PROMOTER SEQUENCE ALIGNMENT

ALIGNMENT

CONVERVED TFBS

Two sequences

Local

Blastz
(zPicture-dcode.org)

rVista 2.0
http:// rvista . dcode . org

Global

LAGAN
(mVISTA)

rVISTA at VISTA

Multiple sequences

Global

Local

TBA/Multiz
(Mulan-dcode.org)

MULTITF at dcode.org

Daniel Rico, PhD.
drico@cnio.es

UBio

cnio

VISTA

VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species.

VISTA Servers

mVISTA

Align and compare your sequences from multiple species

Submit **Instructions** **Download** **About** **Cite**

rVISTA

Regulatory VISTA combines transcription factor binding sites database search with a comparative sequence analysis. It can be used directly or through mVISTA, Genome VISTA, or VISTA Browser

Submit **Instructions** **About** **Cite**

GenomeVISTA

Compare your sequences with whole genome assemblies. It will automatically find the ortholog, obtain the alignment and VISTA plot. View your alignment together with pre-computed alignments of other species in the same interval.

Submit **About** **Cite**

Phylo-VISTA

Analyze multiple DNA sequence alignments of sequences from different species while considering their phylogenic relationships.

wgVISTA

Align sequences up to 10Mb long (finished or draft) including **microbial whole-genome assemblies.**

Submit **Instructions** **About** **Cite**

Precomputed Whole Genome Alignments

VISTA Browser

Allows the user to examine pre-computed alignments of whole genome assemblies. Pairwise and multiple alignments are available.

Whole Genome rVISTA

Whole Genome rVISTA evaluates which **conserved** between pairs of species transcription factor binding sites (**TFBS**) are **over-represented in upstream regions in a group of genes**.

Microbial genomes

Pre-computed full scaffold alignments for **microbial genomes** are available as the VISTA component of IMG (Integrated Microbial Genomes) developed in DOE Joint Genome Institute.

Other Projects

PGA

Berkeley PGA uses a comparative genomic approach first to identify, and then to determine the function of elements regulating the expression of genes affecting the cardiovascular system.

SNP-VISTA

Visualization of mutations in genes and discovery of recombination points in microbial populations.

TreeQ-Vista

Interactive tree visualization tool with functional annotation query capabilities.

VISTA Enhancer Browser

- Enhancer Browser
- Combines computational and experimental data

VISTA Introduction

```
104637349  GTAGTGCCACTGAGTGTGACAGGGATGGCAACAAAAGCATTAAGTTCCAAGGGGAAAGAA  104637408
>>>>>>>>    |   ||  |||  |||   ||||  ||||||||||      ||  ||||  |   ||||||||||  <<<<<<<<
052290302  GAGATGTCACCAAGTA-AACAGAGATGGCAACAGGACCAATAGGTTCTAGTGGGAAAGAC  052290360
```

"sliding window" to measure sequence conservation
(default window size 100bp)

