to launch locally installed programs. To facilitate motif scanning, PWM collections from TRANSFAC Public (Matys *et al*., 2006) and JASPAR (Portales-Casamar *et al*., 2010) have been included, and users can also import their own PWMs or define new collections based on subsets of the available PWMs.

Constructing priors tracks and performing motif discovery analyses can be tedious, especially when it involves many datasets and requires several steps to complete. If a user discovers a good combination of features to use for priors, it may be desirable to repeat the same procedure to analyze other sequence sets as well. PriorsEditor allows such repetitive tasks to be automatized through the use of protocol scripts. Protocol scripts describe a list of operations to be performed along with any specific parameter settings that apply for these operations. They can be programmed manually in a simple command language or be constructed using a 'macro recording' function which logs all operations the user carries out while in recording mode. With protocol scripts these same series of operations can be automatically applied to new sequence sets simply by the click of a button. These scripts can also be set up so that users can provide values for certain settings during the course of an execution, enabling users to select for instance a different background model or PWM threshold value to use in the new analysis.

By providing a protocol script describing the operations to be performed along with a file specifying the target sequences, it is possible to run PriorsEditor from a command-line interface instead of starting up the normal graphical interface. This allows the construction and use of positional priors to be incorporated into a batch-processing pipeline.

Funding: The National Programme for Research in Functional Genomics in Norway (FUGE) in The Research Council of Norway.

Conflict of Interest: none declared.

REFERENCES

Bailey,T.L. *et al.* (2010) The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*, **11**, 179.

Bellora,N. *et al.* (2007) Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics*, **8**, 459.

Duret,L. and Bucher,P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7**, 399–406.

Ernst,J. *et al.* (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.

Griffith,O.L. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.

Kolbe,D. *et al.* (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.*, **14**, 700–707.

Lähdesmäki,H. *et al.* (2008) Probabilistic inference of transcription factor binding from multiple data sources. *PLoS One*, **3**, e1820.

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.* (2006) Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, **22**, e384–e392.

Portales-Casamar,E. *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.

Ravasi,T. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.

Rhead,B. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.

Sandve,G.K. and Drabløs,F. (2006) A survey of motif discovery methods in an integrated framework. *Biology Direct.*, **1**, 11.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Whitington,T. *et al.* (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.