BMC Bioinformatics

Research article

Open Access

NemaFootPrinter: a web based software for the identification of conserved non-coding genome sequence regions between *C. elegans* and *C. briggsae*

Davide Rambaldi, Alessandro Guffanti, Paolo Morandi and Giuseppe Cassata*

Address: IFOM-FIRC Institute of Molecular Oncology Foundation, Milan, Italy

Email: Davide Rambaldi - davide.rambaldi@ifom-ieo-campus.it; Alessandro Guffanti - alessandro.guffanti@ifom-ieo-campus.it; Paolo Morandi - paolo.morandi@ifom-ieo-campus.it; Giuseppe Cassata* - giuseppe.cassata@ifom-ieo-campus.it

* Corresponding author

from Italian Society of Bioinformatics (BITS): Annual Meeting 2005
Milan, Italy, 17–19 March 2005

Published: 1 December 2005

BMC Bioinformatics 2005, **6**(Suppl 4):S22 doi:10.1186/1471-2105-6-S4-S22

Abstract

Background: NemaFootPrinter (Nematode Transcription Factor Scan Through Philogenetic Footprinting) is a web-based software for interactive identification of conserved, non-exonic DNA segments in the genomes of *C. elegans* and *C. briggsae*. It has been implemented according to the following project specifications:

a) Automated identification of orthologous gene pairs.

b) Interactive selection of the boundaries of the genes to be compared.

c) Pairwise sequence comparison with a range of different methods.

d) Identification of putative transcription factor binding sites on conserved, non-exonic DNA segments.

Results: Starting from a *C. elegans* or *C. briggsae* gene name or identifier, the software identifies the putative ortholog (if any), based on information derived from public nematode genome annotation databases. The investigator can then retrieve the genome DNA sequences of the two orthologous genes; visualize graphically the genes' intron/exon structure and the surrounding DNA regions; select, through an interactive graphical user interface, subsequences of the two gene regions. Using a bioinformatics toolbox (Blast2seq, Dotmatcher, Ssearch and connection to the rVista database) the investigator is able at the end of the procedure to identify and analyze significant sequences similarities, detecting the presence of transcription factor binding sites corresponding to the conserved segments. The software automatically masks exons.

Discussion: This software is intended as a practical and intuitive tool for the researchers interested in the identification of non-exonic conserved sequence segments between *C. elegans* and *C. briggsae*. These sequences may contain regulatory transcriptional elements since they are conserved between two related, but rapidly evolving genomes. This software also highlights the power of genome annotation databases when they are conceived as an open resource and the possibilities offered by seamless integration of different web services via the http protocol.

Availability: the program is freely available at http://bio.ifom-firc.it/NTFootPrinter

