Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W217–W221***
DOI: 10.1093/nar/gkh383

rVISTA 2.0: evolutionary analysis of transcription factor binding sites

Gabriela G. Loots and Ivan Ovcharenko[1],*

Genome Biology Division and [1]EEBI Computing Division, Lawrence Livermore National Laboratory,
7000 East Avenue, L-441, Livermore, CA 94550, USA

Received February 10, 2004; Revised March 11, 2004; Accepted March 24, 2004

ABSTRACT

Identifying and characterizing the transcription factor binding site (TFBS) patterns of *cis*-regulatory elements represents a challenge, but holds promise to reveal the regulatory language the genome uses to dictate transcriptional dynamics. Several studies have demonstrated that regulatory modules are under positive selection and, therefore, are often conserved between related species. Using this evolutionary principle, we have created a comparative tool, rVISTA, for analyzing the regulatory potential of noncoding sequences. Our ability to experimentally identify functional noncoding sequences is extremely limited, therefore, rVISTA attempts to fill this great gap in genomic analysis by offering a powerful approach for eliminating TFBSs least likely to be biologically relevant. The rVISTA tool combines TFBS predictions, sequence comparisons and cluster analysis to identify noncoding DNA regions that are evolutionarily conserved and present in a specific configuration within genomic sequences. Here, we present the newly developed version 2.0 of the rVISTA tool, which can process alignments generated by both the zPicture and blastz alignment programs or use pre-computed pairwise alignments of several vertebrate genomes available from the ECR Browser and GALA database. The rVISTA web server is closely interconnected with the TRANSFAC database, allowing users to either search for matrices present in the TRANSFAC library collection or search for user-defined consensus sequences. The rVISTA tool is publicly available at http://rvista.dcode.org/.

INTRODUCTION

Unlike most prokaryotic genomes, which are composed of tightly packed gene units with limited intergenic regions, eukaryotic genomes are rich in noncoding sequences of unknown functions. Extensive annotation of the human and mouse genomes has predicted in the vicinity of ~40 000 genes, exons of which account for <5% of the genome. An additional 40–45% of the mammalian genome is comprised of repetitive DNA, while the remaining 50% is noncoding in nature (1,2). First glimpses of the human genome have revealed very few insights regarding new RNA coding genes, transcriptional regulatory elements or any other biologically relevant sequences present in noncoding regions. Although some parts of the noncoding genome will likely demonstrate no measurable biological function, it is widely assumed that much of our genetic complexity is due to sophisticated regulatory noncoding signals that determine when, where and to what extent each gene displays transcriptional activity. Despite the importance of noncoding sequences in gene regulation, our ability to computationally identify and characterize these elements is very limited.

In multi-cellular organisms, modulation of gene expression is accomplished through the compound interaction of regulatory proteins (transcription factors, TFs) and the specific DNA regions (*cis*-regulatory sequences or modules) with which they physically interact. Numerous DNA footprinting and biochemical studies carried out over the last decade have identified close to 500 vertebrate-specific TFs, and information regarding the DNA sequences they recognize. The TRANSFAC database (http://www.biobase.de) (3,4) represents the most comprehensive collection of TF binding specificities, summarized as position weight matrices (PWMs). A major limitation in using PWMs to computationally identify functional transcription factor binding sites (TFBSs) is that many TFs bind to short degenerate sequence motifs (6–12 bp in length). Such sequences occur very frequently in a genome, and experimentally it has been shown that only a very small fraction of these predicted TFBSs are functionally relevant.

We have previously shown that the rVista tool combines pattern recognition with comparative sequence analysis to dramatically reduce the number of false positive TFBS matches and enrich for functional sites (5). These results suggest an alternative strategy for sequence-based discovery of

*To whom correspondence should be addressed. Tel: +1 925 422 5035; Fax: +1 925 422 2099; Email: ovcharenko1@llnl.gov
Correspondence may also be addressed to Gabriela G. Loots. Tel: +1 925 423 0923; Fax: +1 925 422 2099; Email: loots1@llnl.gov

biologically relevant regulatory elements. To increase its versatility, and create a more efficient and user-friendly tool, we developed rVISTA 2.0, an improved web-based server that interconnects TFBS motif searches and cross-species sequence analysis with several comparative sequence analysis tools to significantly simplify and expedite its use. Originally, rVISTA required external alignment files to be submitted for analysis and was limited to only one alignment format. Currently, rVISTA accepts blastz alignments submitted at the rVISTA homepage or alignment and gene annotations automatically forwarded from (1) the ECR Browser, (2) zPicture and (3) GALA database. Also, we designed a new program for detecting TFBSs that is significantly faster than the MATCH program originally accompanying the TRANSFAC database (4,6). This new development significantly decreases the processing time, enabling the analysis of much larger genomic intervals.