Graphical presentation of sequence conservation
as "peaks-and-valley" curve

% identity

>70% identity
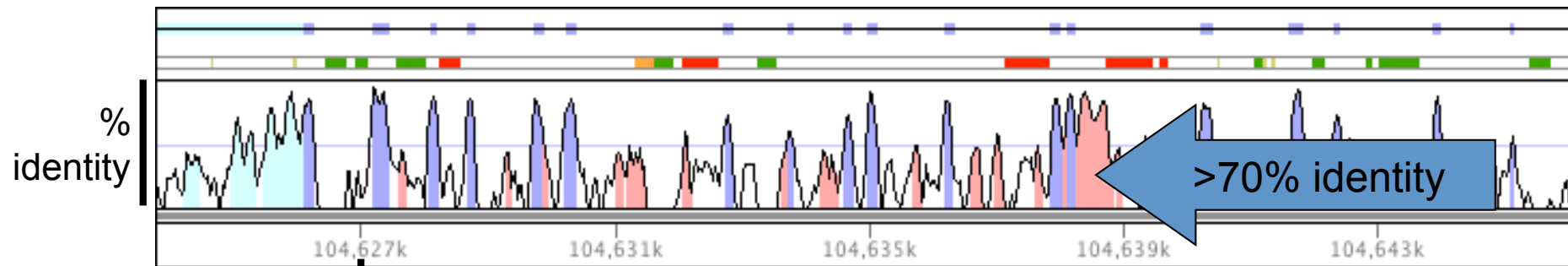
base sequence coordinates

http://dcode.org/

ECRs: Evolutionary Conserved Regions with Mulan
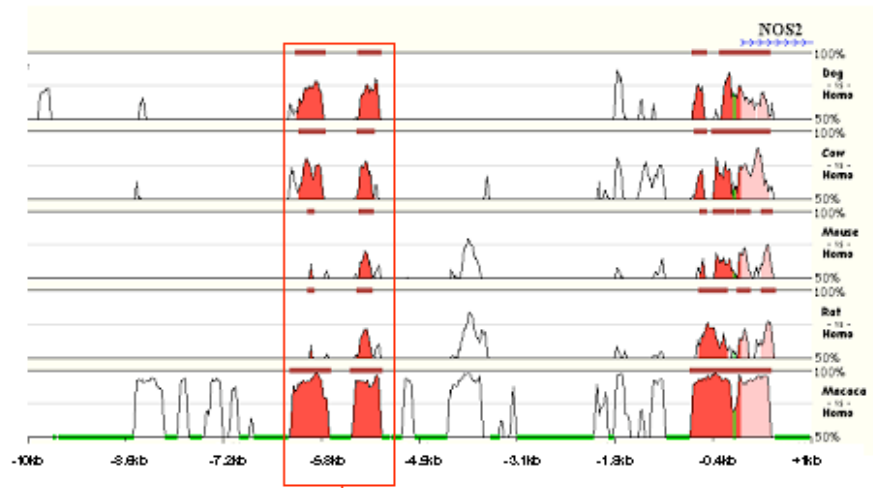
(A) Standard stacked-pairwise visualization (smooth graph) of Mulan alignments of NOS-2 gene promoter. The human sequence (from -10 kb to +1 kb) was selected as the reference species. Repeats were masked in all species with RepeatMasker (Mulan settings); green regions in the base sequence indicate the human repeats. The graphical representations of the other sequences are displayed according to their similarity to the base sequence: the closer they are to human, the higher is the conservation (top sequences are less conserved). Parameters selected for detection of evolutionarily conserved regions (ECR) were 90 bp minimum length and minimum similarity of 65% (50% bottom cut-off). Red indicates regions that are upstream from the transcription start site; pink regions are downstream from it. Two conserved motifs in rodent NOS-2 promoters indicate the presence of distal and fragmented sequences that are very similar to the unique enhancer region conferring NF-κB regulation in human NOS-2. (B) A schematic representation of the hypothetical translocation of these sequences in human and rodents; double head arrows indicate the positional translocation.
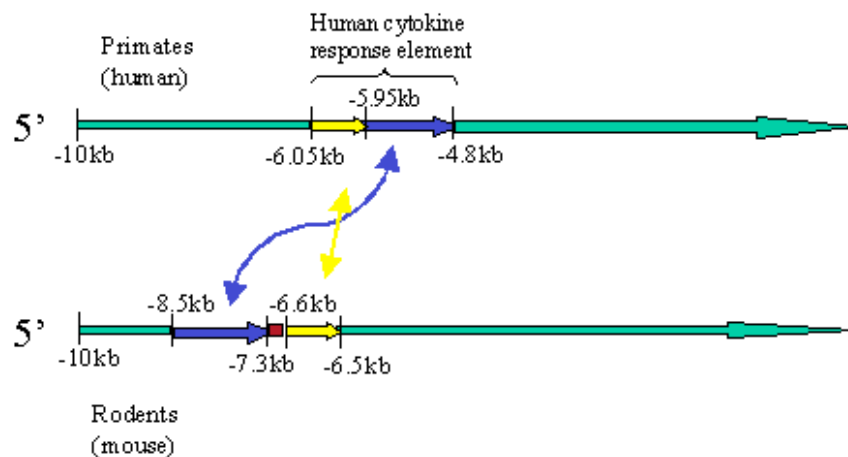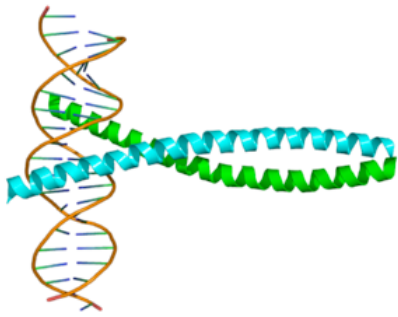
PROMOTER SEQUENCE ALIGNMENT
Guided exercise

1. Mask your fosB (mouse and human) promoter sequences for repeats: http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker Hint: mask repeats using lower case letters.

