BMC Bioinformatics

Methodology article

SeqVISTA: a graphical tool for sequence feature visualization and comparison

Zhenjun Hu[1], Martin Frith[1], Tianhua Niu[2] and Zhiping Weng*[1,3]

Address: [1]Bioinformatics Program, Boston University, Boston, MA, 02215, USA, [2]Program for Population Genetics, Harvard University, Boston, Mass, 02115, USA and [3]Department of Biomedical Engineering, Boston University, Boston, MA, 02215, USA

Email: Zhenjun Hu - zjhu@bu.edu; Martin Frith - mfrith@bu.edu; Tianhua Niu - niu@bioinfo.stat.harvard.edu; Zhiping Weng* - zhiping@bu.edu

* Corresponding author

Published: 4 January 2003

BMC Bioinformatics 2003, **4**:1

Received: 15 September 2002
Accepted: 4 January 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/1

Abstract

Background: Many readers will sympathize with the following story. You are viewing a gene sequence in Entrez, and you want to find whether it contains a particular sequence motif. You reach for the browser's "find in page" button, but those darn spaces every 10 bp get in the way. And what if the motif is on the opposite strand? Subsequently, your favorite sequence analysis software informs you that there is an interesting feature at position 13982–14013. By painstakingly counting the 10 bp blocks, you are able to examine the sequence at this location. But now you want to see what other features have been annotated close by, and this information is buried several screenfuls higher up the web page.

Results: SeqVISTA presents a holistic, graphical view of features annotated on nucleotide or protein sequences. This interactive tool highlights the residues in the sequence that correspond to features chosen by the user, and allows easy searching for sequence motifs or extraction of particular subsequences. SeqVISTA is able to display results from diverse sequence analysis tools in an integrated fashion, and aims to provide much-needed unity to the bioinformatics resources scattered around the Internet. Our viewer may be launched on a GenBank record by a single click of a button installed in the web browser.

Conclusion: SeqVISTA allows insights to be gained by viewing the totality of sequence annotations and predictions, which may be more revealing than the sum of their parts. SeqVISTA runs on any operating system with a Java 1.4 virtual machine. It is freely available to academic users at http://zlab.bu.edu/SeqVISTA.

Background

A significant portion of modern biological research involves the identification of the biochemical and biological functions associated with one or multiple positions of a sequence. Numerous databases have been constructed to store these sequence regions and their associated functions, defined as sequence features. An example compilation of such databases is available at http://zlab.bu.edu/~mfrith/tools.shtml. Common features for DNA sequences include introns, exons, 3' or 5' untranslated regions, transcription start sites, cis-elements and other protein binding sites, repeats, low complexity regions and single nucleotide polymorphisms (SNPs). Protein sequence features include secondary structures (α-helices and β-strands), transmembrane regions, and post-translational modifications such as phosphorylation and glycosylation

sites. There can be dozens of features associated with a single sequence. Frequently, features can be nested; for example, a SNP can reside within a cis-element, which can be in an intron. Therefore, it is extremely difficult for a text record in a database to reveal all of the salient features of a sequence to the user in an intuitive fashion.

The human genome project has motivated substantial scientific and technological developments in sequencing large eukaryotic genomes. Among the many tools developed in the course of the project, web-based graphical viewers facilitate the search, display and retrieval of sequences and annotations associated with a genome. Such viewers are typically integrated with the databases that store the genomes. They are not only extremely important for delivering the final results of a sequencing project to lab-bench biologists but also indispensable in the assembly and annotation of genome drafts, since assorted evidence must be integrated. Three well-known genome viewers are available for the public working draft of the human genome: the viewers developed by the Ensembl project http://www.ensembl.org; [1], the human genome browser at UCSC http://genome.ucsc.edu; [2] and the NCBI map viewer http://www.ncbi.nlm.nih.gov. The focus of genome viewers is typically above the gene level, with the most common use of searching for evidence of novel genes. VISTA is another user-friendly program for visualizing the alignment of very long DNA sequences [3]. With the rapid enrichment of annotated sequence features, there is a need for sequence viewers at the nucleotide or amino acid level, targeting lab-bench experimentalists. An example of such a sequence viewer, viewGene, focuses on polymorphism visualization [4].

