BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 6 2012, pages 763–770
doi:10.1093/bioinformatics/bts024

Genome analysis

Advance Access publication January 16, 2012

ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites

Santi González[1†], Bàrbara Montserrat-Sentís[1†], Friman Sánchez[2], Montserrat Puiggròs[1,3], Enrique Blanco[4], Alex Ramirez[2] and David Torrents[1,5*]

[1]Joint IRB-BSC program on Computational Biology, BSC c/ Jordi Girona 29, [2]Department of Computer Architecture, Campus Nord UPC D6-117, Jordi Girona 1-3, 08034 Barcelona, [3]Computational Bioinformatics, National Institute of Bioinformatics, [4]Departament de Genètica i Institut de Biomedicina (IBUB), Universitat de Barcelona, Diagonal 643, 08028 Barcelona, Catalonia and [5]Institució Catalana de Recerca i Estudis Avançats (ICREA) Pg. Lluís Companys 23, 08010 Barcelona, Catalonia

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The prediction and annotation of the genomic regions involved in gene expression has been largely explored. Most of the energy has been devoted to the development of approaches that detect transcription start sites, leaving the identification of regulatory regions and their functional transcription factor binding sites (TFBSs) largely unexplored and with important quantitative and qualitative methodological gaps.

Results: We have developed ReLA (for REgulatory region Local Alignment tool), a unique tool optimized with the Smith–Waterman algorithm that allows local searches of conserved TFBS clusters and the detection of regulatory regions proximal to genes and enhancer regions. ReLA's performance shows specificities of 81 and 50% when tested on experimentally validated proximal regulatory regions and enhancers, respectively.

Availability: The source code of ReLA's is freely available and can be remotely used through our web server under http://www.bsc.es/cg/rela.

Contact: david.torrents@bsc.es

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

Received on September 28, 2011; revised on December 19, 2011; accepted on January 9, 2012

1 INTRODUCTION

The identification of the genomic regions that control the transcription of genes still remains a challenge despite the recent and continuous development of new experimental and computational methodologies (Tompa *et al.*, 2005). Multiple automatic approaches have been proposed, ranging from those that search for phylogenetic conservation of sequence or transcription factor binding motifs in non-transcribed DNA regions (Blanchette and Tompa, 2003; Van Loo and Marynen, 2009) to the analysis of DNA physical properties characteristic of regions expected to bind transcription factors (Abeel *et al.*, 2008; Goñi *et al.*, 2007). However, the incorporation of novel biological knowledge into these programs is not necessarily improving the quality of their predictions, which still contain a substantial fraction of false positives.

Currently, methods that *de novo* detect and characterize proximal regulatory regions show specificity levels <50% (Van Loo and Marynen, 2009). Even though phylogenetic footprinting using pre-aligned homologous regulatory regions offers promising results in the identification of Conserved Regulatory Modules (CRMs) of transcription factor binding sites (TFBSs) (Blanchette and Tompa, 2003; Blanco *et al.*, 2007; Pavesi *et al.*, 2007; Sebestyen *et al.*, 2009; Tokovenko *et al.*, 2009; Tonon *et al.*, 2010), they cannot define the regulatory region itself in most real scenarios because they are based on global alignment strategies and require that all matching binding sites across all compared sequences are located in the same (or similar) position within each sequence, i.e. they require predefined and pre-aligned regulatory regions. As a result, in spite of the existing methodologies, still the most common and reliable way to identify proximal regulatory regions in genomes is the analysis of a few nucleotides (typically up to 1000) immediately upstream of annotated transcription start sites (TSSs), which likely constitutes the proximal promoter. But the annotation of gene starts is still unsolved, particularly for non-human species. For example, a simple search in the ENSEMBL database (Hubbard *et al.*, 2009) identified substantial fractions of vertebrate genes without annotated 5′UTR (from 17% in mouse to 91% in opossum, 42% for human). This result is even more dramatic within invertebrates. Other problems that constitute a barrier for the automatic inference of promoters (even in human or mouse) are the abundance and overprediction of alternative transcripts. Taken together, most computational methods that detect or align promoters strongly depend on or are coupled with the annotation of untranslated gene regions, which is generally insufficient for this purpose (Guigo *et al.*, 2006).

On the other hand, the computational identification of enhancers is even more complex. These regulatory regions that work in cooperation with promoters throughout multiple structural constraints are, apparently, delocalized relative to the genes that are controlling (Arnosti and Kulkarni, 2005). Therefore, their identification through computational methods requires strategies

*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 Authors.

based on local alignments. Some existing methods, like rVISTA, look for conserved TFBS clusters between regions that have been pre-aligned with local alignment tools, such as BLASTz in rVISTA (Loots and Ovcharenko, 2004), while others use directly local alignment-based search strategies, like the Enhancer Element Locator (EEL) that uses the Smith–Waterman algorithm (Palin *et al.*, 2006). These tools have shown good prediction rates on enhancers, and also show important limitations regarding the number of species that they can analyse, the required parametrization and the accuracy of the prediction.