2. Submit you two sequences to zPicture. Hint: uncheck masking!

3. See the ECRs in graphical display and as alignments.

4. Send the blastZ alignment to rVISTA. Select all AP-1 matrices.

5. See graphical results and highlight conserved sites in the alignment.

RepeatMasker Web Server

RepeatMasker screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches. RepeatMasker also generates a table annotating the masked regions.

Reference: A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.2.7 (RMLib: 20090120)

Check Current Queue Status

Basic Options

Sequence:

[file input] Examinar...
or
[text area]

Select a sequence file to process or paste the sequences(s) in FASTA format. Large sequences will be queued, and may take a while to process.

Search Engine: ● wublast ○ cross_match

Select the search engine to use when searching the sequence. Cross_match is slower but often more sensitive than WUBlast.

Speed/Sensitivity: ○ rush ○ quick ● default ○ slow

Select the sensitivity of your search. The more sensitive the longer the processing time.

DNA source: [Human ▼]

Select a species from the drop down box or select "Other.." and enter a species name in the text box. Try the protein based repeatmasker if the repeat database for your species is small.

Return Format: ● html ○ tar file

Select the format for the results of your search. The "tar" option will return the results as a compressed archive file, and "html" will present the results as a summary web page with links to the individual data files.

Return Method: ● html ○ email [Your email address]

The "HTML" return method will run RepeatMasker on your sequence and return the results immediately to your web browser, provided your sequences are short. The "email" return method will email you when your results are ready.

(Restablecer) (Submit Sequence)

Lineage Annotation Options

If your query sequence is mammalian, RepeatMasker can determine if a repeat instance is expected to be present in one or more other mammalian species. This information can be used to annotate the RepeatMasker output or control the masking process.

http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

Instructions   **Example:** human-rat-fugu and human-rat *GATA3* alignment   Description

zPicture is a dynamic alignment and visualization tool that is based on **blastz** alignment program utilized by **PipMaker**. zPicture alignments can be automatically submitted to **rVista 2.0** to identify conserved transcription factor binding sites.

Genome Research, 14(3), 472-477, (2004)

multi-zPicture: multiple sequence alignment tool

1 SEQUENCE 1

Upload sequence and gene annotation from
UCSC Genome Browser
- or -

○ Paste sequence
(in FASTA format ❓)

- or -

○ FASTA file (.fa)   [   ] Examinar...

- or -

○ NCBI accession # [   ]

2 SEQUENCE 2

Upload sequence and gene annotation from
UCSC Genome Browser
- or -

○ Paste sequence
(in FASTA format ❓)

- or -

○ FASTA file (.fa)   [   ] Examinar...

- or -

○ NCBI accession # [   ]

3 OPTIONAL :: ANNOTATION 1

Repeats:

● Repeats are identified by lower-case letters
○ Mask repetitive elements [no masking ▼]

Gene annotation (if any): ❓

● Paste [   ]

○ File [   ] Examinar...

4 OPTIONAL :: ANNOTATION 2

Repeats:

○ Repeats are identified by lower-case letters
● Mask repetitive elements [no masking ▼]

Gene annotation (if any): ❓

● Paste [   ]

○ File [   ] Examinar...

Select ☐ to run **"fast"** BlastZ on microbial-size genomes
Select ☐ to perform **"chained"** (global) blastz alignment

http://zpicture.dcode.org/

zPICTURE RESULTS

Request ID: **03231218110348**   http://zpicture.dcode.org/

Dynamic visualization:
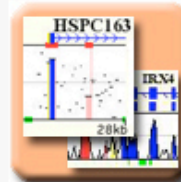
Dot-plot:

Update annotation:

edit   anno1   anno2
sequence titles

rVista 2.0 portal:

submit alignment to   RVISTA

Output files:

list of ECRs	in seq1 or seq2
blast-type alignment	seq1_seq2.blast
blastz alignment	seq1_seq2.blastz

Input files:

	1	2
sequence	seq1.fa	seq2.fa
seq. masked	seq1.txt	seq2.txt
repeats	seq1.reps	seq2.reps
annotation	anno1.txt	anno2.txt

Contact dcode@ncbi.nlm.nih.gov if you have any questions or suggestions

Picture settings	Smooth graph	Base-top switch	Width	ECR length	ECR similarity	Bottom cut-off	Graph height	Remove legend
	☑	☐	3000 bases	at least 100 bases	at least 70 %	50 %	120 pixels	☐

