

Published in final edited form as:

*Methods Mol Biol.* 2007 ; 395: 237–254.

## Mulan:

### Multiple-Sequence Alignment to Predict Functional Elements in Genomic Sequences

Gabriela G. Loots and Ivan Ovcharenko

#### Summary

Multiple sequence alignment analysis is a powerful approach for translating the evolutionary selective power into phylogenetic relationships to localize functional coding and noncoding genomic elements. The tool Mulan (<http://mulan.dcode.org/>) has been designed to effectively perform multiple comparisons of genomic sequences necessary to facilitate bioinformatic-driven biological discoveries. The Mulan network server is capable of comparing both closely and distantly related genomes to identify conserved elements over a broad range of evolutionary time. Several novel algorithms are brought together in this tool: the tba multisequence aligner program used to rapidly identify local sequence conservation and the multiTF program to detect evolutionarily conserved transcription factor binding sites in alignments. Mulan is integrated with the ERC Browser, the UCSC Genome Browser for quick uploads of available sequences and supports two-way communication with the GALA database to overlay GALA functional genome annotation with sequence conservation profiles. Local multiple alignments computed by Mulan ensure reliable representation of short- and large-scale genomic rearrangements in distant organisms. Recently, we have also introduced the ability to handle duplications to permit the reliable reconstruction of evolutionary events that underlie the genome sequence data. Here, we describe the main features of the Mulan tool that include the interactive modification of critical conservation parameters, visualization options, and dynamic access to sequence data from visual graphs for flexible and easy-to-perform analysis of differentially evolving genomic regions.

#### Keywords

Multiple alignment; alignment tool; evolutionary conservation; conserved elements; conserved transcription factor binding sites

## 1. Introduction

It has been determined that blocks of evolutionary conservation identified through cross-species comparisons correlate with functional DNA segments such as protein coding genes (1,2) and transcriptional regulatory elements (3,4). Several available web-based tools implement multiple sequence analysis either as a series of pairwise alignments with a selected reference sequence (5–7) or as a full multisequence global or pseudo-global alignment (8–12). Applications of these tools differ by the type of sequences (nucleotide or amino acid) they are capable of processing, as well as by the maximum length and number of allowable input sequences.

The Mulan alignment engine consists of several data analysis and visualization schemes for high-throughput identification of functional sequences conserved across large evolutionary distances. Mulan (1) determines phylogenetic relationships among the input sequences and

generates phylogenetic trees, (2) constructs graphical and textual alignments, (3) dynamically detects evolutionary conserved regions (ECR) in alignments, and (4) presents users with dynamic visual display options for flexible generation of conservation profiles. In addition, this tool is capable of implementing the phylogenetic shadowing strategy for identifying slow-mutating elements in comparisons of multiple closely related species (11). Alignments generated by the Mulan tool can be directly processed by the MultiTF program to identify evolutionarily conserved transcription factor binding sites (TFBS) shared by all analyzed species. This feature allows users to derive useful information toward decoding the sequence structure of regulatory elements that are functionally conserved among different species. Mulan is publicly available at <http://mulan.dcode.org>.

## 2. Methods

### 2.1. Alignment Strategy

Mulan provides two complementary alignment strategies for performing comparative sequence analysis of multiple sequences that are either (1) “finished” or (2) “draft” quality. The first approach operates with multiple high quality single-contig (finished) sequences, whereas the second method allows the construction of an alignment of multiple draft-quality sequences to a base (or reference) finished-quality sequence by effectively ordering-and-orienting draft sequences based on homology to the base sequence. Genomic sequences submitted to Mulan are aligned by the tba program (13) for “finished” sequences and by the refine program for “draft” sequences. The local alignment approach utilized for both sequence types reassures reliable representation of inversions and genomic reshuffling events that have occurred in a subset of lineages since the last common ancestor. It is important to mention that colinearity between input sequences (as in the case of a global alignment) is not required. Mulan generates different projections of the “threaded block-set alignment or tba” to different reference sequences that are selected by the user to ensure the detection of evolutionarily conserved elements throughout the alignment in the event orthologous regions have been repositioned or inverted in a subset (*see* Note 1).

