BMC Bioinformatics

BioMed Central

Software

Genomorama: genome visualization and analysis

Jason D Gans* and Murray Wolinsky

Address: Biosciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA

Email: Jason D Gans* - jgans@lanl.gov; Murray Wolinsky - murray@lanl.gov

* Corresponding author

Published: 14 June 2007

BMC Bioinformatics 2007, **8**:204 doi:10.1186/1471-2105-8-204

Received: 4 January 2007

Accepted: 14 June 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/204

Abstract

Background: The ability to visualize genomic features and design experimental assays that can target specific regions of a genome is essential for modern biology. To assist in these tasks, we present Genomorama, a software program for interactively displaying multiple genomes and identifying potential DNA hybridization sites for assay design.

Results: Useful features of Genomorama include genome search by DNA hybridization (probe binding and PCR amplification), efficient multi-scale display and manipulation of multiple genomes, support for many genome file types and the ability to search for and retrieve data from the National Center for Biotechnology Information (NCBI) Entrez server.

Conclusion: Genomorama provides an efficient computational platform for visualizing and analyzing multiple genomes.

Background

With the rapid growth in the number of sequenced genomes has come a corresponding proliferation of computational tools for viewing, comparing and searching genome sequences and annotations. Tools can be divided into two broad categories [1], database-client and stand-alone. In general, database-client tools offer static (or semi-static) visualizations of small sets of predefined genomes, while stand-alone tools allow interactive visualizations of locally stored genomes. Stand-alone tools can serve as graphical front ends for displaying the output of locally run calculations. A high level comparison of common features for these stand-alone tools [2-19] is shown in Table 1 and reveals several trends and patterns. Almost all of the tools are implemented in an interpreted language (i.e. Java, Perl, Tcl/Tk). While this provides for cross platform portability, the responsiveness (i.e. rendering speed, file loading speed) of these applications is poor. While all of the tools can display genome annotations, additional functionalities (i.e. sequence and annotation based searching, multiple sequence alignment, annotation editing, etc.) vary widely between programs.

Not content with the performance or feature set of existing programs, we wrote Genomorama, a stand-alone tool originally developed to assist in computational signature design for bacterial and viral pathogen detection. Genomorama allows users designing DNA-based hybridization assays, such as PCR or DNA probes, to easily identify the regions of a genome targeted by a given assay. It is distinguished from existing tools by DNA hybridization-based sequence searching, its rapid execution speed, and ability to read and export a diverse set of common file formats. Despite its origins as a viewer for viral and bacterial genomes, Genomorama can also visualize large eukaryotic genomes (e.g. human chromosomes).

Table 1: Comparing features of freely available, stand-alone genome viewers

Program	Platforms[a]	Input formats[b]	Graphic output formats[c]	Source code available	Circular view	Linear view	Real time navigation	Multiple genomes	Annotation editing and creation	Annotation searching	Sequence searching
Apollo [2]	Java	GAME XML, GFF, GBK, EMBL, FASTA	PS	✔		✔	✔		✔	✔	✔
Argo [3]	Java	GFF, GBK, GENSCAN, BLAST	Printer			✔	✔	✔ (2)	✔	✔	✔
Artemis [4]	Java	EMBL, GBK, FASTA, GFF	JPG, PNG	✔		✔	✔	✔ (via ACT)	✔	✔	✔
Bluejay [5]	Java	XML	Printer, SVG		✔		✔		✔		
CGView [6]	Java	PTT, XML	PNG, JPG, SVG	✔	✔						
DNAvis [7]	Windows, Linux	GFF, FASTA				✔	✔	✔			
GATA [8]	Java	GFF	PNG	✔		✔	✔	✔ (2)			
GeneViTo [9]	Java	PTT+FFN+FNA	JPG		✔	✔			✔	✔	✔
GenoMap [10]	Tcl/Tk	GRS	PS	✔	✔						
Genome2D [11]	Windows	GBK, FASTA, GLIMMER, PARADOX	Printer, WMF, BMP			✔	✔	✔	✔	✔	✔
GenomeComp [12]	Perl/Tk	EMBL, GBK, FASTA	PS	✔		✔	✔	✔ (2)		✔	
GenomePlot [13]	Tcl/Tk/Perl	tab delimited	PS, GIF, TIFF, JPG	✔	✔	✔	✔				
GenomeViz [14]	Tcl/Tk/Perl (no Windows)	tab delimited	PS	✔	✔		✔	✔		✔	
Genome Workbench [15]	OS X, Windows, Linux	ASN.1, XML, FASTA, GFF		✔		✔	✔		✔	✔	
Genomorama	OS X, Windows, Linux	EMBL, GBK, ASN.1, FASTA, PTT	PS, GIF	✔		✔	✔	✔	✔	✔	✔
IGB [16]	Java	GFF, FASTA, PSL, DAS	Printer	✔		✔	✔		✔	✔	✔
Mauve [17]	Java	GBK, FASTA, SEQ	PNG, JPG	✔		✔	✔	✔			
SeqVISTA [18]	Java	EMBL, FASTA	JPG	✔		✔	✔	✔	✔		✔
Sockeye [19]	Java	EMBL (via server), GFF	JPG			✔	✔	✔			