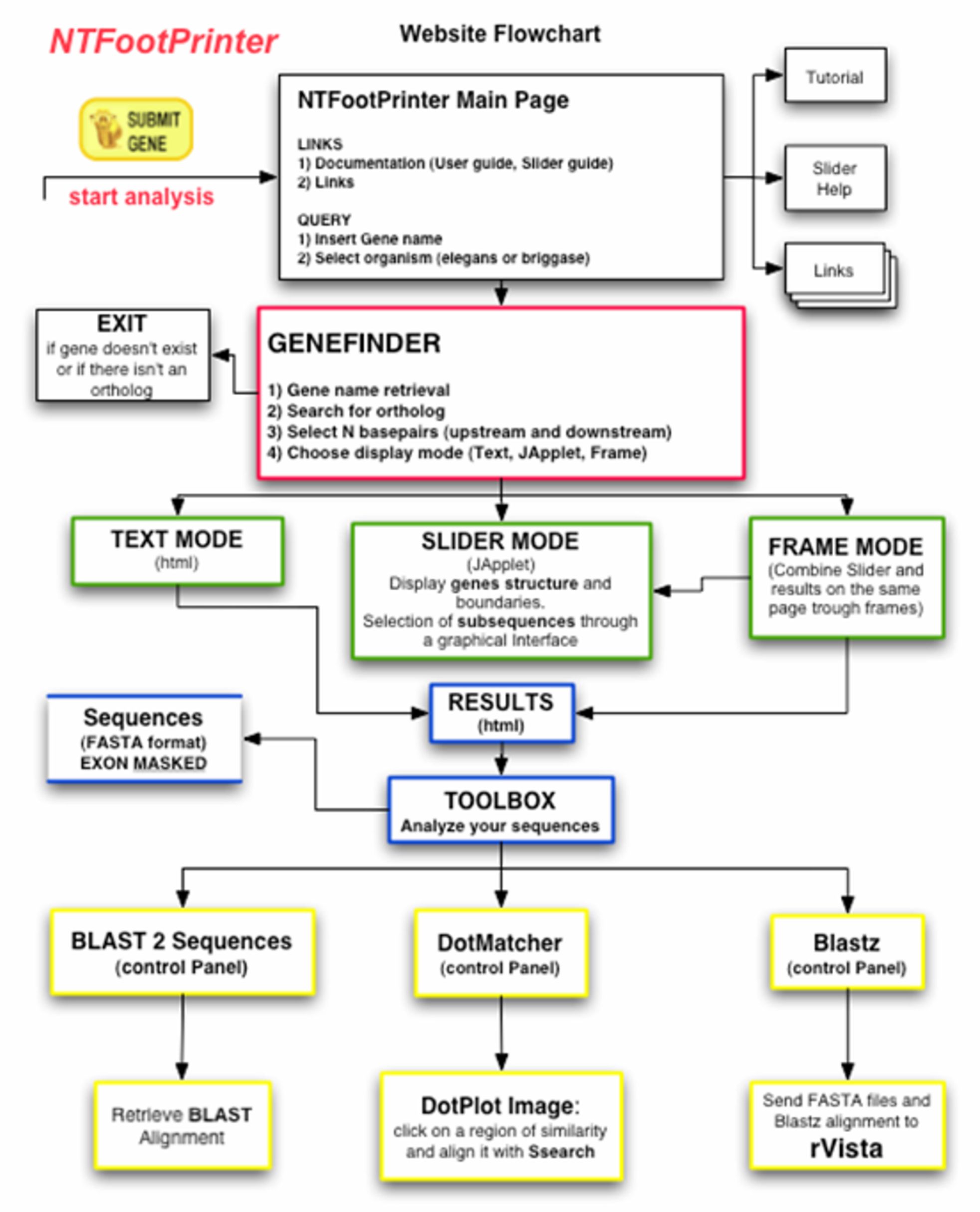

NTFootPrinter
Website Flowchart
SUBMIT GENE
start analysis
NTFootPrinter Main Page
LINKS
1) Documentation (User guide, Slider guide)
2) Links
QUERY
1) Insert Gene name
2) Select organism (elegans or briggase)
Tutorial
Slider Help
Links
EXIT
if gene doesn't exist or if there isn't an ortholog
GENEFINDER
1) Gene name retrieval
2) Search for ortholog
3) Select N basepairs (upstream and downstream)
4) Choose display mode (Text, JApplet, Frame)
TEXT MODE
(html)
SLIDER MODE
(JApplet)
Display genes structure and boundaries.
Selection of subsequences through a graphical Interface
FRAME MODE
(Combine Slider and results on the same page trough frames)
RESULTS
(html)
Sequences
(FASTA format)
EXON MASKED
TOOLBOX
Analyze your sequences
BLAST 2 Sequences
(control Panel)
DotMatcher
(control Panel)
Blastz
(control Panel)
Retrieve BLAST Alignment
DotPlot Image:
click on a region of similarity and align it with Ssearch
Send FASTA files and Blastz alignment to
rVista

Figure 1

Web Interface Scheme. NemaFootPrinter scheme: starting from gene name submission ('Start analysis' on the top-left of the scheme), gene name and organism are submitted to the GENEFINDER script (red-border box) that verifies if the given gene has an ortholog and displays the clone name. GENEFINDER also allows not-interactive selection of *n* base pairs upstream and downstream of the given genes. After gene-name retrieval the user can choose a display mode (green-border boxes): **TEXT-MODE** allows non-interactive selection of subsequences; the **SLIDER-MODE** uses an Applet Java to select interactively sub-sequences; the **FRAME-MODE** combines the slider and the result page on two horizontal frames. On the **RESULTS** page (blue-border boxes) users can find images of genes structure and boundaries generated on the fly, links to the sequences (FASTA format) and links to a series of bioinformatics tools (yellow-border boxes): **BLAST 2 sequences** is used for a first screening of the two sequences for similarities; **Dotmatcher** generates a dot plot and associate an image for direct graphical visualization of regions of similarity. By clicking on a point into the Dotmatcher image and extending the selection for *n* base pairs, it is possible to align small regions with the *Smith and Waterman* algorithm. One can then send Fasta files and Blastz alignments to the **rVista** server (the two subsequences and the relative **Blastz** alignment generated on the local server). Through this public database it is possible to identify the transcriptional regulatory elements, if any, associated with conserved subsequences.