To overcome these limitations, and to provide novel and improved solutions to the prediction of regulatory regions, we have developed ReLA, a public local-based alignment tool that is capable of detecting promoters and enhancers by identifying clusters of regulatory motifs conserved in any position within large homologous DNA regions (i.e. independently of gene annotation). Considering the wide range of potential users of this tool, we have also developed a user-friendly web server for remote predictions. The source code of ReLA is distributed also as a standalone program that can be used locally in Unix-based computational platforms.

2 MATERIALS AND METHODS

2.1 From DNA to TFBS sequences

The first step of our method consists on the mapping of putative TFBSs sequences along the input sequences according to a certain catalogue of position frequency matrices (PFMs) (in the documentation included with the program and in the web server, we provide guidance for selecting a set of homologous sequences). Users locally running ReLA should provide their own PFM files (information about accepted file formats is also provided with the program and in the web server). It has been shown that the selection of particular collections of matrices yields slightly different results (Blanco *et al.*, 2006b). After evaluating different options (data not shown), we obtained the optimal results by using PFMs from TRANSFAC (Matys *et al.*, 2006). We classified this collection of models into three subsets: whole collection of TRANSFAC PFM, the first 600 and the first 400 PFMs ranked by their information content. These three sets are included as the default option in the web server. The identification of potential TFBSs is performed using the MATSCAN software (Blanco *et al.*, 2006b) at high levels of stringency: 80% of similarity threshold. For this study, we calculate the similarity score using SS = [(current – min)/(max – min)], where, 'current' is the actual matching score, 'min' is the minimum possible matching score and 'max', the maximum possible matching score of a particular PFM, as described elsewhere (Kel *et al.*, 2003). Since the next Smith–Waterman step requires a single sequence of TFBSs, and because MATSCAN results usually contain PFM that overlap in all possible ways, we next simplify this output. We remove this overlap by sliding a window of n bp ($n = 3$ or 5, both conditions are included in the global search, see below). For each overlapping PFM starting at the first position of each of the 3 or 5 nt sliding window, we systematically kept the most informative one, maximizing the overall information content of the sequence. To preserve the relative distance between motifs during the comparison, we insert a 'spacer box' every n consecutive nucleotides in regions free of predictions. In summary, we convert each input DNA sequence into a sequence of highly informative non-overlapping TFBS, which is used next in the comparative searches.

2.2 Comparative searches

For our searches, we have modified the classical local alignment Smith–Waterman algorithm (Smith and Waterman, 1981) to deal with sequences of TFBSs (associating a unique three-letter combination to each TFBS) and to provide the best scoring local alignment (i.e with the highest density of conserved TFBSs) for each of the comparisons performed between the reference sequence with each of the others (see pseudocode in Supplementary Material). The overall stringency of the searches, the reliability of the resulting predictions and the conservation of the TFBSs between the input sequences can produce different predictions. Instead of selecting a universal and fixed set of parameters for each of the searches, which would yield one unique prediction, we chose to run recursively each pairwise comparison (reference sequence against each of the other input sequences) with a different set of parameters generating a collection of preliminary predictions. A set of posterior selection filters (see below) is then applied on these preliminary predictions to come up with a final prediction of the promoter region.

Each of these pair-wise comparisons is carried out in two different scoring scenarios (10/−1 and 20/−1 match/mismatch scores, both with an open and extension gap penalties of −2 and with two overlapping thresholds to remove redundant sites (using a window of three or five nucleotides, see above); i.e. a total of four comparisons are performed on each pair of sequences and each set of matrices defined. These specific combinations of parameters were determined by monitoring and maximizing the relationship between sensitivity and specificity using a collection of 10 known promoters of the ABS database (Blanco *et al.*, 2006a) (see Supplementary Material and www.bsc.es/cg/rela/downloads). These 10 regions were excluded from the performance evaluation. We also observed that the best predictions obtained during the benchmarking were those with sizes between 200 and 600 nt. Preliminary predictions covering shorter regions usually involved too few conserved TFBSs, while larger predictions tend to connect distant and, apparently, unrelated binding sites. For this reason, preliminary predictions outside this range of sizes are not considered during the generation of the final prediction.

2.3 Output generation

As part of the results, ReLA generates a graph showing the distribution of all accepted preliminary predictions on the reference sequence. From the analysis of the overlap among these preliminary predictions, we generate the final prediction by selecting the region, between 200 and 1000 nt long that contains the highest number of preliminary candidate predictions. Together with the final prediction on the reference sequence, additional consensus regions are also defined in each of the other sequences, which correspond to those (if any) that match the final predicted promoter region. We also provide the list of all conserved TFBSs. From this list, a subset of the most conserved TFBSs (specifically, those within the top 10% of conservation) is selected and used to generate a multiple alignment in graphical format.

2.4 Web server

We have implemented ReLA as a web server that can be accessed at http://www.bsc.es/cg/rela. The underlying search engine is distributed also as a standalone program. We have designed the ReLA website to meet the requirements of non-expert users. The web version provides a graphical representation of the putative promoter region predicted in all the input sequences (Fig. 5) and a plain text description of the results. As many as two suboptimal solutions can be provided through the web on each set of input sequences to potentially predict alternative promoters.