### 2.2. Generating Alignments

- Access Mulan via the internet at <http://mulan.dcode.org/> (Fig. 1A). Alternatively access Mulan via the ECR. Browser at <http://ecrbrowser.dcode.org>, through the “Synteny/Alignments” link. Click on each box next to the sequence to be aligned and then click on the “Mulan” button provided at the bottom of the page (Fig. 1B).
- At the Mulan homepage, indicate the number of species that will be used in the analysis, and select the desired alignment type: tba-based (left button) or refine-based (right button) (Fig. 1A). (It is advised to select the tba-based approach if the user is unsure of which option is best suited, or have sequences in single-contig format. The tba-method includes more options and provides more sensitive alignments than refine.)
- Sequence input.
  - a. Submit sequences in FASTA format and gene annotation in format described in the Mulan documentation, and select the appropriate option for masking repetitive elements (Fig. 2A). Although Mulan is capable of

---

<sup>1</sup> The Mulan tool is capable of producing fast and accurate multiple alignments for both distantly and closely related organisms, properly taking into account the complexity of evolutionary sequence rearrangements such as inversions, transpositions, and reshuffling.

running Repeat-Masker locally (<http://www.repeatmasker.org/>) to mask repetitive elements in input sequences, submitting premasked sequences will significantly reduce the total processing time.

- b. If sequences of interest are available from fully sequenced genomes, Mulan can automatically fetch these sequences from the University of California Santa Cruz (UCSC) Genome Browser (<http://genome.ucsc.edu/>). To do so, the user needs to click “Upload” (Fig. 2A), and provide the necessary information for the automated upload feature to fetch the sequences directly from the UCSC Genome Browser (Fig. 2B). The required information includes: (1) the organism or genome to be used, (2) the assembly version, (3) the type of annotation, and (4) genome coordinates. Once Mulan downloads the sequence along with its annotation onto the server, the successful upload is acknowledged (Fig. 2C), and the alignment engine proceeds to create alignments between the input sequences. Note that the upload feature automatically extracts information on repetitive elements along with the sequence data; it also permits selection of different gene annotation sources. This automated upload can be combined with manual input of sequences missing representation in fully sequenced genomes.
  - c. Alternatively, if all the sequences of interest are available from fully sequenced genomes, the “Batch Upload System or BUS” integrated into Mulan can be used to simultaneously fetch all the sequences at once. Follow the BUS link on top of the sequence upload page to access this feature (Fig. 2A).
- Step-by-step specifics for the tba-based alignment approach. Upon generating a set of preliminary pairwise alignments, a phylogenetic tree is presented to the user, who has the option to accept it by clicking the “Continue” button, or refine it, if it is believed that the tree does not accurately depict the relationship between the input sequences (Fig. 3). This phylogenetic tree will be guiding the construction of the full tba-based multiple-sequence alignment.
  - Once the alignment request is completed, Mulan presents results data analysis on an interactive summary page (Fig. 4). The summary page consists of multiple links to the dynamic conservation profile visualization module, textual multiple sequence alignment (with a dynamically modified base sequence; specific to tba-based alignments), hot-link to multiTF detection of evolutionary conserved TFBS (specific to tba-based alignments), dot-plots describing sequence rearrangements, interactive selection of ECRs, etc. One also has the option to adjust annotation files and sequence titles from this page.
  - The processing information is stored on our servers for a limited amount of time (usually up to 3 months) and the data can be reaccessed anytime from the homepage (Fig. 1A) by providing the job identification number (request ID) listed on the top left corner of the summary page.

### 2.3. Visualization and Data Analysis Strategies for Multisequence Local Alignments

Multiple-sequence comparative analysis is a challenging task in terms of generating highly reliable alignments and graphically displaying the alignment results. To address the complexity stemming from user input sequence files that potentially consist of a large number of sequences of varying lengths and different phylogenetic relationships, we provide a set of different visualization options in Mulan. In general, Mulan alignment visualization is based on the zPicture display design (6), where the reference sequence is linear along the

horizontal axis and the percent identity is plotted along the vertical axis. All the dynamic visualization options can be accessed through the summary page (Fig. 4). The “Dynamic Visualization” link directs the user to the interactive alignment display (Fig. 5). At this page the top bar (Fig. 5A) allows the user to customize the visual display by selecting the desired:

1. The Graphical type of evolutionary conservation profile (smooth or percent identity plot).
2. The length of the sequence to be displayed per each line.
3. The size and percent identity of the ECR to be highlighted in the graphical alignment display.
4. The percent identity for the bottom cut-off.
5. The subregion to be indicated as “from” – “to” coordinates.