[a]Programs that use Java, Tcl/Tk and Perl are expected to run on any operating system. [b]Common file formats include the GenBank flat file (GBK), EMBL flat file (EMBL), nucleic acid sequence file (FASTA), general feature format (GFF) and protein table file (PTT). A complete list of genome annotation file formats can be found on the Genomorama project webpage. [c]The graphic output format labeled "Printer" indicates direct output to an attached printer.

Implementation

Genomorama is a software program for interactively displaying and analyzing multiple genomes. It provides a powerful yet easy to use interface that leverages the visualization power of modern computers (via OpenGL) and the substantial bioinformatic infrastructure provided by the NCBI (via the NCBI C toolkit). Genomorama is written in portable, highly optimized C++ and comes in three "flavors" that allow it to run natively on (most) modern operating systems: OS X (using Carbon), Microsoft Windows (using the Microsoft Foundation Classes) and Linux (using Motif). The Motif version allows any X-windows client that supports OpenGL to remotely run Genomorama. Executables and source code are freely provided for all flavors.

Results and discussion

To visualize and compare annotated genome features at all relevant size scales, genomes are displayed on the computer screen as linear, scale-dependent maps. The user interacts with a map using the mouse, keyboard and scroll bars. Semantic zooming [20] is used to display genomic features which occur at a wide range of scales, i.e. $\sim 10^5$ bases for a mammalian gene, $\sim 10^4$ bases for a pathogenicity island, $\sim 10^3$ bases for a bacterial gene, $\sim 10^2$ bases for a tRNA, $\sim 10^1$ bases for a transcription factor binding site and 10^0 for a single nucleotide polymorphism. Optional 2D graphs, including %G+C, GC skew (automatically computed from the genome sequence) and external data sets (provided by the user in a separate file), can be superimposed on genome maps. Publication quality, WYSIWYG ("What You See Is What You Get") images can be saved in either GIF or PostScript formats.

Genome annotations and sequences are available in a large number of file formats and Genomorama can read a substantial subset of these formats, including GenBank (GBK), European Molecular Biology Laboratory (EMBL), Abstract Syntax Notation One (ASN.1), Protein Table (PTT) and FASTA. Unlike existing programs, Genomorama can read the multi-part GBK, EMBL and ASN.1 files used to store annotations and sequence for partially assembled sequences for both prokaryotic and eukaryotic organisms. The ability to load multipart annotation files allows access to preliminary annotation information provided by sequencing centers during the whole genome shotgun sequencing of an organism (these files are available from the NCBI ftp site [21]). A screen shot of five contigs and associated sequencing quality scores from the genome *Sphingopyxis alaskensis* RB2256 is shown in Figure 1.

Genomorama can load large (> 10^8 bases) genomes. Support for large genomes is crucial for visualizing entire eukaryotic chromosomes. A comparison between loading times for Genomorama and two Java-based visualization tools is shown in Figure 2. Conservative memory usage and efficient C++ implementation enable Genomorama to load the sequence and annotations for human chromosome 1 substantially faster (more than an order of magnitude) than either of the Java-based programs on a range of desktop computers.