2.5 Selection of databases for evaluation

To validate the results obtained and to compare our method with other existing programs in similar searching conditions, three different working subsets or reported regulatory regions were generated from three different databases: Annotated regulatory Binding Sites database [ABS; (Blanco *et al.*, 2006a)], Eukaryotic Promoter Database [EPD; (Schmid *et al.*, 2006)] and Vista Enhancer Database Browser [VISTA; (Visel *et al.*, 2007)] To follow common criteria and to be consistent with the annotation of ABS (as 500 nt promoters), we transformed these TSS into regions by considering as promoter region 500 bp upstream from the EPD TSS. VISTA Enhancer

Browser is a database of regions containing experimentally validated human and mouse enhancers tested in transgenic mice.

The working subsets were generated according to three different filters to facilitate the automation of the validation process and to ensure reliability of the evaluation protocol: (i) genes associated to selected regulatory regions must have at least three orthologous one-to-one genes according to ENSEMBL orthology data; (ii) the promoter fragments selected should not overlap with coding regions; and (iii) they have to be in our scanned region: as described in the Section 3, for ABS and EPD it is 500 bp upstream of the first codon of the gene, and for Vista, 50000 bp upstream of the closest gene (see below).

Applying these three filters we obtained 75 human and mouse promoters from ABS, and 740 from EPD. In both cases, 5000 bp upstream from the first methionine were used to check and compare the accuracy of the method. From VISTA Enhancer Browser, we ended up with 44 enhancers laying in the 5000 bp upstream of a known gene.

2.6 Evaluation protocol

For the evaluation of the promoter prediction programs, we followed a modified version of the Distance-based validation evaluation protocol from Abeel *et al.* (2009). Taking into account that we were evaluating promoter genes and we were considering distances, we calculated recall and precision values as

$$\text{Precision} = \text{correct predictions}/\text{total predictions}$$

$$\text{Recall} = \text{discovered genes}/\text{total genes}$$

For those programs that provide single positions as outputs (TSS predictors, ARTS, Eponine and Promoter Explorer), we considered the sequence ± 500 from TSS for the evaluation. In the case of TFM, we obtained conserved binding sites as result and considered the fragment between the first and the last one for evaluation. A correct prediction was considered if there was an overlap between the prediction and the 500 bp upstream of the defined TSS. For all programs, we considered the unique or the best prediction, except for Promoter Explorer that does not rank the results. Since we were using already filtered promoter genes instead of big DNA fragments, we did not discarded any prediction further of 500 nt from the TSS as it is done elsewhere (Abeel *et al.*, 2009). For all programs of our evaluation, we used default settings defined by the corresponding developers. For EEL runs, we used the mouse and human sequences for each of the orthologous groups and applied the parameters described for this species pair elsewhere (Palin *et al.*, 2006). All the data used for the validation procedure are available at www.bsc.es/cg/rela/downloads.

3 RESULTS

3.1 Rationale and underlying search strategy of ReLA

The goal of this study is to develop a novel methodology that would overcome the current limitations mentioned above by focusing on: (i) the detection of conserved TFBS; (ii) the use of local search strategies; (iii) simplicity of use; and (iv) a low computational cost to perform genome-wide searches. For this, we decided to use the same strategies that have been used for fast local sequence comparisons of protein sequences. In particular, we adapted the Smith–Waterman algorithm (Smith and Waterman, 1981) to make it capable of comparing and detecting the best optimal local alignment of regions with similar sequences of TFBSs. In our procedure, each TFBS is internally transformed into symbols of an arbitrary alphabet, as if they were amino acid or nucleotides in traditional protein–protein and DNA–DNA comparative searches. This search algorithm is the core of a pipeline, referred from now on as ReLA (for REgulatory region Local Alignment tool). The complete procedure can be divided into three major steps. First, input DNA sequences are transformed into sequences of TFBSs by mapping with the MATSCAN software (Blanco *et al.*, 2006b) all the PFMs provided; secondly, the resulting TFBS sequences are compared with each other to identify conserved groups of TFBSs using the modified Smith–Waterman algorithm under different scoring scenarios. Finally, all the resulting preliminary alignments are evaluated to produce the final prediction of the promoter region.

3.2 Evaluation of ReLA's performance

We first applied ReLA to a collection of experimentally validated promoter and enhancer regions, both to define its internal search parameters and to evaluate its performance. Despite ongoing efforts of acquiring experimental data, on functional TFBS, still the vast majority of detailed and reliable data can only be retrieved from the literature. In this direction, the ABS database (Blanco *et al.*, 2006a) is the result of one of the few initiatives to gather, from the literature, promoters with two or more of their TFBSs experimentally validated. For this reason we used this database for ReLA's evaluation. We selected the subset of 73 (35 human and 38 mouse) promoters from this database that showed, in ENSEMBL (Hubbard *et al.*, 2009), one-to-one orthologous relationship with at least three out of seven chosen vertebrate species (human or mouse, rat, horse, dog, cow, opossum and chicken). By reproducing a common and realistic search scenario, where the TSS and 5′UTRs of query homologous regions are not known, we collected the putative upstream region of these genes and their corresponding orthologous regions. These upstream regions comprise 5000 bp upstream DNA, from the first annotated codon according to the encoded ENSEMBL protein. From the measurement of the length of 5′UTRs regions of 'known' ENSEMBL genes, we previously had estimated that this selection of 5000 bp is sufficient to capture the proximal promoters for $>85\%$ of known vertebrate genes (data not shown). In addition, we have also compared the resulting performance of ReLA with the prediction ratio of other reported TFBS-based search tools: TFM (Tonon *et al.*, 2010) and rVISTA (Loots and Ovcharenko, 2004), as well as with that of TSS predictors: ARTS (Sonnenburg *et al.*, 2006); Eponine (Down and Hubbard, 2002) and PromoterExplorer (Xie *et al.*, 2006), all of them run on the same regions (Table 1).