ALGORITHM

There are four main ways to access the rVISTA tool: (i) submitting a blastz alignment file (7) at the rVISTA homepage (http://rvista.dcode.org/), (ii) dynamically generating and automatically forwarding (with a single mouse button click) zPicture alignments (http://zpicture.dcode.org/) (8), accessing pre-computed multiple genome alignment data available at (iii) the ECR Browser website (http://ecrbrowser.dcode.org/) (Figure 1A) and (iv) the GALA database. All these tools providing alignments for rVista 2.0 use the blastz program (7) to identify homologous regions and to produce local sequence alignments between the reference sequence and one or more other orthologous sequences. The local alignment method used by zPicture and the ECR Browser tools provides a careful assessment of the evolutionary rearrangements, ensuring the ability of rVISTA to detect TFBSs that have undergone positional changes relative to nearby genes and other features over the course of evolution.

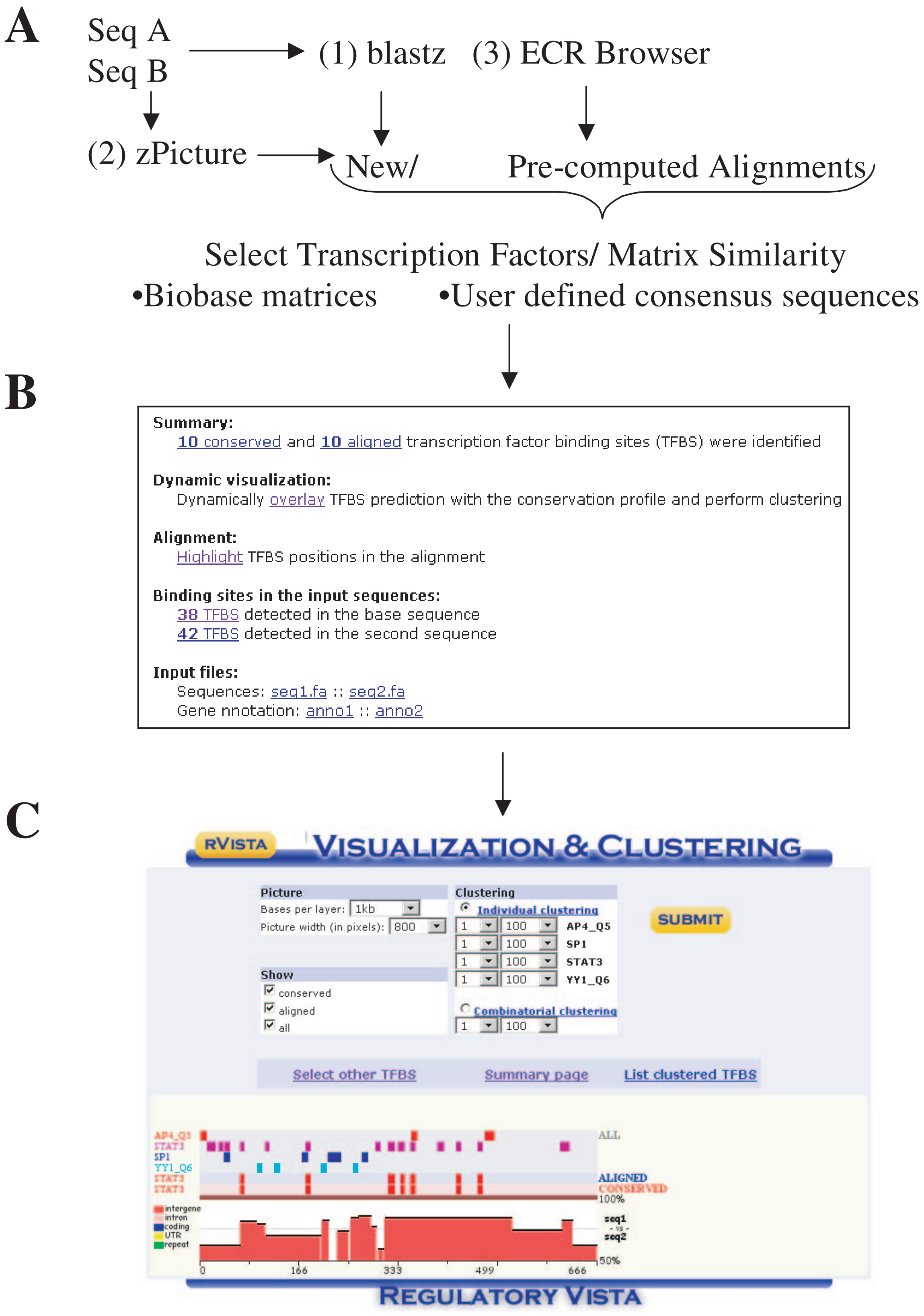
rVISTA analysis proceeds in four main steps: (i) detect TFBS matches in each individual sequence using PWMs from the TRANSFAC database, (ii) identify pairs of locally aligned TFBSs, (iii) select TFBSs present in regions of high DNA conservation and (iv) create a graphical display that dynamically overlays individual or clustered TFBSs with the conservation profile of the genomic locus. Users have the option of either selecting matrices from the TRANSFAC library or inputting their own TFBS consensus sequences. TRANSFAC professional library includes matrices from vertebrates, plants, nematodes, insects, fungi and bacteria. The current TRANSFAC library utilized by rVista 2.0 contains representatives from ~500 vertebrate TF matrices that comprise ~400 TF families. Selected matrices from this library are additionally verified and improved. Users selecting the TRANSFAC library have the option to specify the stringency to be used for the PWM identification.