Refresh

intergene
intron
coding
UTR
repeat

100%

Fosb:chr7:13
- vs -
FOSB:chr19:5
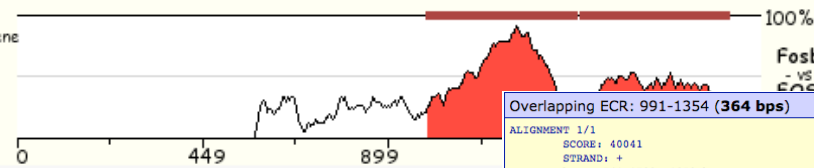
To extract an alignment a
To save the image, right mouse button c

Overlapping ECR: 991-1354 (364 bps)

```
ALIGNMENT 1/1
        SCORE: 40041
        STRAND: +
        base: 541-1799 (1259 bps)
        second: 345-1647 (1303 bps)
        CONSERVATION (:base: stats)
                Matches = 78.57% (286/364)
                Mismatches = 20.88% (76/364)
                Gapped nucleotides = 0.55% (2/364)
                base: 991-1354 (364 bps)
                second: 776-1155 (380 bps)

          1000      1010      1020      1030      1040
base    GAACAGGGTAGGTGGGGGCGCCAGAGGTG-AAGGGGACCTGCAGGCTGGGGT-----CTT
        |::::|||:|||:||||:||||||||::  |||::|||||||:::|||||:|     |||
second  GGTTTGGGGAGGGTGGGGTCCCAGGGGTATAAGCAGACCTGGGATCTGGAGTTGCACCTT
          780       790       800       810       820       830

          1050      1060      1070      1080      1090      1100
base    CCCCG-CCCGGGTCAGCGGGGTCCCTGCGGGGCTAGTCTAAGCGCCTATTATTACCAGCC
        |:||: ||||||||||||:|:|||:|:||:|| :||||:||||||||:|:|||:|
second  CTCCAACCCGGGTCAGCAGGGGCCTTCTGAGG-GAGTTTAGGCGCCTGTCAATCTCAGCC
          840       850       860       870       880       890

          1110      1120      1130      1140      1150
base    CCCGGGGCGGCGTTGCACTGCGCAGGCGCGGGCGGG---------GCGCGG--GCGCGCG
        :|||||||:||:|||||||||||||||||||||||         |||||:: ||:|||
second  TCCGGGACAGCGTGGAACTGCGCAGGCGCGGGCGGGTTCCGCACAGCGCAACAGCGGGCG
          900       910       920       930       940       950

          1160      1170      1180      1190      1200      1210
base    CGCGAGCGAGCGAGGGATTCCCTCTGACGTCATTGCTAGGATACCAAACAAACACTCCGC
        |||:::|||||||||||||||||||||||||||||||||||||||||||||||::|
second  CGCGCAGGAGCGAGGGATTCCCTCTGACGTAATTGCTAGGATACCAAACAAACACTGGGGC
          960       970       980       990       1000      1010

          1220      1230      1240      1250      1260      1270
base    CGCGCCGGCCGAGCTCCTTATATGGCTAATTGCGTCACAGGAACTCCGGGAAGGCGGGGC
        ||||:|||||||||||||||||||||||||||||||||||||||||:||::::|||:
second  CGCGCTGGCCGAGCTCCTTATATGGCTAATTGCGTCACAGGAACTCCGGGGAGGGCGGGG
          1020      1030      1040      1050      1060      1070

          1280      1290      1300      1310      1320      1330
base    CGGGATCCCCTCCCGCCGAGTGGCCCGGAACGCAACCCCCGAGACCCCCAGGGCCCCGAG
        ||||||||||||||||::|:|||::|:|:||||||||::|||||||||::|:||||||
second  CGGGATCCCCTCCCGCGAAGCCCCTCAGAACGCAGCCTTGGGGACCCCCGAGACCCCCAG
          1080      1090      1100      1110      1120      1130

          1340      1350
base    GGTCATGCAAGTGACCAGATCG
        ||||||:|:| ||::|||:|:|
second  GGTCACACTA-TGGGCAGGTGG
          1140      1150

#############################################################################
```