Computational analysis of DNA and protein sequences is among the most frequently encountered activities in bioinformatics research. Computational tools for sequence analysis are often specialized in producing only one kind of feature, and frequently in text output format. The most widely used tools detect genes [5–8], cis-elements and general promoter regions [9–12], repeats [13,14], protein secondary structures [15–17] and protein transmembrane helices [18,19]. Currently, there is no visualization tool to easily compare the output of a sequence analysis program with the annotations of this sequence stored in a database, as well as to compare the outputs of multiple analysis programs. THEATRE [20] is an attempt to combine the sequence features produced by widely used sequence analysis tools or sequence databases; however, it only produces a static postscript graph.

We have developed SeqVISTA with the exact goal of facilitating the visualization of sequence features in annotation records such as those of GenBank [21] and Swiss-Prot [22], as well as the comparison of multiple sequence feature sets, produced by different sequence analysis programs, with the annotation record. We take advantage of the observation that all sequence features are indexed with one or several positions of the sequence, and construct a coherent framework for the representation of virtually unlimited feature types and feature sets. SeqVISTA can be a general platform for integrating numerous sequence analysis tools, and thus alleviate the need of developing program-specific visualization software. More importantly, with careful programming design and implementation, SeqVISTA targets the broad community of experimental biologists. All features are linked directly and dynamically to the sequence itself, and a user is presented with the global view of the most salient features. Furthermore, the user can extract any feature-containing sequence region easily and precisely for performing further experiments.

Application

Use SeqVISTA to display sequence records

The most basic application of SeqVISTA is to display sequence records. Figure 1 is a screen shot of the SeqVISTA window displaying the human ALAD gene. The window contains three panels: the left panel (*tree panel*) is a tree structure of all features with each type in a different color. Such a tree structure is our approach to a clear presentation of the complex organization of sequence features. It combines the flexibility of operating on feature types and the precision of selecting individual features. In Figure 1, the first intron is expanded to show a number of features nested within it, including an alternatively spliced exon. The second, fourth and seventh introns also overlap with other features (but not expanded). The top-right panel (*graphics panel*) contains graphical depictions of the gene and mRNA structures. The gene is represented by a thick line, and each feature is indicated by a color box, with features on the + strand drawn above the line and – strand below the line. Features associated with a single nucleotide (such as SNPs) or a single amino acid (such as variations or post-translational modifications) are indicated with thumb pin symbols. The location and width of each color box represent the location of the feature in the gene and the number of bases the feature corresponds to. The user can zoom in any region of the gene. The mRNA structure is illustrated using only the exons, including untranslated regions such as the poly-A tail if applicable. All exons in the mRNA structure are aligned with the exons in the gene structure. For the ALAD gene, there are two mRNAs, corresponding to alternative splicing. The lower-right panel (*sequence panel*) contains the nucleotide sequence for the gene (or amino acid sequence in the case of a protein).

The three panels of SeqVISTA collectively present all aspects of an annotated sequence. The tree panel emphasiz-