Table 1. Performance results on ABS promoters

	ReLA	TFM	rVISTA[(1)]	PromoterExplorer	Eponine	ARTS
Recall	0.81	0.6	0.37	0.51	0.2	0.14
Precision	0.81	0.61	0.46	0.69	0.21	0.14
Prediction type	Defined regions with conserved TFBS		Conserved TFBS		TSS	

Table 2. Performance results on EPD TSSs

	ReLA	TFM	PromoterExplorer	Eponine	ARTS
Recall	0.56	0.49	0.78	0.23	0.17
Precision	0.56	0.51	0.67	0.27	0.17

Following the same strategy, we alternatively evaluated ReLA using 740 regions derived from the Eukaryotic Promoter Database [EPD; (Schmid *et al.*, 2006)], which, despite not being ideal for this purpose, sets our tool into the context of previous evaluations of these other existing search strategies: ARTS, Eponine and PromoterExplorer against which we have also compared ReLA's predictions (Table 2).

From all resulting predictions, we calculated different performance parameters, such as recall and precision by adapting an evaluation protocol used for the comparison of a large number of TSSs predicting methods (Abeel *et al.*, 2009). This adaptation is necessary because the different methods we used provide different type of outputs, e.g. ARTS, Eponine, PromoterExplorer provide single TSS positions, TFM and rVISTA lists of conserved TFBS, and ReLA delimited regions of conserved TFBSs.

From the results of this evaluation, we observe that the overall performance is different between the different methods and databases: while ReLA's performance was the best on ABS entries, PromoterExplorer on EPD regions outperformed it. Interestingly, ReLA's precision values for EPD regions are lower than those shown with ABS. To discard a possible bias in the performance of ReLA towards specific promoter types that are more abundant in the ABS database, we divided all EPD and ABS regions in different promoter classes (Section 2) and calculated the same precision values for each promoter type and each prediction method separately (Supplementary Tables S1 and S2). This analysis has shown that, despite ABS appears to be enriched in TATA-box containing promoters (a 42% versus a 20% in EPD), ReLA's performance is not affected by this, as precision values are similar among most of the promoter types present in ABS and EPD. It is also worth noticing that predictors based on TFBSs show a better performance on the ABS, which is also based on TFBS, than with the TSS-based EPD entries, where TSS predictors tend to do better. We did not find any significant difference when comparing performance values for Human or Mouse entries (data not shown).

In order to have a sense of the TFBS conservation levels, upon which ReLA is able to build predictions, we have also analysed the distribution of the number of conserved boxes detected within all ABS and EPD results. This analysis shows that, indeed, there is a wide range of TFBS conservation, both in number and in composition (Supplementary Fig. S1). Similarly, from the analysis of the contribution of each of the species in ReLA's performance, we observe that all the other vertebrates used in this study contribute substantially to the final prediction in human: for example, cow and dog contribute to ∼60% of the predictions whereas opossum and chicken to 36 and 32%, respectively (Supplementary Table S3).

A detailed inspection of ReLA's results on ABS entries uncovered some interesting features. Often, the promoters that we identified on each of the species present different locations within the input 5000 bp region (Fig. 1). This typical scenario, which must be necessarily solved with local-based comparative approaches, is

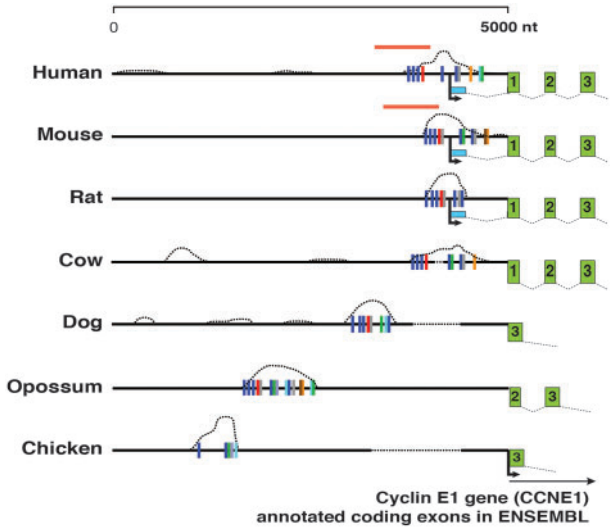

Fig. 1. Prediction of the proximal regulatory region of the Cyclin E1 (*CCNE1*) gene in seven vertebrate species. Typical search scenario where the TSS for most of the species compared is not known or missannotated: no TSS information was available for cow, dog and opossum, whereas in chicken is wrongly placed. The predicted promoter for these species appears in different location within the input sequence. Dashed lines show the distribution of all primary predictions along these regions. Consensus predictions are delimited by the first and the last coloured box (each box corresponds to a conserved TFBS). Red horizontal lines indicate the experimentally characterized regulatory region. Initial fragments of each transcript are shown on the right: non-coding regions in blue, and coding exons in green. The numbers indicate the position of the coding exons in the human mRNA. ENSEMBL TSSs are indicated with arrows.

observed when the annotation of orthologous gene 5′UTRs and first exons is practically absent, as occurs for most of the available genomes. These results highlight the potential of using ReLA for the systematic identification and annotation of regulatory regions in non-model organisms, such as chicken, cow, dog, opossum and any other that has incomplete gene and cDNA data.