To assist in the visual analysis of conservation, Mulan has several additional options available.

1. The user can choose to color code ECRs that are present in a particular number of species (Fig. 5). This option will dynamically prioritize regions with variable degree of phylogenetic occurrence (*see* Note 2).
2. The user has the option to change the base genome in the visualization of multi-species sequence evolution. This provides the option to study conservation of regions specific to different lineages and closely related groups of species. By changing the base species, the new stacking order of conservation profiles with the rest of the species will be automatically determined using the evolutionary relationship of each sequence to the reference sequence, where more closely related species are at the bottom. (Option specific to the tba-based alignment.)
3. Visualization scheme provides the means to include or remove the legend in the display as well as to adjust the graph height.
4. Contig names and alignment blocks can be visualized as tracks on top of the conservation profile (Fig. 6). In this situation, syntenic blocks are color-coded based on their orientation in respect to the base sequence thus allowing for easy ordering-and-orienting of draft sequences by using the base sequence as the architectural guide. This feature can be used as a preassembly tool when multiple overlapping contigs are available from a homologous interval in a new species with detectable sequence similarity to the base sequence. (Option specific to the refine-based alignment.)
5. “Color density by interspecies conservation” illustrates the relationship between a conserved element and the number of species that share a particular region (Fig. 7A) such that, the more species share a sequence, the darker the conservation profile will be displayed. (This analysis is performed for every pixel-wide region of the conservation plot. The number of ECRs from different species that overlap with a particular pixel count toward the number of species sharing this region.)
6. Similar to Picture, Mulan allows interactive and customized ECR analysis. Users can select the evolutionary criteria (length and percent identity) for graphical

---

<sup>2</sup>Mulan provides users with a versatile visualization platform that allows interactive manipulation of both textual alignments and graphical conservation displays to differently address the conservation structure of either closely or distantly related species. In particular, the option to color conserved regions using a gradient based on the depth of conservation, coupled with a module that filters out ECRs that are shared by a requested number of species, permits the user to control the type of analysis performed to identify elements shared by a subset of input sequences.

identification of ECRs from the conservation plot. We have previously shown that longer and well conserved ECRs can be indicators of functional elements in genomic alignments (14) and this option permits the user to prioritize and define the optimal amount of ECRs in the studied locus—to adjust for highly conserved vs poorly conserved loci.

7. Two additional data representation modules are implemented in the Mulan tool: phylogenetic shadowing and summary of conservation. Summary of conservation collects shared similarities from all the pairwise comparisons into a single conservation profile (Fig. 7B), the phylogenetic shadowing option effectively collects pairwise mismatches (11). Thus, the summary of conservation option will aid in reconstructing conservation profiles in cases of highly diverged sequences, whereas the phylogenetic shadowing option will facilitate the identification of the most conserved elements in alignments of closely related species with a limited number of mismatches (*see* Note 3).

#### 2.4. Multisequence Conservation of TFBS

The ability to accurately predict active TFBS is a powerful approach for sequence-based discovery of gene regulatory sequences and for elucidating gene regulatory mechanisms (*see* Note 4). To combat the overabundance of false-positive computational predictions stemming predominantly from the small size of TFBS footprints and from poorly defined position weight matrices (PWM), evolutionary sequence analysis has been proposed as a robust strategy for filtering out false-positive sites (15–18). Methodologically, multiTF is similar to the rVista 2.0 tool (16,17), but implements a different strategy of detecting TFBS present in a multiple alignment. rVista 2.0 works only with pairwise sequence alignments, and requires each site to be present in a short island of high sequence conservation. In contrast, multiTF does not rely on preferential local conservation of functional binding sites vs neutrally evolving background as rVista does, instead it requires a binding site to be present in all the species at the same position as dictated by the alignment. Putative TFBS are identified in all the input sequences by using TRANSFAC PWM matrices to define consensus sequences and the tfSearch utility is used to map these consensus sequences to the genomic sequence of each input species (17,19). MultiTF excludes all TFBS predictions that overlap with exons. Gene annotation for only one of the sequences (the reference sequence) is sufficient to carry out this step. In the final step, multiTF detects TFBS predictions that are shared by all the species and are located at the same position as defined by the alignment. This is achieved by scanning through all the “anchors” or fully conserved nucleotide blocks (nucleotides that are identical in all species in the multiple-sequence alignment; Fig. 9B). If a TFBS from the reference sequence is found to overlap with an “anchor” nucleotide, we project this TFBS position to all the other species by using the alignment and excluding gaps (Fig. 9B). Starting and ending positions of the footprint of the reference sequence TFBS are compared to the starting and ending position for the same TFBS on the same strand as detected by the initial TFBS annotation. If corresponding TFBS can be identified in all the species in the alignment, this is reported by the multiTF.