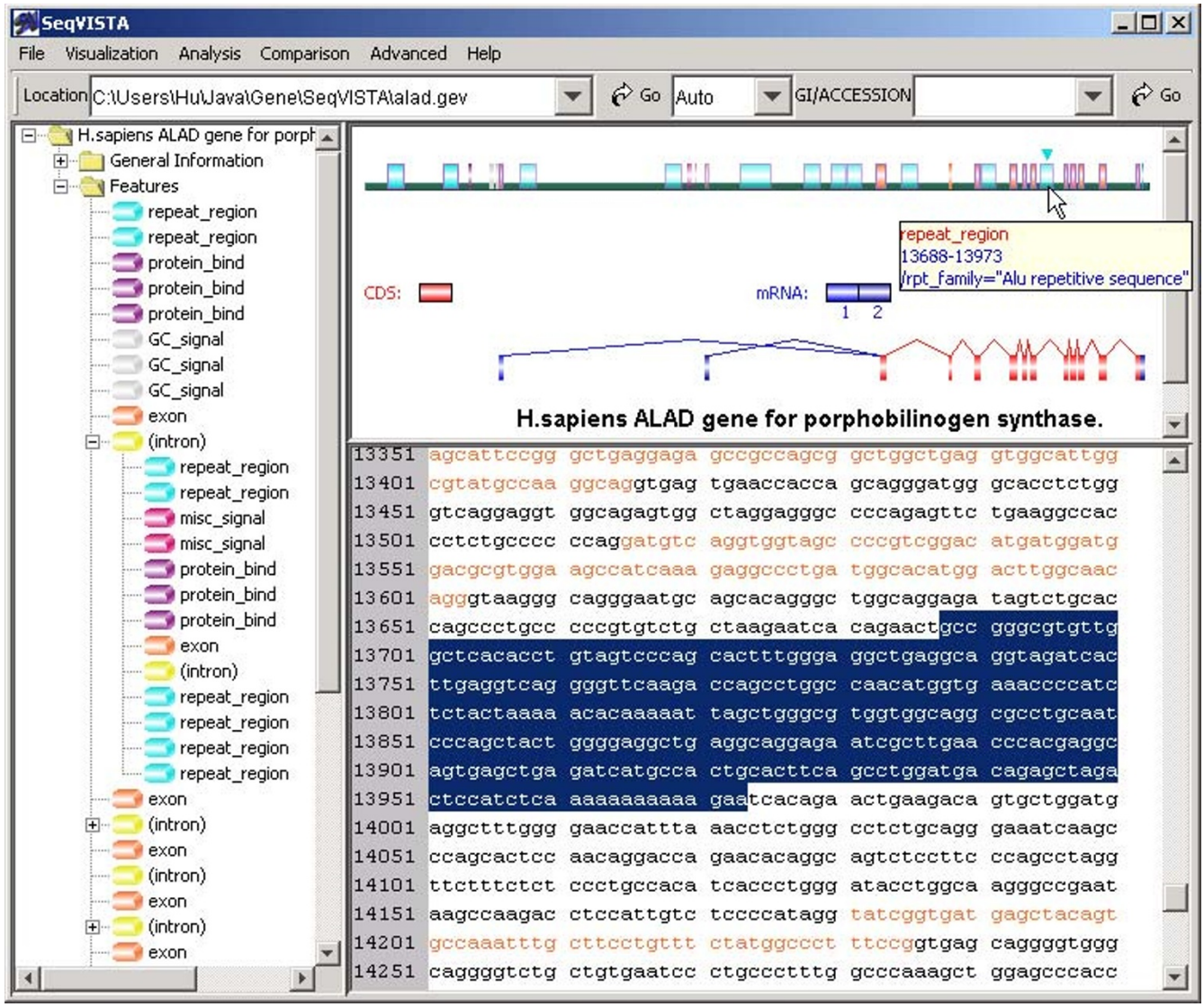

Figure 1
Screen shot of SeqVISTA displaying the human ALAD gene.