3.3 Prediction of multiple alternative promoters

During the evaluation of ReLA, we also observed that, in some cases, the distribution of preliminary predictions along the reference sequence highlighted two regions with similar scores, which could correspond to alternative promoters. From these two options, ReLA selects the one that generated more preliminary predictions (Section 2). Suboptimal solutions, i.e. potential alternative promoters, can be obtained by simply masking the previously predicted regions in the reference sequence and running ReLA again. For a number of such cases, we confirmed the presence of two TSSs through the analysis of ESTs or known alternatively transcribed full-length mRNAs. For this reason, we have implemented this option in the web server, where the user can launch a second search run to find suboptimal solutions. Figure 2 shows the best two predictions of regulatory regions located upstream from *SLC7A7*, an amino acid transporter gene, which has been experimentally proven to have two alternative promoters (Puomila *et al.*, 2007).

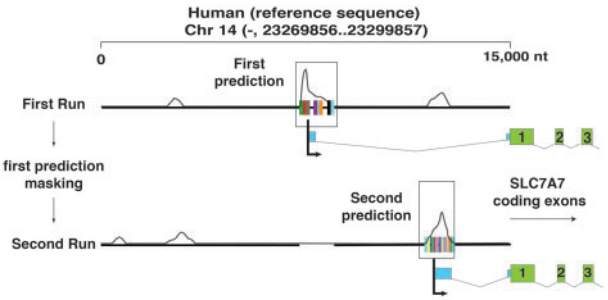

Fig. 2. Prediction of alternative known promoter regions of the solute carrier family 7 member 7 (SLC7A7). Searches were done on 15000 bp region upstream of the first amino acid annotated in the ENSEMBL database for the human SLC7A7 gene. Dashed lines show the distribution of preliminary predictions in the first and second run. Final first and second predictions are enclosed in a box and delimited by the first and the last conserved TFBS (each designed by a coloured box). Initial fragments of each transcript are shown on the right: non-coding regions in blue, and coding exons in green. The numbers indicate the position of the coding exons in the mRNA. ENSEMBL TSSs are indicated with arrows. Distances are not drawn at real scale.

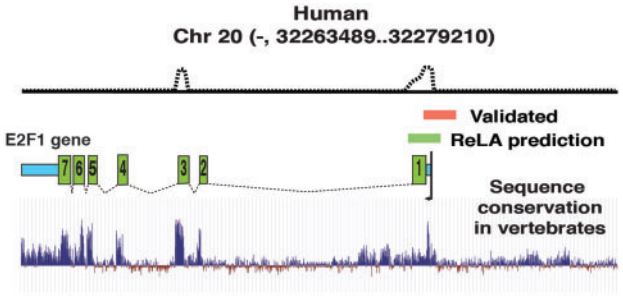

Fig. 3. Prediction of the proximal regulatory region of the E2F transcription factor 1 (E2F1) using the sequence of the whole gene. Predicted regulatory region along a highly conserved region of 15 722 bp that includes the E2F1 complete gene transcript and 5000 bp upstream of the first amino acid annotated in ENSEMBL. Dashed line shows the distribution of all preliminary predictions along this region. A schematic representation of the structure of this gene is shown: non-coding regions are in blue, and coding exons in green. The numbers designate the position of the coding exons in the mRNA, according to human. Data related to DNA conservation from UCSC are also included [http://genome.ucsc.edu; (Kent *et al.*, 2002)].

The finding of two high scoring regions in any ReLA prediction could suggest, instead of the presence of an alternative promoter, the existence of highly conserved coding exons, which would constitute a false positive prediction. Thus, the identification of regulatory regions with ReLA would be based only on the level of sequence conservation and the presence of highly conserved non-regulatory DNA, like coding exons, could negatively influence the results. To discard this, we studied how this scenario can affect ReLA predictions. The example in Figure 3 shows a positive promoter prediction when all seven orthologous input sequences include the complete region of the *E2F1* gene and the additional upstream untranslated regions (a total of 15000 nt each). In this case, the distribution of hits along the human sequence shows two high scoring regions that appear to share similar conservation levels of nucleotides. One of these fragments constitutes the third exon of this gene, while the other matches the 5′UTR, the TSS and the core promoter. ReLA is able to successfully discriminate the correct promoter region, including sites that have been experimentally proven (Blanco *et al.*, 2006a). In particular, the two TFBS that ReLA scores highest in conservation among input sequences are precisely described in the ABS database as two E2F1 factor binding sites necessary for self-regulation during the transition from G1 to S phase in the cell cycle (Johnson *et al.*, 1994). Interestingly, the third TFBS following the conservation ranking corresponds to ADF1, which was located within the 5′UTR and is known to bind the same motifs recognized by the E2F1 factor in mice (Hsiao *et al.*, 1994) Despite these results, we cannot discard the possibility that exons are wrongly predicted as promoters in certain situations. Therefore, we recommend performing preliminary evaluations of the coding potential within the input sequences, for example, by comparing them against protein sequence databases with BLASTX (Altschul *et al.*, 1997). Putative coding regions should be preferentially masked from the input sequences to ensure the correct prediction.