To analyze Mulan alignments for the presence/absence of conserved TFBS shared among all provided, sequences the user needs to follow these steps:

1. Click on the multiTF button on the summary page (Fig. 4) to forward the alignments to the multiTF program (Fig. 8A).

<sup>3</sup>Mulan is capable of handling large genomic sequences within minutes of processing time (up to megabases in length).

<sup>4</sup>The dynamic interconnection of Mulan with multiTF presents an effective way to identify TFBS shared by multiple species. In combination, these tools can be used to predict and prioritize functional elements in otherwise anonymous sequences, a method that has been shown to be highly effective in identifying novel genes and regulatory sequences.

2. Upon forwarding to the multiTF analysis initiation page, the user selects from methods and parameters to identify TFBS in individual sequences. First, the user has to choose between the use of the TRANSFAC database of TFBS (<http://www.biobase.de/>) or user-defined consensus sequences (Fig. 8A).
3. Assuming the most common use of TRANSFAC PWM matrices in description of TFBS binding specificities to scan for binding sites, the user selects the appropriate library of phylogenies (including vertebrate, plant, fungi, nematodes, insects, and bacteria).
4. Two different options are available for detecting TFBS through the use of TRANSFAC libraries. The default option is to use the “optimized for function” search option, which weights individual PWM matrices differently by minimizing and balancing out the abundance of false-negative hits from different matrices. The alternative option is to manually specify matrix similarity cut-off for the annotation of candidate TFBS. The “optimized for function” option utilizes different cut-off parameters for different TFBS, such that no more than three TFBS per 10 kb are predicted in a random sequence (20). Manually selected cut-offs measure sequence similarity to TRANSFAC PWM; the higher the cut-offs are, the fewer sites are predicted.
5. The final option permits the selection of only “high-specificity” matrices in the TFBS annotation. This option subselects a list of TFBS matrices that have  $\leq 0.85$  cut-off similarity to the TRANSFAC PWM. These are the matrices with the most reliable definitions in the TRANSFAC database.
6. Upon submitting a request, the user is directed to a page that lists all the available transcription factor families alphabetically, where one has to choose the matrices to be used for the analysis by clicking on the provided boxes (Fig. 8B) Alternatively, the user can “select all” to obtain a full repertoire of conserved TFBS.
7. A summary page will comprehensively display the results of the TFBS analysis (Fig. 8C). Here, users can access position and matrix information provided for each sequence independently, as well as the sites can be visualized “on top of” the alignment and used in subsequent clustering analysis (Figs. 8C and 9A). The clustering options are similar to the ones available for the rVISTA 2.0 tool, TFBS can be clustered “individually” or “combinatorially” and the sites can be visualized as a summary of conservation (show binding sites by multispecies) or in each sequence individually (show all) (Fig. 9A).

## 2.5. Mulan-GALA Interconnection and Finding Orthologous Regions

The database of genomic DNA sequence alignments and annotations (GALA) allows users to find genomic intervals that meet defined conservation thresholds, alignment-based scores, and gene annotation criteria, TFBS patterns, expression profiles, and other features (21). Once a region of interest has been found, a user may wish to examine it using the Mulan tool. Likewise, once an ECR element has been identified by using Mulan, users have the option to utilize GALA to find additional information about the region containing it. Thus, two-way data flow has been established between the GALA database and the Mulan server. The interconnection link of GALA to Mulan is established through forwarding a list of homologous regions in different species from GALA to Mulan. Once a DNA interval is specified in GALA, the user can easily access a page to find estimated orthologous positions in other species.