We have replaced the MATCH (6) program accompanying the TRANSFAC (3,4) database with a recently developed tfSearch tool for detecting TFBS (I. Ovcharenko, unpublished data). tfSearch combines ‘suffix tree’-based fast substring searches (9) with PWM scoring of substring similarities. Transforming the original sequence into the suffix tree may use extensive memory (requiring a memory allocation ~100 times larger than the size of the sequence), but it greatly raises the efficiency in localizing substrings. A substring of size N will require O(N) operations with the suffix tree in order to localize all the matches. PWM searches that use the suffix tree require a scan of the suffix tree at a depth $\leq N$ and stop when the count at the node is below the PWM matrix similarity threshold selected by the user. Table 1 summarizes results of PWM-detecting TFBSs in two genomic loci, 100 kb and 1 Mb long, utilizing MATCH and tfSearch tools. The gain in speed obtained with use of the tfSearch tool varies from 10- to 100-fold in comparison with the time required by the MATCH program. It is especially pronounced when a large number of PWMs is used. The speed improvement thus introduced into the rVista 2.0 tool significantly decreases the tool’s response time due to the fact that detecting TFBSs in the sequence file is the performance bottleneck of this approach.

After localizing the TFBSs in both sequences, rVISTA proceeds to identify pairs of *aligned* TFBSs that are interconnected in the local blastz alignment. Genomic DNA insertions and deletions in either of the sequences (identified as gaps in the alignment) that occur in the core region of a TFBS disqualify the prediction. Subsequently, rVISTA requires aligned TFBS predictions to be locally highly conserved. Local conservation of at least 80% sequence identity in a 20 bp sliding window spanning the binding site (and always including the core of the binding site) selects *aligned-and-conserved* TFBSs (that are also referred to as *conserved* in the rVISTA output).

The rVista web page that is returned to the user contains detailed information on rVista processing results. This includes positional information on TFBS predictions in both sequences, and distribution of *aligned* and *aligned-and-conserved* TFBSs. The report includes data on the location, percentage identity and strand (Figure 1B) (reference sequence only). *Conserved* sites can also be visualized in the textual blast-like alignment, and are highlighted in blue. Finally, rVISTA results provide an interactive visualization module that overlays positional information on TFBS predictions above a graphical conservation profile that includes annotation of protein coding features for the locus. Clustering analysis of TFBSs permits the search and subsequent visualization of complex TFBS modules consisting of multiple different TFBSs (Figure 1C). For more informative analysis, users have the option to select for visualization only a subset of TFs from the initial list provided.