> base [#1]
GAACAGGGTAGGTGGGGGCGCCAGAGGTGAAGGGGACCTGCAGGCTGGGG
TCTTCCCCGCCCGGGTCAGCGGGGTCCCTGCGGGGCTAGTCTAAGCGCCT
ATTATTACCAGCCCCCGGGGCGGCGTTGCACTGCGCAGGCGCGGGCGGGG
CGCGGGCGCGCGCGCGAGCGAGCGAGGGATTCCCTCTGACGTCATTGCTA
GGATACCAAACAAACACTCCGCCGCGCCGGCCGAGCTCCTTATATGGCTA
ATTGCGTCACAGGAACTCCGGGAAGGCGGGGCGGGATCCCCTCCCGCCG
AGTGGCCCGGAACGCAACCCCCGAGACCCCCAGGGCCCCGAGGGTCATGC
AAGTGACCAGATCG

> second [#1]
GGTTTGGGGAGGGTGGGGTCCCAGGGGTATAAGCAGACCTGGGATCTGGA
GTTGCACCTTCTCCAACCCGGGTCAGCAGGGGCCTTCTGAGGGAGTTTAG
GCGCCTGTCAATCTCAGCCTCCGGGACAGCGTGGAACTGCGCAGGCGCGG
GCGGGTTCCGCACAGCGCAACAGCGGGCGCGCGCAGGAGCGAGGGATTCC
CTCTGACGTAATTGCTAGGATACCAAACAAACACTGGGGCCGCGCTGGCCG
AGCTCCTTATATGGCTAATTGCGTCACAGGAACTCCGGGGAGGGCGGGGC
GGGATCCCCTCCCGCGAAGCCCCTCAGAACGCAGCCTTGGGGACCCCCGA
GACCCCCAGGGTCACACTATGGGCAGGTGG

Promoter Sequence Alignment

UBio

Total number of transcription factor families: 467

SELECT TRANSCRIPTION FACTORS

SELECT SEPARATE TRANSCRIPTION FACTORS

A

☐ ACAAT_B	☐ AFP1_Q6	☐ AHR	☐ AHRARNT	☐ AHRHIF_Q6	☐ AHR_Q5	☐ AIRE	☐ ALPHACP1
☐ ALX4	☐ AMEF2_Q6	☐ AML1	☐ AML1_Q6	☐ AML_Q6	☐ AP1	☑ AP1FJ_Q2	☑ AP1_C
☑ AP1_Q2	☑ AP1_Q4	☑ AP1_Q6	☐ AP2ALPHA	☐ AP2GAMMA	☐ AP2REP	☐ AP2_Q3	☐ AP2_Q6
☐ AP3_Q6	☐ AP4	☐ AP4_Q5	☐ AP4_Q6	☐ APOLYA_B	☐ AR	☐ AREB6	☐ ARNT
☐ ARP1	☐ AR_Q2	☐ AR_Q6	☐ ATATA_B	☐ ATF	☐ ATF1_Q6	☐ ATF3_Q6	☐ ATF4_Q2
☐ ATF6	☐ ATF_B						

B .. C

☐ BACH1	☐ BACH2	☐ BARBIE	☐ BEL1_B	☐ BLIMP1_Q6	☐ BRACH	☐ BRCA	☐ BRN2
☐ CAAT	☐ CAAT_C	☐ CACBINDING	☐ CACCCBINDI	☐ CACD	☐ CAP	☐ CART1	☐ CBF
☐ CDC5	☐ CDP	☐ CDPCR1	☐ CDPCR3	☐ CDPCR3HD	☐ CDX2_Q5	☐ CDXA	☐ CDX_Q5
☐ CEBP	☐ CEBPA	☐ CEBPB	☐ CEBPDELTA_	☐ CEBPGAMMA_	☐ CEBP_C	☐ CEBP_Q2	☐ CEBP_Q3
☐ CETS168_Q6	☐ CETS1P54	☐ CHCH	☐ CHOP	☐ CHX10	☐ CIZ	☐ CLOCKBMAL_	☐ CLOX
☐ CMAF	☐ CMYB	☐ COMP1	☐ COREBINDIN	☐ COUP	☐ COUPTF_Q6	☐ COUP_DR1_Q	☐ CP2