## References

1. Pennacchio LA, Olivier M, Hubacek JA, et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*. 2001; 294:169–173. [PubMed: 11588264]
2. Gilligan P, Brenner S, Venkatesh B. Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene*. 2002; 294:35–44. [PubMed: 12234665]
3. Elnitski L, Li J, Noguchi CT, Miller W, Hardison R. A negative cis-element regulates the level of enhancement by hypersensitive site 2 of the beta-globin locus control region. *J. Biol. Chem.* 2001; 276:6289–6298. [PubMed: 11092897]
4. Loots GG, Locksley RM, Blankespoor CM, et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*. 2000; 288:136–140. [PubMed: 10753117]
5. Mayor C, Brudno M, Schwartz JR, et al. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*. 2000; 16:1046–1047. [PubMed: 11159318]
6. Ovcharenko I, Loots GG, Hardison RC, Miller W, Stubbs L. zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res*. 2004; 14:472–477. [PubMed: 14993211]
7. Schwartz S, Zhang Z, Frazer KA, et al. PipMaker: a web server for aligning two genomic DNA sequences. *Genome Res*. 2000; 10:577–586. [PubMed: 10779500]
8. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 1994; 22:4673–4680. [PubMed: 7984417]
9. Bray N, Dubchak I, Pachter L. AVID: A global alignment program. *Genome Res*. 2003; 13:97–102. [PubMed: 12529311]
10. Brudno M, Do CB, Cooper GM, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. 2003; 13:721–731. [PubMed: 12654723]
11. Ovcharenko I, Boffelli D, Loots GG. eShadow: a tool for comparing closely related sequences. *Genome Res*. 2004; 14:1191–1198. [PubMed: 15173121]
12. Schwartz S, Elnitski L, Li M, et al. NISC Comparative Sequencing Program. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res*. 2003; 31:3518–3524. [PubMed: 12824357]
13. Blanchette M, Kent WJ, Riemer C, et al. Aligning multiple genomic sequences with the threaded blocks aligner. *Genome Res*. 2004; 14:708–715. [PubMed: 15060014]
14. Ovcharenko I, Stubbs L, Loots GG. Interpreting mammalian evolution using Fugu genome comparisons. *Genomics*. 2004; 84:890–895. [PubMed: 15475268]
15. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*. 2003; 31:1753–1764. [PubMed: 12626717]
16. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*. 2002; 12:832–839. [PubMed: 11997350]
17. Loots GG, Ovcharenko I. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*. 2004; 32:W217–W221. [PubMed: 15215384]
18. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* 2003; 2:13. [PubMed: 12760745]
19. Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*. 1996; 24:238–241. [PubMed: 8594589]
20. Ovcharenko I, Loots GG, Giardine BM, et al. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res*. 2005; 15:184–194. [PubMed: 15590941]
21. Giardine B, Elnitski L, Riemer C, et al. GALA, a database for genomic sequence alignments and annotations. *Genome Res*. 2003; 13:732–741. [PubMed: 12671007]

**A**

**B**

Base sequence :: Human [hg17]

Position	Length		
<input checked="" type="checkbox"/> chr3:129460969-129494726	33758 bps	--	--

-- syntenic counterparts --

Mouse [mm5]				
Homology	Syntenic link	Length	Alignment/Graphs	Binding sites
<input checked="" type="checkbox"/>	chr6:8528100-8854166	33507 bps	<a href="#">picture_page</a>	<a href="#">fasta_2.0_analysis</a>

Frog [xt4]				
Homology	Syntenic link	Length	Alignment/Graphs	Binding sites
<input checked="" type="checkbox"/>	sc04946_48:2624810-2636302	11493 bps	<a href="#">picture_page</a>	<a href="#">fasta_2.0_analysis</a>

Chicken [gg2]				
Homology	Syntenic link	Length	Alignment/Graphs	Binding sites
<input checked="" type="checkbox"/>	chr12:3235027-3245650	10644 bps	<a href="#">picture_page</a>	<a href="#">fasta_2.0_analysis</a>

Fugu [fu3]				
Homology	Syntenic link	Length	Alignment/Graphs	Binding sites
<input checked="" type="checkbox"/>	sc04946_376:128783-135343	6564 bps	<a href="#">picture_page</a>	<a href="#">fasta_2.0_analysis</a>

Mulan :: send selected regions to [Dulan](#) to generate phylogenetic trees and identify multi-species transcription factor binding sites

**Fig. 1.** Accessing the Mulan tool from the homepage at <http://mulan.dcode.org> (A) or from the ECR Browser “Syntenic/Alignments” link (B).