Several visualization parameters can be adjusted by the user: (i) alignment size (in bp) per layer, (ii) window resolution, (iii) types of site to be displayed (*all*, *aligned*, *conserved*) and (iv) the type of clustering analysis to be used. Two clustering options are also available, individual and combinatorial. Individual clustering is used for identifying groups of TFBSs belonging to the same TFs. Users have the option to indicate the number of sites and the size of the TFBS module they wish to identify. Combinatorial clustering is carried out for groups of TFBSs belonging to two or more different TFs. For example, if the visualization module is selected to display binding sites for TFs Hnf1, Tbx5 and Nkx2.5, and the user is interested in finding 100 bp regions that contain clusters comprised of at least five sites from this selected subset, rVISTA will identify all evolutionary conserved regions with any combination of these sites. In the visual display rVISTA will present only sites that fit the selected criteria (Figure 1C).



Summary:
10 conserved and 10 aligned transcription factor binding sites (TFBS) were identified

Dynamic visualization:
Dynamically overlay TFBS prediction with the conservation profile and perform clustering

Alignment:
Highlight TFBS positions in the alignment

Binding sites in the input sequences:
38 TFBS detected in the base sequence
42 TFBS detected in the second sequence

Input files:
Sequences: seq1.fa :: seq2.fa
Gene nnotation: anno1 :: anno2

Picture
Bases per layer: 1kb
Picture width (in pixels): 800

Show
☑ conserved
☑ aligned
☑ all

Clustering
● Individual clustering

1	100	AP4_Q5
1	100	SP1
1	100	STAT3
1	100	YY1_Q6

○ Combinatorial clustering
1 100

Select other TFBS Summary page List clustered TFBS

Figure 1. rVISTA 2.0 analysis data flow. (**A**) The rVISTA tool can process blastz alignments submitted at the rVISTA homepage (http://rvista.dcode.org/), or alignments automatically forwarded from the zPicture alignment program (http://zpicture.dcode.org/), the ECR Browser (http://ecrbrowser.dcode.org/) or the GALA database (http://globin.cse.psu.edu/gala/). (**B**) Users select the search criteria, and the results are returned in the same page as the downloadable static data files and dynamic links to visual analysis of TFBS distribution. (**C**) TFBSs for pre-selected TFs can be visualized above the conservation profile as tick marks, and the clustering module can detect user-specified groups of TFBSs.

Table 1. Comparative detection of PWMs in long genomic intervals performed by MATCH (6) and tfSearch programs

Region/PWMs	MATCH(s) (cut-off = 0.75)	tfSearch(s) (cut-off = 0.75)	Speed increase	MATCH(s) (cut-off = 0.85)	tfSearch(s) (cut-off = 0.85)	Speed increase
1 Mb/491 PWMs	12243.0	708.4	17×	4029.5	54.6	74×
100 kb/491 PWMs	1235.5	15.3	81×	405.1	3.9	105×
1 Mb/GATA3	40.1	4.4	9×	39.9	3.2	13×
100 kb/GATA3	4.0	0.2	20×	4.0	0.2	20×

Two different PWM matrix cut-offs (with equivalent core cut-offs in the case of the MATCH tool), 0.75 and 0.85, were analyzed. Analysis for all the 491 available TRANSFAC (4) PWMs is compared with the analysis performed with a single *GATA3* PWM. The test was performed on a 2.2 GHz Dell PC running RedHat Linux 7.3. Two loci were analyzed: 1 Mb at chr20:10,000,000–11,000,000 (human genome, NCBI Build 34) containing *ANKDR5*, *SNAP25*, *MKKS* and *JAG1* genes, and 100 kb at chr20:10,000,000–10,100,000 (human genome, NCBI Build 34) containing *ANKDR5* gene.