es the organization of features. The graphics panel focuses on the locations and sequence lengths of features. The sequence panel illustrates the sequence details. The three panels are dynamically linked in two ways. First, each type of feature adopts the same color in all panels. Second, if the user selects a feature in one panel by mouse clicking, the corresponding feature or sequence region in the other two panels will be highlighted accordingly (the sequence panel will be scrolled automatically to show the highlighted region, if it was not visible in the previous scroll). For clarity in the graphics panel, as well as for accommodating nested features, we allow any type of feature to be hidden (by right-clicking any instance of the feature in the tree panel). Hidden features are shown in parentheses in the tree panel (for example, intron in Figure 1), and they are not shown in the graphics panel. The nucleotides or amino acids in the sequence panel are colored according to the outermost layer of non-hidden features. SeqVISTA responds to all user requests, such as selecting or hiding features, by updating the display in all panels accordingly. The user can also mouse over a feature to obtain its annotation without updating the display in other panels (e.g., the Alu repeat in Figure 1). In every panel, the user can output the content of the entire panel as a color image in jpeg format. The user can also save the image in the graphics panel at higher resolution (300 or 600 dpi), which can be directly incorporated into scientific publications.

An important goal of our design is to make SeqVISTA extremely friendly to experimentalists. An experimentalist frequently needs to locate a fragment in a long sequence according to the starting and ending coordinates of the fragment. Manually counting the positions is a laborious and error-prone process. We have developed several functions in the sequence panel of SeqVISTA to render this task effortless: 1. The user can input the start and end coordinates and the corresponding fragment will be highlighted. 2. The user can also search using the sequence of a fragment, with the option of searching both the forward and the reverse strands of a DNA sequence. 3. We also accept regular expressions for the search, if the exact sequence of the fragment is not available. 4. The user can highlight a region and search for more occurrences in the entire sequence. All these functions can be operated entirely with copying and pasting with the mouse to avoid manually typing sequences. They can be evoked by right-clicking in the sequence panel. They are also available from the Edit tab of the top menu bar.

Another user-friendly aspect of SeqVISTA is that the user can launch the program while browsing a sequence record using Internet Explorer, by clicking the "SeqVISTA" button, which is added to the browser during the installation of SeqVISTA. Another flavor of this function is that the user can load several records into SeqVISTA by opening a text file that contains their GenBank Identification (GI) numbers. Note that even though SeqVISTA can accommodate multiple sequence records, it does not align them automatically. In the future, we plan to implement functions to integrate multiple sequence records associated with the same gene or protein.

One more example involves the dynamic display of SeqVISTA functions that are sequence type specific. SeqVISTA is capable of displaying both DNA and protein sequences and different functions are involved with different sequence types. However this could become confusing to a user if all functions are available at all times. Instead, we have ensured that the functions not applicable to a particular sequence type are grayed out.

Use SeqVISTA to compare multiple sets of annotations

One of our goals is to establish SeqVISTA as a general platform for visualizing the results of sequence analysis software and comparing them to the annotations of the same sequence stored in a public database. We have developed several means to facilitate communication among different software programs: 1. *Common format*. SeqVISTA accepts several formats: GenBank flat file (GBFF) format, GenBank HTML format, FASTA format, and the simple meta-data based SeqVISTA format. In the future, we plan to support EMBL format as well. The user can load a sequence record into SeqVISTA by supplying its GI number, or accession number, or the web address while viewing it at the NCBI website. The user can visualize the outputs of a sequence analysis software package using SeqVISTA, as long as they are in any of the above formats. SeqVISTA format allows the user to save multiple records and outputs into one file, which makes future viewing easy. 2. *Plugin*. Plugins can be developed to recognize the outputs of external software. Thus, any external software can use SeqVISTA as its graphical interface, instead of developing its own specialized graphical modules. Detailed instructions for developing SeqVISTA parser plugins and sample codes can be found at the SeqVISTA web site http://zlab.bu.edu/SeqVISTA. 3. *Direct query*. Most widely used sequence analysis programs support web servers. For these programs, we can develop SeqVISTA functions to directly query a web server and retrieve results. For all of the above three means of retrieving results, we can display the results directly in the graphics panel to facilitate the comparison between the results and database annotations. We use three examples to illustrate the above functions of SeqVISTA.