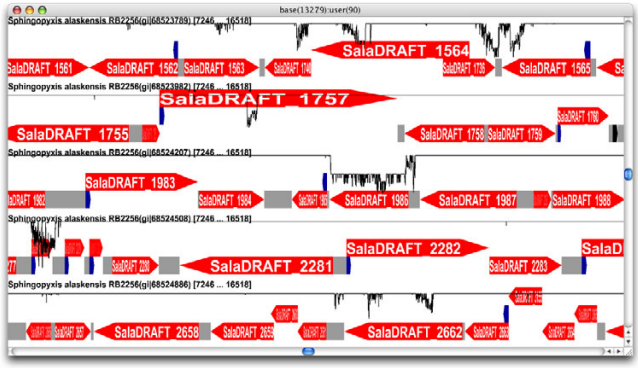

Figure 1
Genomorama can load and display the multiple annotated contigs stored in a whole genome shotgun GBK file. This screen shot shows five contigs from *Sphingopyxis alaskensis* RB2256 (extracted from the NCBI [21] file wgs.AAIP.1.gbff) and the associated sequence quality scores (from the NCBI [21] file wgs.AAIP.1.qscore). Quality scores are proportional to the negative log of the probability that a given base has been incorrectly assigned as an A, T, G or C and are shown as black plots superimposed over each contig track. The value of a quality score for each track is interactively displayed on the menu bar as a user specified score [i.e. "user(90)"] for the annotation track and base currently selected by the cursor.

To assist in experimental design and analysis, Genomorama provides DNA hybridization-based searches to identify probe binding locations and PCR amplification products. Given a pair of PCR primers, Genomorama will display all corresponding PCR amplicons from a target sequence. Both traditional PCR primer and Padlock probe [22] queries are supported. These searches employ a sequence similarity criteria defined by DNA melting temperature [23-28], which allows for non-Watson and Crick base pairing (but currently not gaps or DNA bulges), and an optional number of exact matching bases at the 3' end of each primer. All possible combinations of the forward and reverse PCR primers are tested (i.e. forward-reverse, reverse-forward, forward-forward and reverse-reverse). In contrast, existing in-silico PCR tools are either inflexible (i.e. require a preconfigured server) [29] or rely on heuristic similarity measures (i.e. number of mismatches between primer and template) [30,31].

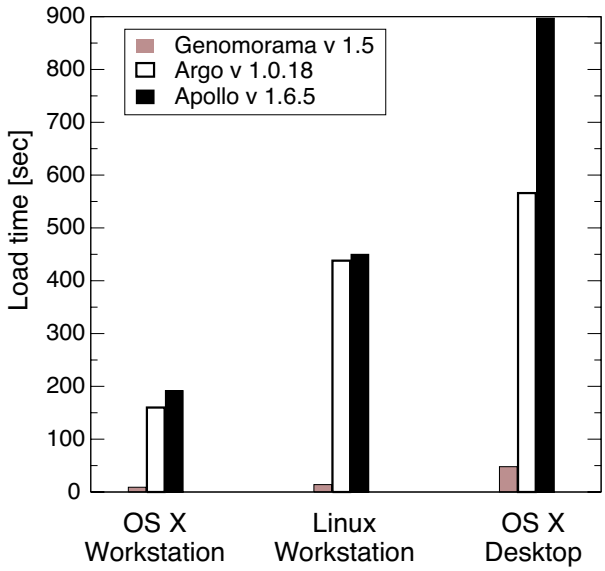

Figure 2
Comparing the time to load human chromosome 1. The time to load *Homo sapiens* chromosome 1 is used to compare the performance of Genomorama and two Java based tools: Apollo [2] and Argo [3]. The time to load the GBK file [GenBank:NC_000001.9] from the local hard drive is shown for three computing platforms: a high-end OS X 10.4.8 workstation (dual 3 Ghz Intel Xeon CPUs, 3 GB ram, Java 1.5.0), a mid-range Linux Red Hat 4.0.1 workstation (dual 2.4 GHz Intel Xeon CPUs, 1 GB ram, Java 1.4.2) and low-end OS X 10.3.9 desktop (single 1.8 GHz G5 PowerPC CPU, 512 MB ram, Java 1.4.2). The Java-based programs were run from the command line with the arguments "-Xms32m -Xmx1024m" to increase the amount of memory allowed to the Java virtual machine. Providing Java with more than 1 GB of memory did not improve performance (results not shown). Each program loaded the genome file twice (to ensure fair OS disk caching) and the second load time is reported. For all platforms, Genomorama loads the genome file more than an order of magnitude faster than either of the Java-based programs.