3.4 Identification of enhancers

The local nature of the underlying search engine and the capacity to compare large DNA sequences makes ReLA a suitable tool for the identification of enhancers, which are often located distant from other functional elements. In order to test ReLA's capabilities in enhancer detection we have gathered a collection of experimentally validated human enhancers from the VISTA database with activity assessed on transgenic mice (Visel *et al.*, 2007). To test ReLA, we selected 44 enhancers that are located within the first 50000 bp upstream of a known gene. In order to search for each of these distal regulatory regions we extracted up to 50000 bp from the most upstream TSS annotated for the closest gene in human and from each of the corresponding one-to-one orthologous genes in mouse, rat, horse, dog, cow, opossum and chicken. The first run of ReLA on these 44 regions generated 40 predictions, from which 11 (28%) overlapped with the annotated enhancer. Considering that the regions selected for the search theoretically contain other unknown regulatory regions (promoters, for instance) that could match with the first prediction, we performed a second run on the remaining 29 cases, which yielded nine other positive hits. In total, with two iterative runs, ReLA showed a positive predictive value of 50% of the screened subset of annotated human enhancers. A similar prediction rate (49%) is obtained over the same enhancer benchmark set when using a specific enhancer locator tool, EEL (Palin *et al.*, 2006) that also relies on local search strategies (EEL searches implied only human and mouse sequences, as it does not accept more than two sequences per search). It is worth mentioning that an important difference between both methods is that ReLA provides more precise results, as the regions predicted are shorter (up to 750 nt long, with an average of 485 nt) than those coming from EEL (up to 11563 nt, with an average of 2644 nt).

These results indicate that ReLA is capable of searching large genomic DNA fragments and identifying multiple proximal and distal regulatory regions, which makes this tool suitable for genome-wide screenings and across several genomes (see an example in Fig. 4).

To further exemplify this feature, we also performed a genome-wide analysis on a 109 kb long ENCODE region [ENm011; chr11:1858751-1968592; (Birney *et al.*, 2007)] that includes six genes coding for, at least, 11 transcripts, with their corresponding intergenic regions. By using SYT8 and MRPL23 flaking genes

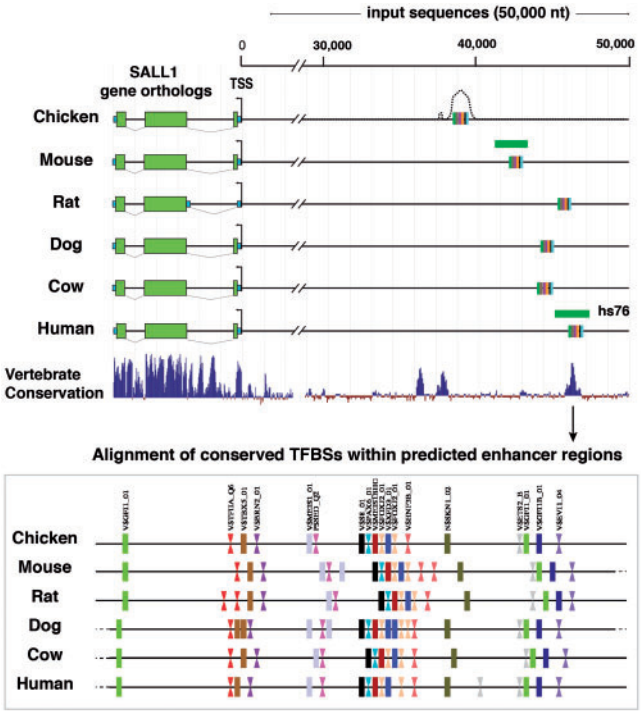

Fig. 4. Prediction of the SALL1 enhancer in six vertebrate genomes. The upper panel shows the SALL1 gene and its corresponding 50 kb upstream region for six vertebrate species. Coordinates and strand of these regions are: Chicken (chr11: 6148098–6213605, −), Mouse (chr8: 91551143–91618061, +), Rat (chr19: 19227337–19293298, −), Dog (chr2: 67080056–67144522, −), Cow (chr18: 18688671–1875278, +) and Human (chr16: 51169886–51235181, +). In each line, we display the structure of the gene (green boxes are coding exons, while blue are untranslated). Known and predicted TSSs are also shown. ReLA's predictions are shown for each species as groups of colored boxes. Please, note that these regions are not drown to scale. ReLA's predicted enhancer regions expanded from 223 and 233 bp for dog and mouse, 301 bp for cow, to 359, 360 and 366 bp for rat, human and chicken, respectively. The locations of the experimentally proven regions (as shown in rVISTA db) are displayed as green boxes. The bottom line of this panel shows the sequence conservation profile (according to human coordinates; http://genome.ucsc.edu). In the bottom panel, we display the alignment of the conserved TFBS detected within each of the predicted enhancer regions. TFBS are labelled (in TRANSFAC format) and differentiated using arbitrary shapes and colours. Coordinates shown here indicate the position of the predicted enhancer within the 50 kb input sequence.

as anchors, we identified and characterized the corresponding orthologous regions for mouse, rat, dog, cow and chicken. The complete analysis consisted in 10 iterative ReLA searches and implied the screening of TFBS sequences in >600 kb of genomic DNA. In order to obtain an estimation of the performance on these genome-wide conditions, we have taken as positive predictions those that match ChIP-Seq transcription evidences (Birney *et al.*, 2007), as well as those falling immediately upstream of annotated gene starts. This count shows that 8 out of 10 predictions have evidence of expression or regulation (Supplementary Fig. S2). Please note that we cannot discard that additional runs would provide other overlooked regulatory regions and, at some point, also false positives.