RepeatMasker

RepeatMasker screens an input DNA sequence against a library of repetitive elements (A.F.A. Smit & P. Green, unpublished data). The program produces a list of identified repeats, their locations in the input sequence, their match scores and three quantities associated with the scores: % substitution, % insertion and % deletion. The repeats identified by RepeatMasker can be displayed as if they are sequence features. The various scores can be displayed as bar graphs. We have developed a SeqVISTA function to query the web server of RepeatMasker directly. The user only needs to right-click in the graphics panel and choose the RepeatMasker option in the nucleotide analysis tab, and SeqVISTA will submit the sequence being viewed to the RepeatMasker server, retrieve the results and display them.

Figure 2 displays the results of RepeatMasker for the human Alad gene. In the Graphics panel of the SeqVISTA window, the locations of predicted repeats are shown as a feature plot and the various scores as bar graphs. The predicted repeats are treated as typical sequence features; therefore, all properties of features apply to them. For example, a tree structure of these repeats is added to the tree panel. A user can hide/show a repeat type in the tree panel, or click a repeat location to highlight its sequence in the sequence panel. Figure 2 indicates that the predicted repeats tend to locate in non-coding regions, and most of the high scoring repeats agree with the GenBank annotation. Interestingly, all of the repeats on the – strand of the sequence are annotated on the + strand in the GenBank record.

Figure 2
Comparing the results of RepeatMasker to GenBank Annotations.

PSIPRED

PSIPRED is one of the most accurate programs for predicting protein secondary structures (α-helices and β-strands) [18]. It produces a confidence score (between 0 and 9) for each position of the input sequence to be in a secondary structure state or otherwise in the coil state. An α-helix is made of a contiguous stretch of positions in the α-helix state, likewise for a β-strand. PSIPRED accepts an input sequence in a web form, and emails the results back to the user.

We have developed a plugin in SeqVISTA to recognize the results that PSIPRED emails to the user. Predicted α-helices and β-strands are displayed as features in the graphics panel of SeqVISTA. The confidence scores are plotted in a separate bar graph. Once again, selecting either an α-helix or a β-strand will then highlight the corresponding sequence region. Figure 4 illustrates the results of PSIPRED on cow phosphodiesterase.