Background

Comparative genomics is a powerful bioinformatics methodology for the identification of conserved genomic DNA segments between two related organisms [1]. Alignment of DNA sequences from different species provides an effective tool to decode genomic information, based on the assumption that functional sequences tend to diverge at a slower rate than non-functional sequences. By comparing the genomic sequences of species at different evolutionary distances, it is possible, besides identifying coding sequences, to recognize conserved non-coding sequences with a potential regulatory function, and determine which sequences are unique for a given species. This procedure is called Phylogenetic Footprinting [2,3]. Alignment algorithms optimize these comparisons so that the regions, that diverge slowly, can be anchored together and highlighted against a background of more rapidly evolving DNA, that is devoid of any functional constraints [4]. On a broader view, the identification of non-exonic Conserved Sequence Elements tags associated with human disease-related genes may open new venues for the interpretation of experimental data [5].

Although *C. elegans* and *C. briggsae* are almost identical in morphology and development [6], their genomes have diverged. Several estimates suggest that separation of the two species occurred 23–40 million years ago [7]. Conservation of DNA sequences is confined largely to protein-coding regions and short flanking sequences; functional conservation between these two species has also been demonstrated by rescue experiments of mutant phenotypes via DNA-mediated transformation [8].

We developed an interactive web-based and user friendly software to help the researchers in the identification of conserved non-coding sequence regions between the genomes of the two nematodes *C. elegans* and *C. briggsae*, starting from a bio-computational project focused on identification of conserved segments in a single pair of orthologous genes.

Implementation

The program developed here is a research tool; hence the design has been sometimes bound to the functionality, in order to optimize the speed and easiness of interaction between all the components. A careful planning of all the required modules has been achieved, however we have used system analysis and design techniques for the most complex parts of this development project.

Documentation has been targeted both at the user with a web page http://bio.ifom-firc.it/NTFootPrinter/howto.html, and to the programmer/maintainer of the software, with internal documentation on the scripts. We always relied on feedback from the interested scientist in designing the user interface and ameliorating the software functionality. From the programming language point of view, we adopted standard solutions that were fit to the problem (bioinformatics development).

The following components have been used to develop the web application:

1) A local mirror of Wormbase [9] database under mySQL. The genome annotation information in Wormbase is maintained in General Feature Format (GFF). GFF is a text-based format for the transfer of genome information, allowing genome researchers to develop tools and have them tested without having to maintain a complete feature-finding system. Documentation on this format is available at http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml

2) An EnsMart table to search orthologs. EnsMart [10] is a data retrieval tool that generates lists of biological objects (e.g. genes, SNPs) from data held in the Ensembl database. EnsMart uses the BioMart system http://www.ebi.ac.uk/biomart/.

3) A Perl web interface used to retrieve gene and sequences, mask exons, launch analysis tools, and generate gene images (GD libraries). Other web based technologies are implemented inside the Perl code: cursor coordinates capture and client-side image maps are implemented in html and JavaScript.

4) A graphical user interface implemented in Java (Swing Applet) for interactive selection of sub-sequences. Using HTTP POST and GET method, the Applet is able to communicate with other elements of the system.

5) A collection of locally compiled C++ software: **dotmatcher** and **extractseq** (EMBOSS package), **blast two sequences** (NCBI), **ssearch** (part of the FASTA program suite written by William Pearson), **blastz**.

6) An User Agent LWP connection (Perl library) to send subsequences and blastz alignment to a transcription factor database web server (**rVista**)

Results and Discussion

The main functional flow of the software can be summarized as follows (Figure 1): Starting from a *C. elegans* or *C. briggsae* gene name or identifier, the software identifies the putative ortholog (if any), based on information derived from one of the EnsMart datamart tables [10]. In this first step, the user has the opportunity to start from either an identifier ('gene model') or from a 'common gene name', following the *CGC* (*Caenorhabditis Genetics Center*) genetic nomenclature [11].