**A**

Batch Upload System  
Concurrently uploads ALL the sequences from the UCSC Genome Browser

**SEQUENCE 1**  
Upload sequence and gene annotation from UCSC Genome Browser

Paste sequence (in FASTA format)

FASTA file (.fa)

NCBI accession #

**OPTIONAL :: ANNOTATION 1**

Repeats:  Repeats are identified by lower-case letters  
 Mask repetitive elements  no masking

Gene annotation (if any):

Paste  File

**B**

Step 1. Select a specie, genomic interval and genome assembly freeze from the UCSC Genome Browser (link will open in a new window)

Step 2. Describe selected region using the form below

Organism: Human

Assembly: May 2004

Annotation: RefSeq Genes

Position:  (Format: chr7:1000-2000)

Step 3. Submit your request for verification

**C**

**SEQUENCE 1**  
Sequence accepted  
seq1.fa 170742 bps

**OPTIONAL :: ANNOTATION 1**

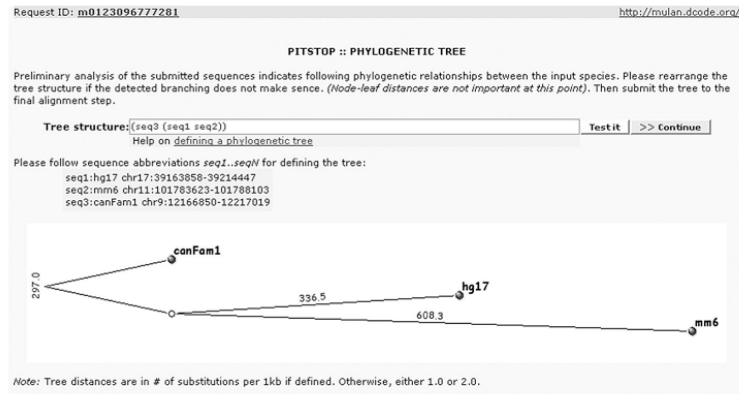
Repeats:  Repeats are identified by lower-case letters  
 Mask repetitive elements  human

Gene annotation (if any):

Paste  File

< 15042 56206 MEOKL  
15042 16552 UTR  
14553 15679 exon  
18151 18303 exon  
25709 26177 exon  
34439 34504 UTR

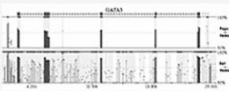
**Fig. 2.** Each sequence can be pasted in, in FASTA format, uploaded as a FASTA file, or entered as an accession number along with the available annotation (A). Alternatively, sequences can be fetched from the UCSC Genome Browser individually using the “Upload” function (A), or in groups (Batch Upload System) Browser (B). Once sequences have been uploaded, the program acknowledges the receipt (C).



**Fig. 3.** Mulan defines a guiding phylogenetic tree before proceeding with the detailed sequence alignment. The user has the option to submit modifications to this tree.

Request ID: [m0123096777281](#) <http://mulan.dcode.org/>

**Dynamic visualization:**



**Transcription factor binding sites conserved across all the species:**  
submit Mulan alignment to multiTF [MULTiTF](#)

**Pairwise dynamic plots:**  
[seq1\\_seq2](#) [seq1\\_seq3](#)

**Dot-plots:**  
[seq1\\_seq2](#) [seq1\\_seq3](#)

**Update annotation:**  
edit [anno1](#) [anno2](#) [anno3](#)  
[sequence titles](#)

**Results & output files:**  
Mulan alignment file with [hg17](#) being the base sequence :: [GENERATE](#)  
[tba\\_maf](#) (120.2 kb) :: final **TBA** multiple alignment file (blasz textual format)  
[Phylogenetic tree](#) :: evolutionary relationship among the sequences  
[refine\\_fasta](#) (160.2 kb) :: intermediate multiple alignment file (FASTA format)