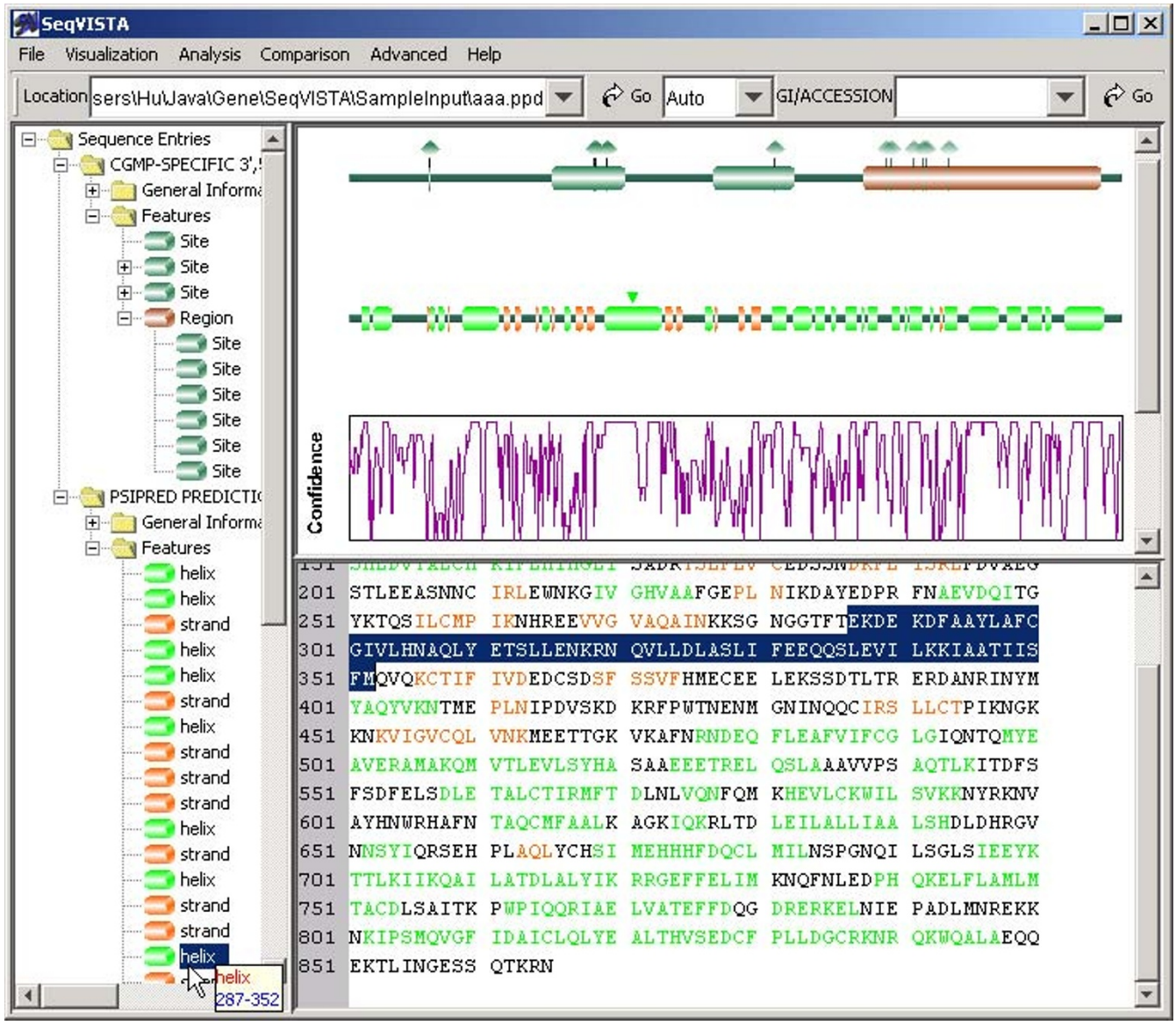

Figure 3
Results of PSIPRED on the cow phosphodiesterase protein.

Cister

Cister is a program that accepts a genomic sequence and a set of cis-element weight matrices and computes the locations of cis-elements and their clusters in the genomic sequence [11]. The outputs of Cister include the scores and locations of cis-elements, and a graph, which plots the probability that a position of the input sequence is in a cis-element cluster. It is helpful for the user of Cister to compare the predicted cis-element locations with the GenBank record of the input sequence, which could contain the experimentally determined promoter region, as well as known cis-elements. Such a comparison would still be valuable even if the promoter region or cis-element locations are not included in the GenBank record, since knowing regions such as exons or repeats could help in analyzing the Cister output.

We have developed a SeqVISTA plugin to interpret the output files of Cister. A user can then add a Cister output while visualizing the GenBank record of the sequence in SeqVISTA. Figure 4 illustrates the Cister output of the human SV40 virus genome. Three figures are added to the