Genomorama also performs primer prediction by computing all potential forward and reverse PCR primers that satisfy primer length, melting temperature, %G+C and heuristic base composition requirements. An example of PCR primer based searching, using the *B. anthracis* specific primers [32], is shown in Figure 3. Finally, sequence searching (both exact and hybridization based) is sensitive to the topology of the target DNA molecule (i.e. either linear or circular) and, as a result, can identify query matches that span the start/stop (i.e. nucleotide 0) of circular genomes.

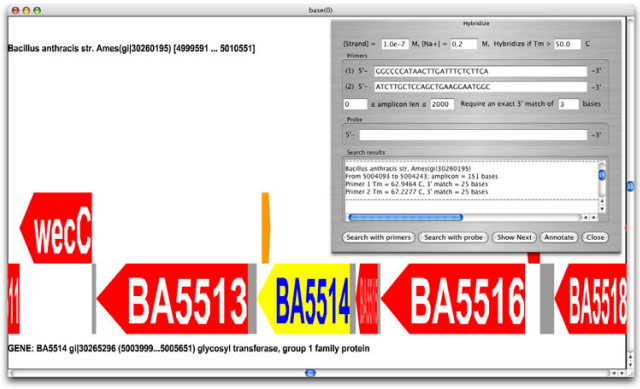

Figure 3
Genomorama supports sequence searching with PCR primers. The genomic neighborhood of the amplicon (shown in orange) produced by the *B. anthracis* [GenBank:NC_003997.3] chromosomal specific PCR primers, M.Ctg032 [32]. The amplicon is contained within a glycosyl transferase (show in yellow). The amplicon annotation was added to the genome by selecting the "annotate" button on the Hybridize dialog box.

Conclusion

Genomorama is an easy to use computational tool for a number of genome comparison tasks, including real time display of multiple genomes, high quality output and novel hybridization based sequence searching.

Availability and requirements

- **Project name:** Genomorama
- **Project homepage:** http://snp.lanl.gov/genomorama
- **Operating systems:** OS X, Windows, Linux
- **Programming language:** C++
- **License:** Freely available
- **Any restrictions on use by non-academics**: No

Authors' contributions

JG wrote the program and documentation. MW oversaw the development process. Both authors prepared and approved the manuscript.

Acknowledgements

This research was supported in part by the DOE/DHS Chemical Biological National Security Program (CBNP), the DOD/USAMRMC Toxin and Virulence Factor Database Effort (MIPR 2MCTC32157) and the Los Alamos National Laboratory Directed Research Development Program (LDRD 20070010DR). The authors would like to thank N. Pawley for helpful discussions and K. Sirotkin and J. Kans for assistance with the NCBI toolkit.