Promoter Sequence Alignment

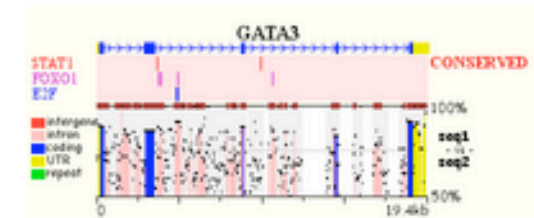**UB**io

RVISTA

RESULTS

Request ID: zpr03232009122315716

Summary:
  6 conserved and **6 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

```
              40           50
AATAAGAGATAATAATCTATT
::|   ||||||||||||||::
GCT--GAGATAATAATCTAAG
)           60
```

Binding sites in the input sequences:
  9 TFBS detected in the base sequence
  33 TFBS detected in the second sequence

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

Rerun rVista using different parameters
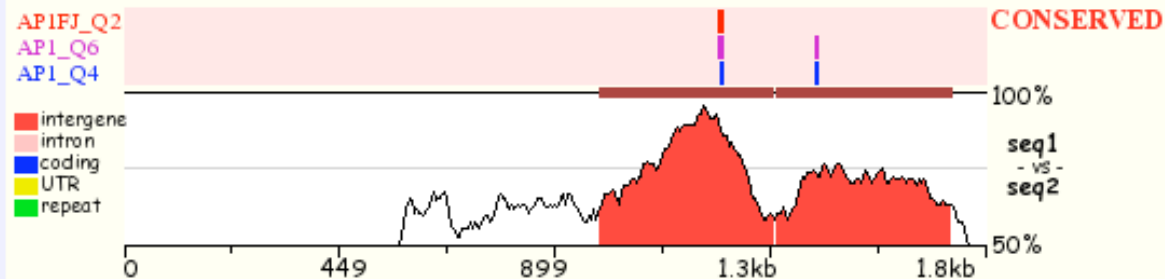
ALIGNMENT

RVISTA

Conserved AP1FJ_Q2 binding sites are in blue:

```
ALIGNMENT 1/1
        SCORE: 40041
        STRAND: +
        base: 541-1799 (1259 bps)
        second: 345-1647 (1303 bps)
        CONSERVATION (:base: stats)
                Matches = 68.71% (865/1259)
                Mismatches =  26.45% (333/1259)
                Gapped nucleotides = 4.85% (61/1259)
                base: 541-1799 (1259 bps)
                second: 345-1647 (1303 bps)


                550       560       570       580       590       600
base    CATTAATTATTATTATTAGAATTAATAATAATATCCCCTCCCTCTTACACATTCTTTGTC
        ||::|||||||||||:||::|:|  ||:||::|||:|:|:|::||||     :|:||||
second  CACCAATTATTATTGTTCAATT--ATTATTGTATACACACTTTCTT-------TTCTGTC
          350       360       370       380               390


                610       620       630       640       650
base    TCCGGGTGGATTAAAAGGTGGAAGGAGAGGCTACCAACACCATCA-----GAAGAGAGGC
        |||   ||:::|||::||  ||||||||||||:|||||||||||||     |||||:
second  TCC---TGAGCTAAGTGG--GAAGGAGAGGCTTCCAACACCATCAAGCCTGAAGAA----
           400       410       420       430       440
```
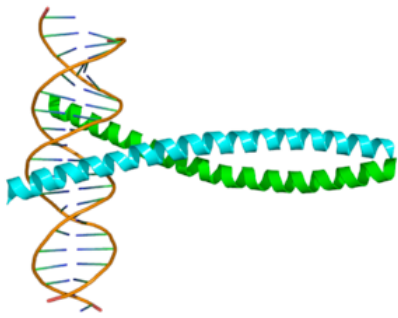
```
                1220      1230      1240      1250      1260      1270
base    GCCGGCCGAGCTCCTTATATGGCTAATTGCGTCACAGGAACTCCGGGAAGGCGGGGCCGG
        ||:|||||||||||||||||||||||||||||||||||||||||||||:|||::|||:|||
second  GCTGGCCGAGCTCCTTATATGGCTAATTGCGTCACAGGAACTCCGGGGAGGGCGGGGCGG
           1020      1030      1040      1050      1060      1070
```

Promoter Sequence Alignment

UBio

PROMOTER SEQUENCE ALIGNMENT

ALIGNMENT

CONVERVED TFBS

Blastz
(zPicture-dcode.org)

rVista 2.0
http:// rvista . dcode . org

Two sequences

Local

Global

LAGAN
(mVISTA)

rVISTA at

Multiple sequences

Global

Local

TBA/Multiz
(Mulan-dcode.org)

MULTITF   at dcode.org

Daniel Rico, PhD.
drico@cnio.es