The user can select a display mode:

1. **TEXTMODE**: html output only

2. **SLIDER MODE**: graphical visualization with a Java Applet (Figure 2)

3. **FRAME MODE**: both results and slider on the same web page using frames

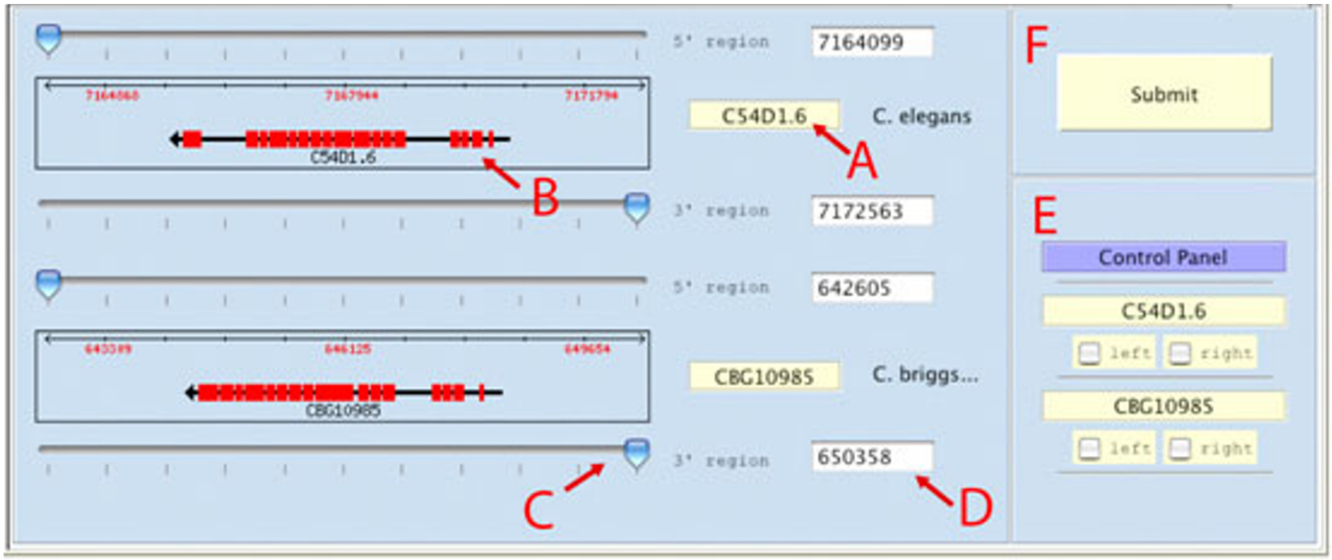
The software retrieves the sequences of the two orthologous genes from the local database. At this step, exons of the two orthologous genes are masked, since similarities between conserved coding regions are not interesting for our purpose. After sequence retrieval, the software generates '*On-the fly*' an image of the gene structures with associated intron / exon structures.

A Java applet (Figure 2) displays gene structures and neighbourhoods, the user can select subsequences. The same operation can be performed in text mode. After this step, the user can start sequence analysis from the results page.

On the results page the investigator can identify and analyze sequence similarities with a number of tools:

- **Pairwise Blast12**[12]**:** While the standard BLAST program is widely used to search for homologous sequences in nucleotide and protein databases, it is necessary to compare only two sequences to ascertain their similarity or common features. In such cases, searching the entire database would be unnecessarily time-consuming. 'BLAST 2 Sequences' utilizes the BLAST algorithm for pairwise DNA-DNA or protein-protein sequence comparison.

- **Dotmatcher13**[13]**:** Dotplot is a graphical representation of the regions of similarity between two sequences. The two sequences are placed along the axes of a rectangular matrix and (subject to threshold conditions) wherever there is equality between the sequences a dot is placed on the image. Where the two sequences have substantial regions of similarity, many dots align to form diagonal lines. It is therefore possible to glance at local regions of similarity, as these will show diagonal lines (Figure 3). In this version of Dotplot the user can control **window size**, **threshold** and **which strand to align**: a small subroutine in the web page named '*strand-helper*' helps in choosing the best configuration (align plus strand against plus, minus strand against minus, plus against minus, etc.). Even if the software selects a default strand configuration, the user can manually choose another strand mode. On the web page, text boxes give the exact position (relative to the effective sequence length) of the cursor on the X and



5' region 7164099

C54D1.6 C. elegans

3' region 7172563

5' region 642605

CBG10985 C. briggs...