References

1. Loraine AE, Helt GA: **Visualizing the genome: techniques for presenting human genome data and annotations.** *BMC Bioinformatics* 2002, **3(19):**.
2. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME: **Apollo: a sequence annotation editor.** *Genome Biology* 2002, **3(12):**1-14.
3. Engels R, Yu T, Burge C, Mesirov JP, DeCaprio D, Galagan JE: **Combo: a whole genome comparative browser.** *Bioinformatics* 2006, **22(4):**1782-1783.
4. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16(10):**944-945.
5. Turinsky AL, Ah-Seng AC, Gordon PMK, Stromer JN, Taschuk ML, Xu EW, Sensen CW: **Bioinformatics visualization and integration with open standards: The Bluejay genomic browser.** In *Silico Biology* 2004, **5(18):**.
6. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21(4):**537-539.
7. Fiers MW, van de Wetering H, Peeters TH, van Wijk JJ, Nap JP: **DNAVis: interactive visualization of comparative genome annotations.** *Bioinformatics* 2006, **22(3):**354-355.
8. Nix DA, Eisen MB: **GATA: a graphic alignment tool for comparative sequence analysis.** *BMC Bioinformatics* 2005, **6(9):**.
9. Vernikos G, Gkogkas C, Promponas V, Hamodrakas S: **GeneViTo: Visualizing gene-product functional and structural features in genomic datasets.** *BMC Bioinformatics* 2003, **4(1):**.
10. Sato N, Ehira S: **GenoMap, a circular genome data viewer.** *Bioinformatics* 2003, **19(12):**1583-1584.
11. Baerends R, Smits W, de Jong A, Hamoen L, Kok J, Kuipers O: **Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data.** *Genome Biology* 2004, **5(5):**R37.
12. Yang J, Wang J, Yao ZJ, Jin Q, Shen Y, Chen R: **GenomeComp: a visualization tool for microbial genome comparison.** *J Microbiol Methods* 2003, **54(3):**423-426.
13. Gibson R, Smith DR: **Genome visualization made fast and simple.** *Bioinformatics* 2003, **19(11):**1449-1450.
14. Ghai R, Hain T, Chakraborty T: **GenomeViz: visualizing microbial genomes.** *BMC Bioinformatics* 2004, **5(1):**.
15. DiCuccio M, Cherry J, Lebedev V, Shomrat M, Smith R, Tereshkov V, Voronov Y, Yazhuk A: **Genome Workbench.** [http://www.ncbi.nlm.nih.gov/projects/gbench/].
16. Affymetrix: **IGB.** [http://www.affymetrix.com/support/developer/tools/download_igb.affx].
17. Darling A, Mau B, Blattner FR, Perna NT: **Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements.** *Genome Res* 2004, **14(7):**1394-1403.
18. Hu Z, Frith M, Niu T, Weng Z: **SeqVISTA: a graphical tool for sequence feature visualization and comparison.** *BMC Bioinformatics* 2003, **4(1):**.
19. Montgomery SB, Astakhova T, Bilenky M, Birney E, Fu T, Hassel M, Melsopp C, Rak M, Robertson AG, Sleumer M, Siddiqui AS, Jones SJM: **Sockeye: A 3D Environment for Comparative Genomics.** *Genome Res* 2004, **14(5):**956-962.
20. Bederson BB, Hollan JD, Perlin K, Meyer J, Bacon D, Furnas G: **Pad++: A Zoomable Graphical Sketchpad For Exploring Alternate Interface Physics.** *Journal of Visual Languages and Computing* 1995, **7:**3-31.
21. **NCBI ftp site** [ftp://ftp.ncbi.nih.gov/genbank/wgs/]
22. Nilsson M, Banér J, Mendel-Hartvig M, Dahl F, Antson DO, Gullberg M, Landegren U: **Making ends meet in genetic analysis using padlock probes.** *Human Mutation* 2002, **19:**410-415.
23. Allawi HT, SantaLucia J: **Thermodynamics and NMR of Internal G-T Mismatches in DNA.** *Biochemistry* 1997, **36:**10581-10594.
24. Allawi HT, SantaLucia J: **Thermodynamics of internal C-T mismatches in DNA.** *Nucleic Acids Research* 1998, **26(11):**2694-2701.
25. Allawi HT, SantaLucia J: **Nearest Neighbor Thermodynamic Parameters for Internal C-A Mismatches in DNA.** *Biochemistry* 1998, **37:**2170-2179.
26. Allawi HT, SantaLucia J: **Nearest-Neighbor Thermodynamics of Internal A-C Mismatches in DNA: Sequence Dependence and pH Effects.** *Biochemistry* 1998, **37:**9435-9444.
27. Bommarito S, Peyret N, SantaLucia J: **Thermodynamic parameters for DNA sequences with dangling ends.** *Nucleic Acids Research* 2000, **28(9):**1929-1934.
28. Peyret N, Seneviratne PA, Allawi HT, SantaLucia J: **Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches.** *Biochemistry* 1999, **38:**3468-3477.
29. Lexa M, Horak J, Brzobohaty B: **Virtual PCR.** *Bioinformatics* 2001, **17(2):**192-193.
30. Bikandi J, Millán RS, Rementeria A, Garaizar J: **In silico analysis of complete bacterial genomes: PCR, AFLP-PCR and endonuclease restriction.** *Bioinformatics* 2004, **20(5):**798-799.
31. Schuler GD: **Sequence Mapping by Electronic PCR.** *Genome Research* 1997, **7(5):**541-550.
32. Radnedge L, Agron P, Hill K, Jackson P, Ticknor L, Kiem P, Anderson G: **Genome differences that distinguish Bacillus anthracis from Bacillus cereus and Bacillus thuringiensis.** *Applied and Environmental Microbiology* 2003, **69(5):**2755-2764.