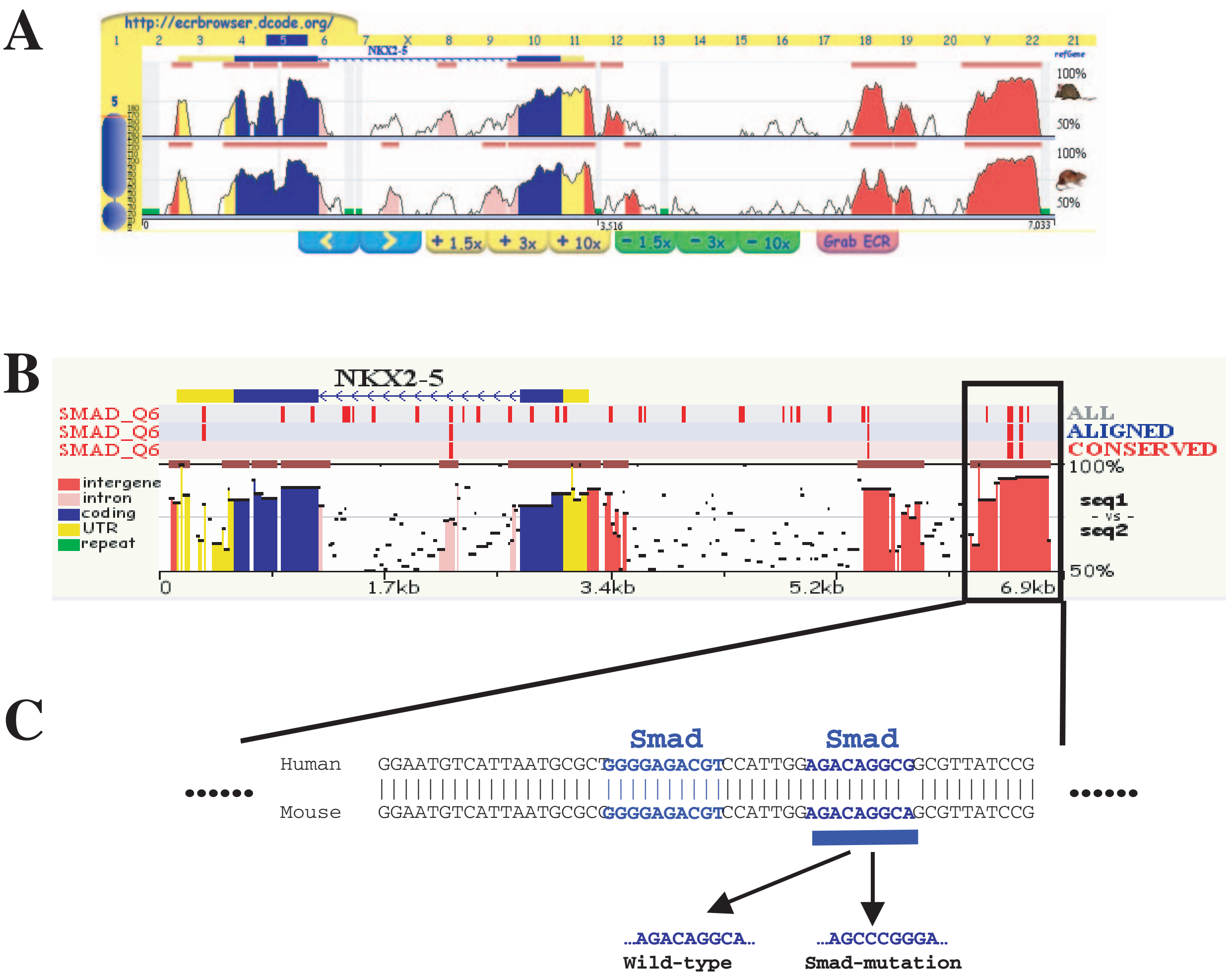


Figure 2. TFBS analysis of the *NKX2.5* genomic locus. (**A**) The *NKX2.5* genomic region was accessed in the ECR Browser. Human/mouse and human/rat alignments are displayed (7 kb in the window). (**B**) Coding exons are in blue, untranslated regions (UTRs) are in yellow, conserved intronic noncoding ECRs are in pink and conserved intergenic ECRs are in red. The alignment was processed for Smad4 binding sites. (**C**) Smad4 TFBS matches to the reference sequence (human) are in blue, aligned pairs in red and aligned-and-conserved in green. *NKX2.5* cardiac enhancer harbors 4 conserved Smad4 sites; one site corresponds with a previously functionally characterized Smad4 site.