4 DISCUSSION

Taking into account the available methods to *in silico* recognize gene regulatory regions, a substantial improvement is necessary to accurately annotate genes and promoter sequences in most genomes. Here we report the development of ReLA, a computational tool to identify such regulatory regions using genome-wide comparisons. ReLA is distributed as a standalone program and as a web server. Our approach is mostly based in an adaptation of the popular Smith–Waterman algorithm that is able to rapidly identify coincidences of TFBSs between two sequences (conceptually similar to traditional protein–protein comparisons). ReLA is able to efficiently process long sequences in standard computational platforms (e.g. less than a minute to obtain the results shown in Fig. 5). We have evaluated the accuracy of ReLA, first in a dataset of experimentally validated human and mouse promoters, on an extensive collection of validated TSS from the Eukaryotic Promoter Database, as well as on an experimentally validated collection of rVISTA enhancers. We have reached maximums of 0.81 of recall and precision levels on ABS sequences On the other hand, and surprisingly, ReLA's performance results lower when using EPD TSS entries. A possible explanation for this observation could be that ReLA performs better on certain types of promoter regions. But, after we classified all ABS and EPD entries into different promoter types according to their composition and evaluated their associated performance obtained with all the methods used here for the validation, we observed that ReLA's accuracy is similar among most of the identified promoter types. We cannot discard though that other uncontrolled biases present either inside the underlying search methodology of each of the protocols used here or in the used databases could actually explain the different behaviour observed. It is worth noticing that overall, TFBS-based prediction methods perform better on the TFBS-based ABS database than on the TSS-based EPD, where TSS predictors are doing better. In any case, the levels of precision and recall obtained with ReLA are sufficient to provide reliable predictions that guide posterior experimental validation. This study also demonstrates the benefits of using the Smith–Waterman algorithm to directly search for conserved binding sites, as it outperforms other methods like rVISTA and TFM, that are based on pre-aligned DNA and global search strategies. See the example in Figure 1, which cannot be solved using global alignment approaches. Please note that other methods based on similar strategies could not be included in the comparison, as they did not provide results on our benchmark set because of limitations in the size (MMETA) and on the number of sequences [Conreal (Berezikov *et al.*, 2005)].

Furthermore, we also show that ReLA is able of predicting alternative promoters and even enhancer regions, dealing with multiple suboptimal solutions in most cases. Our approach is suitable for integrating a computational annotation pipeline in which other predictive methods such as homology searches (e.g. BLAST against protein databases) can assist in the improvement of the final predictions.

In summary, we believe that the development of ReLA constitutes a significant step forward in the field of the prediction of regulatory regions, as it shows the highest predictive power reported so far. ReLA is able to locally compare multiple large genomic regions and identify non-alignable conservation events across different genomes. This is relevant if we consider that the limited information regarding regulatory regions in eukaryotes is restricted to human