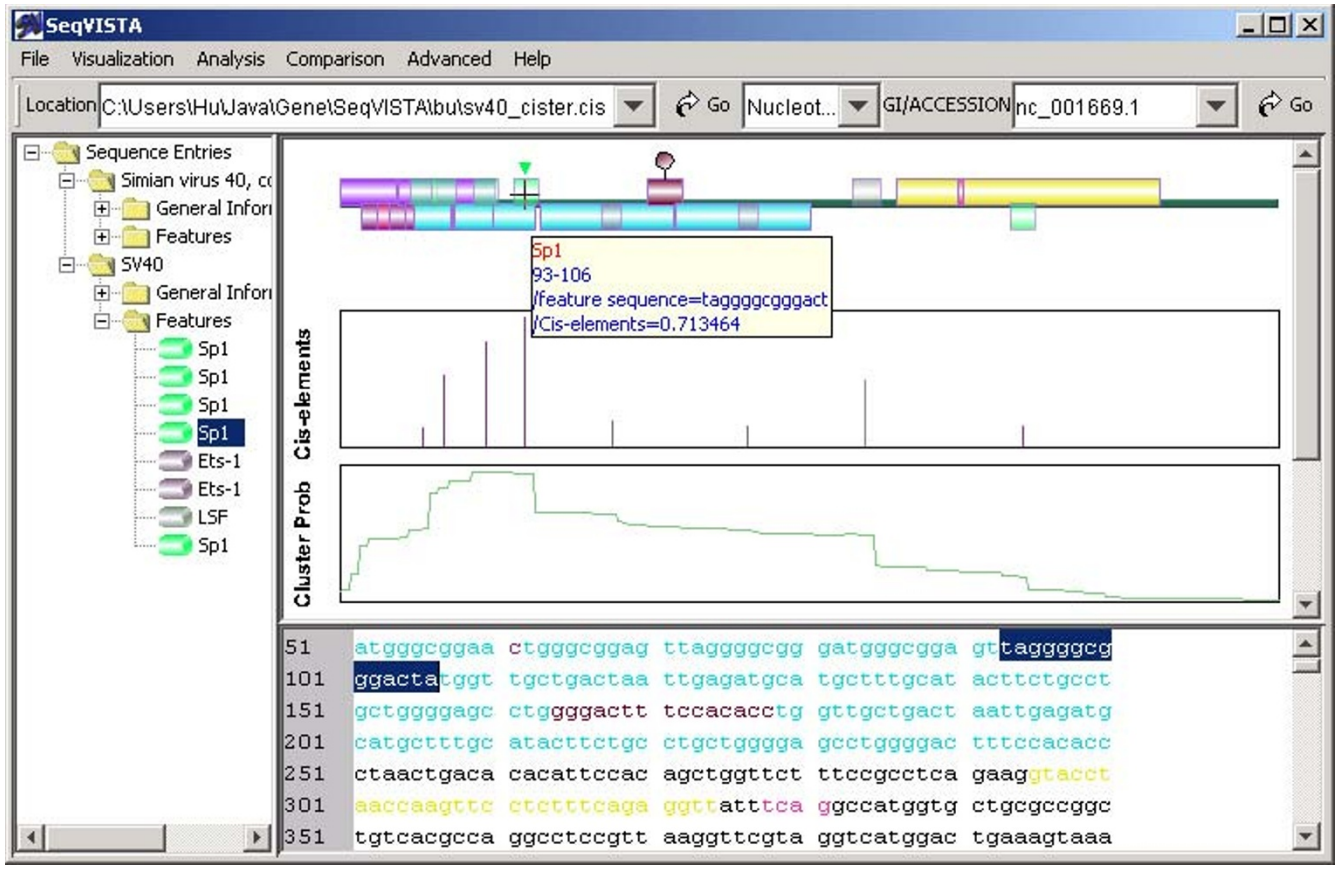

Figure 4
Results of Cister on the SV40 genome

graphics panel of SeqVISTA. The top figure includes the predicted locations of three cis-element types (LSF, Sp1 and Ets-1) in the SV40 genome. These predicted cis-elements are treated as typical sequence features; therefore, they are included in the tree panel as well as linked to the sequence panel. The middle bar graph in the graphics panel indicates the scores of predicted cis-elements. The bottom figure plots the probability that each base in the SV40 genome is in a cluster composed of these four cis-element types, judged according to the strengths of individual cis-elements and their local concentration [11]. By juxtaposing the Cister output and the GenBank annotation of the same sequence, the user can easily examine the context of the Cister predictions and select the most plausible ones for experimental testing.