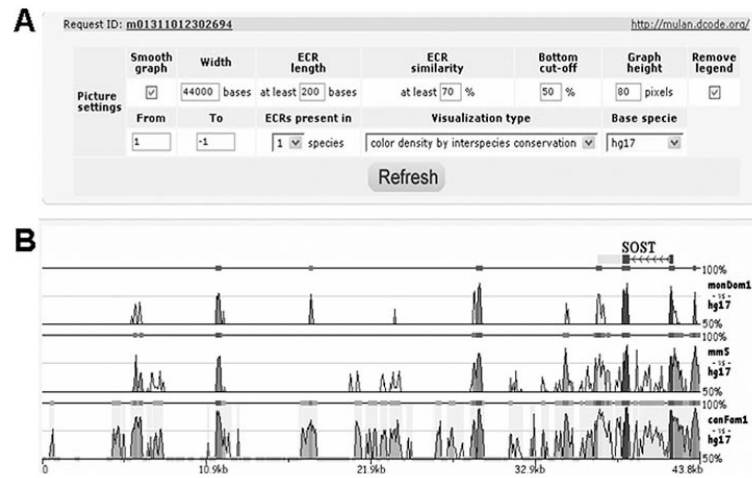
pairwise data	list of ECRs	blast-type alignment	blastz alignment
<a href="#">seq1_seq2</a>	<a href="#">detect ECRs</a>	<a href="#">seq1_seq2.blast</a>	<a href="#">seq1_seq2.blastz</a>
<a href="#">seq1_seq3</a>	<a href="#">detect ECRs</a>	<a href="#">seq1_seq3.blast</a>	<a href="#">seq1_seq3.blastz</a>

**Input files:**

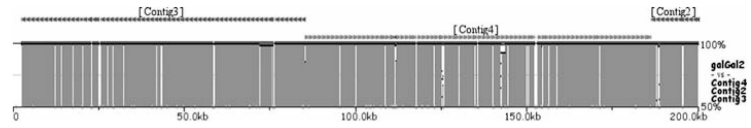
	sequence	seq_masked	repeats	annotation
1	<a href="#">seq1.fa</a>	<a href="#">seq1.txt</a>	<a href="#">seq1.reps</a>	<a href="#">anno1.txt</a>
2	<a href="#">seq2.fa</a>	<a href="#">seq2.txt</a>	<a href="#">seq2.reps</a>	<a href="#">anno2.txt</a>
3	<a href="#">seq3.fa</a>	<a href="#">seq3.txt</a>	<a href="#">seq3.reps</a>	<a href="#">anno3.txt</a>

**Fig. 4.**

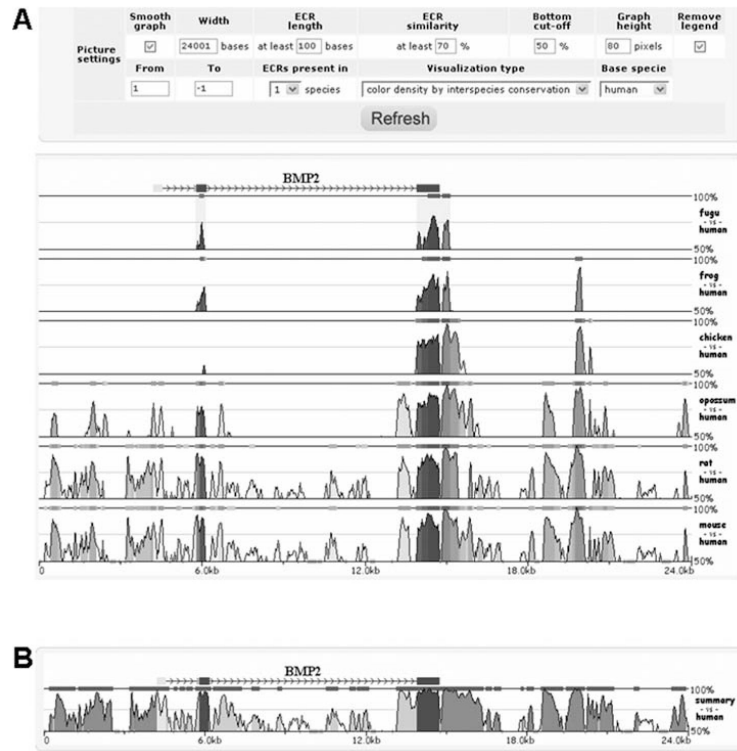
A completed alignment request results in a “summary page” that provides links to the interactive visualization platform, pairwise dynamic plots, dot plots, annotation files, sequence files, and a portal to the transcription factor binding site analysis software, MultiTF.



**Fig. 5.** Mulan interactive alignment customization options (**A**) and graphical display of alignments (**B**).



**Fig. 6.** Contig ordering based on homology to the reference sequence. The top layer of shaded lines indicates the location of contigs from a second sequence aligned to the base sequence where right-turned triangles specify forward strand alignments, and left-turned triangles correspond to reverse strand alignments.



**Fig. 7.** Mulan alignment analysis options: color density by interspecies conservation (**A**) and summary of conservation display (**B**).