APPLICATION

To illustrate the application of the rVISTA tool, we have carried out an unbiased analysis for the *NKX2.5* human locus with the intention of detecting the regulatory element known to play a key role in cardiac development. The conservation profile available for this gene in the ECR Browser revealed several upstream and intronic noncoding elements in this locus (Figure 2A). Experimental evidence suggests that Smad proteins are involved during cardiac development, and in particular a direct link has been established between *NKX2.5* induction and Smad consensus sequences. Therefore, Smad transcription factors were selected as ideal candidates for *NKX2.5* TFBS analysis (10–11). rVISTA analysis of the ECR Browser alignment spanning ∼7 kb including the *NKX2.5* region was performed using TFBS matrices for the transcription factor, Smad4. A TFBS search with a 0.85 PWM matrix cut-off identified 43 PWM matches across the locus, four of which are highly conserved in the human–mouse alignment (Figure 2B). All four Smad4 TFBSs are localized inside the single conserved element located ~2 kb upstream of the *NKX2.5* transcription start site. This highly conserved element has been previously shown to function as a cardiac enhancer in transgenic mice. In particular, one of these conserved Smad4 TFBSs coincides with the site mutated by Lien *et al*. (12) and has been shown to be required for the proper activity of the *NKX2.5* cardiac enhancer (Figure 2C). It was also demonstrated that a 2 basepair mutation (from A to C) in the most highly conserved Smad4 TFBS was able to diminish the cardiac enhancer properties of this regulatory element (Figure 2C) (12).

CONCLUSIONS

Understanding the function of noncoding DNA and identifying and characterizing the structure of transcriptional regulatory elements embedded in the human genome create continuing challenges. We present a completely redeveloped rVISTA 2.0 web server, for high-throughput discovery of *cis*-regulatory elements. By combining interspecies sequence conservation, reliable TF matrices and combinatorial clustering of TFBSs, rVISTA 2.0 maximizes the probability of identifying functional TFBSs. The novel features and programs implemented into rVISTA 2.0 make this tool very powerful for identifying and analyzing TFBSs in long genomic intervals. The interconnectivity with blastz, zPicture and the ECR Browser tools for genome comparative sequence analysis makes rVISTA 2.0 a valuable resource for establishing a direct link between the language of noncoding DNA and biological function of genomes.

ACKNOWLEDGEMENTS

We are grateful to Lisa Stubbs for constant support, critical comments and suggestions on the manuscript. The work was performed under the auspices of the US Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory Contract No. W-7405-Eng-48.

REFERENCES

1. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
2. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
3. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*., **24**, 238–241.
4. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*., **31**, 374–378.
5. Loots,G.G., Ovcharenko,I., Pachter,L., Dubchak,I. and Rubin,E.M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*., **12**, 832–839.
6. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.
7. Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human–mouse alignments with BLASTZ. *Genome Res*., **13**, 103–107.
8. Ovcharenko,I., Loots,G.G., Hardison,R.C., Miller,W. and Stubbs,L. (2004) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res*., **14**, 100.
9. Delcher,A.L., Phillippy,A., Carlton,J. and Salzberg,S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*., **30**, 2478–2483.
10. Liberatore,C.M., Searcy-Schrick,R.D., Vincent,E.B. and Yutzey,K.E. (2002) Nkx-2.5 gene induction in mice is mediated by a Smad consensus regulatory region. *Dev. Biol*., **244**, 243–256.
11. Yamada,M., Szendro,P.I., Prokscha,A., Schwartz,R.J. and Eichele,G. (1999). Evidence for a role of Smad6 in chick cardiac development. *Dev. Biol*., **215**, 48–61.
12. Lien,C.L., McAnally,J., Richardson,J.A. and Olson,E.N. (2002) Cardiac-specific activity of an Nkx2-5 enhancer requires an evolutionarily conserved Smad binding site. *Dev. Biol*., **244**, 257–266.