Summary

We have developed a sequence visualization tool called SeqVISTA. It focuses on the detailed base-by-base or residue-by-residue level of a sequence and its annotations. Our first goal is to enable a user to grasp the most salient features of a sequence at a glance, and extract the corresponding bases or residues precisely and painlessly. To this end, we have made a conscious effort to make the user interface of SeqVISTA simple, intuitive and coherent. While searching GenBank, the user can load a record into SeqVISTA by a single click. The user can also save all contents of a SeqVISTA window as publication quality images. Our second goal is to establish SeqVISTA as a general platform for visualizing the results of sequence analysis software, as well as for comparing these results to the annotations of the same sequence. We have devised three ways to achieve this goal: common file format, parser plugin and direct query of software with a web server. SeqVISTA is written in Java and has been extensively tested on Windows and Linux. It should run on any operating system with a Java 1.4 virtual machine. It is freely available to academic users at http://zlab.bu.edu/SeqVISTA.

Authors' contributions

ZH created SeqVISTA and performed all programming work. MF and ZW tested SeqVISTA extensively, and sug-

gested many functions of the program. TN participated in the earlier development of the program. ZW oversaw this study. All authors have read and approved the final manuscript.

Web Site References

http://genome.ucsc.edu, the UCSC genome browser.

http://www.ensembl.org, the Ensembl genome viewer.

http://www.ncbi.nlm.nih.gov, Entrez map viewer at NCBI.

http://www-gsd.lbl.gov/vista/, VISTA (visualization tools for alignments)

http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker, RepeatMasker

http://bioinf.cs.ucl.ac.uk/psipred/, PSIPRED

http://zlab.bu.edu/~mfrith/cister.shtml, Cister

Acknowledgements

We thank Kevin Wiehe for proof reading the manuscript. This work was funded by NSF grant DBI-0078194 and NIH grant 1P20GM066401-01.

References

1. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I and Clamp M **The Ensembl genome database project** *Nucleic Acids Res* 2002, **30:**38
2. WJ Kent, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996
3. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS and Dubchak I **VISTA : visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16:**1046
4. Kashuk C, SenGupta S, Eichler E and Chakravarti A **ViewGene: a graphical tool for polymorphism visualization and characterization.** *Genome Res* 2002, **12:**333
5. Burge C and Karlin S **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78
6. Lukashin AV and Borodovsky M **GeneMark.hmm: new solutions for gene finding.** *Nucleic Acids Res* 1998, **26:**1107
7. Reese MG, Kulp D, Tammana H and Haussler D **Genie – gene finding in Drosophila melanogaster.** *Genome Res* 2000, **10:**529
8. Delcher AL, Harmon D, Kasif S, White O and Salzberg SL **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27:**4636
9. Scherf M, Klingenhoff A and Werner T **Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach.** *J Mol Biol* 2000, **297:**599
10. Quandt K, Frech K, Karas H, Wingender E and Werner T **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23:**4878
11. Frith MC, Hansen U and Weng Z **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17:**878
12. Frith MC, Spouge JL, Hansen U and Weng Z **Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences.** *Nucleic Acids Res* 2002, **30:**3214
13. Smit AFA and Green P **personal communication.** *RepeatMasker*
14. Benson G **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573
15. Rost B and Sander C **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232:**584
16. Cuff JA, Clamp ME, Siddiqui AS, Finlay M and Barton GJ **JPred: a consensus secondary structure prediction server.** *Bioinformatics* 1998, **14:**892
17. Jones DT **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292:**195
18. Jones DT, Taylor WR and Thornton JM **A model recognition approach to the prediction of all-helical membrane protein structure and topology.** *Biochemistry* 1994, **33:**3038
19. Moller S, Croning MD and Apweiler R **Evaluation of methods for the prediction of membrane spanning regions** *Bioinformatics* 2001, **17:**646
20. Edwards YJK, Frith M, Elgar G and Bishop MJ **Theatre: a novel tool for the comparative investigation and display of evolutionary diversity of functional and structural features in DNA sequences.** *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure* 1998,
21. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA and Wheeler DL **GenBank.** *Nucleic Acids Res* 2002, **30:**17
22. Bairoch A and Apweiler R **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45