3' region 650358

Figure 2
Select subsequences. CBrothersSlider.java is a small application written in Java to interactively display gene structures with intron/exon structure and to select subsequences. The interface display clone identifiers (A) and gene images generated "On-the fly" (B). Shifting the sliders (C) or submitting directly chromosome coordinates (D), the user is able to select a subsequence. A small control panel (E) can be used to select only the region on the left of the gene (*left* checkbox) or on the right of the gene (*right* checkbox); if both the checkboxes are selected, the application selects only the gene sequence. After sequence manipulation, using the Submit button (F), user can post the selected subsequence coordinates to the main script that generates new FASTA files and display the Results page.

Y-axis; this helps the user in choosing the sequence stretch of interest.

- **Ssearch 14**[14]**:** Ssearch uses Pearson's implementation of the method of Smith and Waterman [15] to search for similarities between one sequence (the query) and any group of sequences of the same type (nucleic acid or protein). After the Smith-Waterman score for a pairwise alignment is determined, Ssearch uses a simple linear regression against the natural log of the search set sequence length to calculate a normalized z-score for the sequence pair [16]. The distribution of the z-scores tends to closely approximate an extreme-value distribution; using this distribution, the program can estimate the number of sequences that would be expected to produce a z-score greater than or equal to the z-score obtained in the search. This is reported as the E() score. When all of the search set sequences have been compared to the query, the list of best scores is printed. In our implementation Ssearch is used for aligning two subsections isolated from the Dotplot output.

- **BlastZ17**[17]**and rVista 18**[18]**: BlastZ** computes local alignments for sequences of any length based on the assumption that the input sequences are related and share blocks of high conservation that are separated by regions that lack similarity and vary in length. Regions of homology are displayed collinear only to the reference sequence, while the order and orientation of the conserved elements is not necessarily the same in the second sequence.

Identifying transcriptional regulatory elements represents a significant challenge in annotating genomes. Our bioinformatics procedure has been transparently connected trough the http protocol to a computational tool, **rVista**. rVista is aimed at high-throughput discovery of *cis*-regulatory elements, combining clustering of predicted transcription factor binding sites (TFBSs) and maximizing the identification of functional sites.

A continuous exchange of ideas and information about this software and its interface occurred between the software developers and the nematode researchers. This user feedback has hence been fundamental in tailoring the graphical interface in functions of his needs.

A PAIRWISE alignment performed with **blast2seq** (blast two sequences), a heuristic algorithm (BLAST), is used for fast comparison of two sequences.

Dotmatcher uses a simple but slow algorithm; for this reason we have chosen this as a tool to verify the alignments identified in the first step. When two regions of similarity have been found, this similarity can be quantified with the **Smith and Waterman algorithm** (best local alignment). This is the most sensitive method available for pairwise sequence comparison, but it works slowly, therefore it is more appropriate for in depth analyses than primary searches.

The last component of the toolbox is a transparent connection through the http layer to a database search tool (**rVista**) focused on the identification of transcription factor binding sites related to conserved sequence elements. Continuous update of the transcription factors for vertebrates and, in particular, for nematodes is thus guaranteed. Local execution of **blastz** algorithm permits a fast genome sequence alignment, while the remote server (**rVista**) performs transcription factor binding site analysis.

Finally, in order to test the quality of the software, we performed the following experiment: Natarajan and colleagues [19] have recently determined enhancer elements in the *C. elegans* gene encoding for a beta catenin homologue: *bar-1*. By classical promoter analysis (creating transgenic lines bearing deletion constructs of the promoter region) they were able to identify two different Cis acting elements [19]. In a second step, by alignment, they show that these enhancers/elements are conserved in *C. briggsae*. By simply using our software we were able to confirm all the elements described in this work without any molecular biology experimentation needed (data not shown). With the help of NemaFootPrinter, from now on, researchers will "first" use our software and "then" test the results by creating just a few transgenic strains in order to just "test" the results and not to "identify" the regions by tedious *in vivo* experiments.

Other tools for performing comparative genomics

CisHorto, another tool for transcription-factor identification [20], uses Position Weight Matrices and a user-provided ungapped multiple alignment to predict new transcription factors binding sites. Our software adopts a simpler strategy for the identification of putative transcription factor sites, based on sequence similarity and exon masking to highlight similarities between non-coding regions.