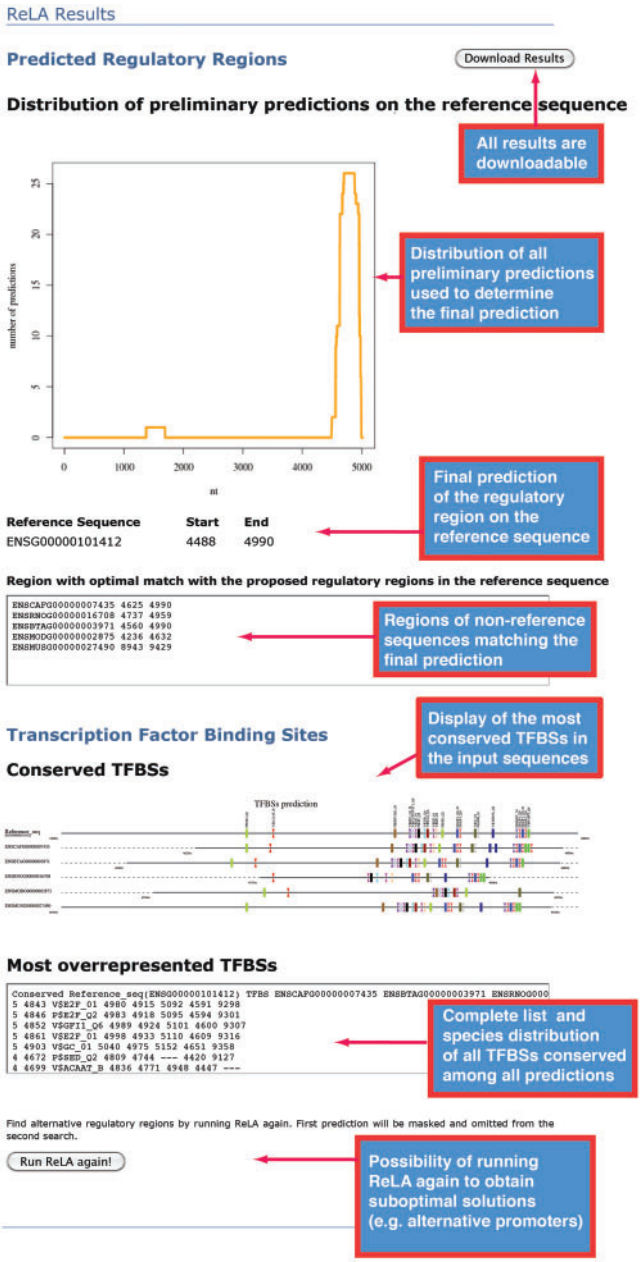

Fig. 5. Snapshot of the ReLA web server output. Graphical representation of the putative promoters and alignments of TFBSs of the human E2F1 gene, as well as the lists of predicted regions and conserved binding motifs. See Sections 2 and 3 for a complete interpretation of each of the results provided.

and mouse, e.g. from 2540 vertebrate entries in the Eukaryotic Promoter Database (Schmid *et al.*, 2006), 2067 (81%) belong to these two species. Thus, with this tool in hand we can now, not only fill missing gaps in the annotation of the genomes of model organisms, mostly with the identification of enhancers and alternative promoters, but also to start a reliable and consistent annotation of conserved promoters throughout the rest of genomes that have little or no information regarding 5′UTRs and often first coding exons (Fig. 1). Beyond the current performance of ReLA and, as we are planning a genome-wide search of regulatory regions across sequenced vertebrate genomes, we are actively searching for ways of improving further its predictive power by, for example, applying more sophisticated scoring systems and accepting even larger DNA regions with low additional computational costs.

ACKNOWLEDGEMENTS

We thank Mar Albà, Steven Laurie and our entire group for constructive feedback during the development of this work and during the writing of the manuscript. We also thank Jan and Aina Sagristà for designing ReLA's logo.

Funding: Ministerio Español de Ciencia e Innovación (BIO2006-15036).

Conflict of Interest: none declared.

REFERENCES

Abeel,T. *et al.* (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.
Abeel,T. *et al.* (2009) Toward a gold standard for promoter prediction evaluation. *Bioinformatics*, **25**, i313–i320.
Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
Arnosti,D.N. and Kulkarni,M.M. (2005) Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.*, **94**, 890–898.
Berezikov,E. *et al.* (2005) CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res.*, **33**, W447–W450.
Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
Blanchette,M. and Tompa,M. (2003) FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.
Blanco,E. *et al.* (2006a) ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. *Nucleic Acids Res.*, **34**, D63–D67.
Blanco,E. *et al.* (2006b) Transcription factor map alignment of promoter regions. *PLoS Comput. Biol.*, **2**, e49.
Blanco,E. *et al.* (2007) Multiple non-collinear TF-map alignments of promoter regions. *BMC Bioinformatics*, **8**, 138.
Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
Goñi,J.R. *et al.* (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.
Guigo,R. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, **7** (Suppl. 1), S2 1–31.
Hsiao,K.M. *et al.* (1994) Multiple DNA elements are required for the growth regulation of the mouse E2F1 promoter. *Genes Dev.*, **8**, 1526–1537.
Hubbard,T.J. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
Johnson,D.G. *et al.* (1994) Autoregulatory control of E2F1 expression in response to positive and negative regulators of cell cycle progression. *Genes Dev.*, **8**, 1514–1525.
Kel,A.E. *et al.* (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
Palin,K. *et al.* (2006) Locating potential enhancer elements by comparative genomics using the EEL software. *Nat. Protocols*, **1**, 368–374.
Pavesi,G. *et al.* (2007) WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics*, **8**, 46.
Puomila,K. *et al.* (2007) Two alternative promoters regulate the expression of lysinuric protein intolerance gene SLC7A7. *Mol. Genet. Metab.*, **90**, 298–306.
Schmid,C.D. *et al.* (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.*, **34**, D82–D85.
Sebestyen,E. *et al.* (2009) DoOPSearch: a web-based tool for finding and analysing common conserved motifs in the promoter regions of different chordate and plant genes. *BMC Bioinformatics*, **10** (Suppl. 6), S6.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Sonnenburg,S. *et al.* (2006) ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, **22**, e472–e480.

Tokovenko,B. *et al.* (2009) COTRASIF: conservation-aided transcription-factor-binding site finder. *Nucleic Acids Res.*, **37**, e49.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Tonon,L. *et al.* (2010) TFM-Explorer: mining cis-regulatory regions in genomes. *Nucleic Acids Res.*, **38** (Web Server Issue), W286–W292.

Van Loo,P. and Marynen,P. (2009) Computational methods for the detection of cis-regulatory modules. *Brief. Bioinform.*, **10**, 509–524.

Visel,A. *et al.* (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.

Xie,X. *et al.* (2006) PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. *Bioinformatics*, **22**, 2722–2728.