CSTminer [21] is an user-friendly tool for generic identification of coding and noncoding conserved sequence tags through cross-species genome comparison, that uses an original algorithm to identify statistically significant conserved blocks and assess their coding or noncoding nature through the measure of a "coding potential score". We focused our development specifically on nematode genetics, leaving to the final user a high degree of interactivity and exploration for the identification of conserved sequence regions.

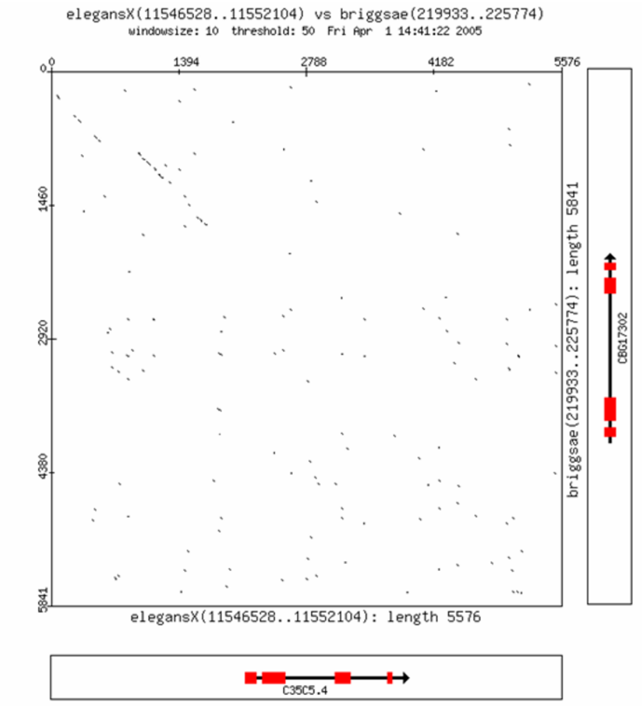


Figure 3
dotmatcher. A Dotplot image: *C. elegans* and *C. briggsae* sequences are placed on the axes. The gene structure generated on the fly help the Investigator to orient itself in the rectangular image. The top of the image shows parameters given by the investigator (windowsize and threshold). This web images are clickable by the user, the single click is extended in both directions for *n* base pairs. Both segments (one for *C. elegans* and one for *C. briggsae*) are sent to the Ssearch control page, in order to align segments with the Smith and Waterman algorithm.

Conclusion

Genome annotation databases such as Wormbase [22] or EnsEMBL [23] have a fundamental role in modern biological research and they also offer a platform to bioinformatics development. We have presented a simple web-based software aimed to identify conserved functional segments outside exons (putative new gene expression control elements) through comparative genomics between the nematodes *C. elegans* and *C. briggsae* [24]. With this project we have highlighted that a specific bioinformatics project can be realized by a sound integration of genome database mirrors, local development and transparent integration with remote resources [25]. Where speed and robustness were needed, we relied on local mirroring of databases and development of software modules, but when other resources already solved the task we integrated seamlessly calls to remote services through the http protocol. As a result, a new resource, aimed at solving a specific biological problem is now freely available at http://bio.ifom-firc.it/NTFootPrinter/howto.html. We have also demonstrated that usability and functionality in bioinformatics development can be achieved only through a strong and continued feedback from the scientist/user.

Availability and requirements

Project name: *NemaFootPrinter*

Project home page: http://bio.ifom-firc.it/NTFootPrinter/index.html

server side: UNIX type platforms

client side: Any operating system

Programming language: SQL, Perl, Java

Other requirements:

The web-based application was tested and is compatible with the more common Internet browser. For the Slider Applet the user must have a Java Virtual Machine installed and configured on the client. User without Java can use the TEXT MODE to analyze genes. Even the older *text-only* browsers like 'lynx' are compatibles with the software (obviously text-only browser must use the TEXT MODE display). For more compatibility information check the help page: http://bio.ifom-firc.it/NTFootPrinter/slider_help.html

Authors' contributions

All authors contributed to the development of NemaFootPrinter.

DR wrote most parts of the NemaFootPrinter core, the CGI scripts, the Java applet and the database handlers. AG was responsible of the main bioinformatics programming strategy.

PM was respnsible of biological applied aspect.

GC made the scientific supervision and interface design. All authors drafted the manuscript and approved the final version.

Acknowledgements

The authors wish to thank all the teams behind the public resources used for this development work: Wormbase, EnsEMBL, EnsMart and rVista. DR is supported by an AIRC (Italian Association for Cancer Research) fellowship in the framework of the AIRC Bioinformatic Center Grant (BICG). AG is supported by FIRC (Italian Foundation for Cancer Research). NemaFootPrinter is a project developed with Perl, MySQL and others open-source software, for these reasons we wish also thank all the Open Source community for their support to the Bioinformatics research. We also acknowledge the members of the Cassata lab for comments on the manuscript.

References

1. Nardone J, Lee DU, Ansel KM, Rao A: **Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA.** *Nat Immunol* 2004, **5**:768-774.
2. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9**:211-223.
3. Zhang Z, Gerstein M: **Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.** *J Biol* 2003, **2**:11.
4. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB: **Benchmarking tools for the alignment of functional noncoding DNA.** *BMC Bioinformatics* 2004, **5**:6.
5. Boccia A, Petrillo M, di Bernardo D, Guffanti A, Mignone F, Confalonieri S, Luzi L, Pesole G, Paolella G, Ballabio A, Banfi S: **DG-CST (Disease Gene Conserved Sequence Tags), a database of human-mouse conserved elements associated to disease genes.** *Nucleic Acids Res* 2005, **33**:D505-510.
6. Nigon V, Dougherty EC: **Reproductive patterns and attempts at reciprocal crossing of Rhabditis elegans Maupas, 1900, and Rhabditis briggsae Dougherty and Nigon, 1949 (Nematoda: Rhabditidae).** *J Exp Zool* 1949, **112**:485-503.
7. Emmons SW, Klass MR, Hirsh D: **Analysis of the constancy of DNA sequences during development and evolution of the nematode Caenorhabditis elegans.** *Proc Natl Acad Sci U S A* 1979, **76**:1333-1337.
8. Maduro M, Pilgrim D: **Conservation of function and expression of unc-119 from two Caenorhabditis species despite divergence of non-coding DNA.** *Gene* 1996, **183**:77-85.
9. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK, *et al.*: **WormBase: a comprehensive data resource for Caenorhabditis biology and genomics.** *Nucleic Acids Res* 2005, **33(Database)**:D383-389.
10. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data.** *Genome Res* 2004, **14**:160-169.
11. Horvitz HR, Brenner S, Hodgkin J, Herman RK: **A uniform genetic nomenclature for the nematode Caenorhabditis elegans.** *Mol Gen Genet* 1979, **175**:129-133.
12. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174**:247-250.
13. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**:276-277.
14. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**:2444-2448.
15. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147**:195-197.
16. Pearson WR: **Comparison of methods for searching protein sequence databases.** *Protein Sci* 1995, **4**:1145-1160.
17. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13**:103-107.
18. Loots GG, Ovcharenko I: **rVISTA 2.0: evolutionary analysis of transcription factor binding sites.** *Nucleic Acids Res* 2004, **32**:W217-221.
19. Natarajan L, Jackson BM, Szyleyko E, Eisenmann DM: **Identification of evolutionarily conserved promoter elements and amino acids required for function of the C. elegans beta-catenin homolog BAR-1.** *Dev Biol* 2004, **272**:536-557.
20. Bigelow HR, Wenick AS, Wong A, Hobert O: **CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting.** *BMC Bioinformatics* 2004, **5**:27.
21. Castrignano T, Canali A, Grillo G, Liuni S, Mignone F, Pesole G: **CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison.** *Nucleic Acids Res* 2004, **32**:W624-627.
22. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J: **WormBase: network access to the genome and biology of Caenorhabditis elegans.** *Nucleic Acids Res* 2001, **29**:82-86.
23. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, *et al.*: **Ensembl 2005.** *Nucleic Acids Res* 2005, **33(Database)**:D447-453.
24. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, *et al.*: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics.** *PLoS Biol* 2003, **1**:E45.
25. Geraghty DE, Fortelny S, Guthrie B, Irving M, Pham H, Wang R, Daza R, Nelson B, Stonehocker J, Williams L, Vu Q: **Data acquisition, data storage, and data presentation in a modern genetics laboratory.** *Rev Immunogenet* 2000, **2**:532-540.