

Software

Open Access

GeneOrder3.0: Software for comparing the order of genes in pairs of small bacterial genomes

Srikanth Celamkoti^{†1}, Sashidhara Kundeti^{†1}, Anjan Purkayastha¹, Raja Mazumder², Charles Buck³ and Donald Seto^{*1}

Address: ¹Bioinformatics and Computational Biology, School of Computational Sciences, George Mason University, 10900 University Boulevard, MSN 5B3, Manassas, VA 20110 USA, ²Biochemistry and Molecular Biology Department, Georgetown University School of Medicine, 4000 Reservoir Road, Washington, D.C. 20057 USA and ³Virology Program, American Type Culture Collection (ATCC), 10801 University Boulevard, Manassas, VA 20110, USA

Email: Srikanth Celamkoti - scelamko@gmu.edu; Sashidhara Kundeti - skundeti@gmu.edu; Anjan Purkayastha - purkayas@hotmail.com; Raja Mazumder - rm285@georgetown.edu; Charles Buck - cbuck@atcc.org; Donald Seto* - dseto@gmu.edu

* Corresponding author †Equal contributors

Published: 05 May 2004

Received: 16 January 2004

BMC Bioinformatics 2004, 5:52

Accepted: 05 May 2004

This article is available from: <http://www.biomedcentral.com/1471-2105/5/52>

© 2004 Celamkoti et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: An increasing number of whole viral and bacterial genomes are being sequenced and deposited in public databases. In parallel to the mounting interest in whole genomes, the number of whole genome analyses software tools is also increasing. GeneOrder was originally developed to provide an analysis of genes between two genomes, allowing visualization of gene order and synteny comparisons of any small genomes. It was originally developed for comparing virus, mitochondrion and chloroplast genomes. This is now extended to small bacterial genomes of sizes less than 2 Mb.

Results: GeneOrder3.0 has been developed and validated successfully on several small bacterial genomes (ca. 580 kb to 1.83 Mb) archived in the NCBI GenBank database. It is an updated web-based "on-the-fly" computational tool allowing gene order and synteny comparisons of any two small bacterial genomes. Analyses of several bacterial genomes show that a large amount of gene and genome re-arrangement occurs, as seen with earlier DNA software tools. This can be displayed at the protein level using GeneOrder3.0. Whole genome alignments of genes are presented in both a table and a dot plot. This allows the detection of evolutionary more distant relationships since protein sequences are more conserved than DNA sequences.

Conclusions: GeneOrder3.0 allows researchers to perform comparative analysis of gene order and synteny in genomes of sizes up to 2 Mb "on-the-fly." Availability: <http://binf.gmu.edu/genometools.html> and <http://pasteur.atcc.org:8050/GeneOrder3.0>.

Background

Whole genome DNA sequence data are being generated and deposited into proprietary and public domain databases at an increasingly rapid rate as a result of more efficient and less expensive DNA sequencing and analysis technology. An on-line database, <http://www.genomes>

[online.org](http://www.genomes)[1], lists a running count of 182 published complete genomes with 488 on-going prokaryotic genome projects as of March 26, 2004.

There are relatively few whole genome comparison software tools owing to the previous paucity of whole genome

data. However, there are a growing number of these tools available, and along with earlier versions of GeneOrder and CoreGenes [2-4], there is increasing interest in developing software comparing the gene order and synteny of whole genomes at the gene level, as evidenced by two recent publications [5,6].

As genome sequencing projects are completed, greater scientific curiosity and attention focuses on the analysis and data mining of these whole genomes. In turn, this stimulates additional interest in obtaining and analyzing more and diverse genomes, as seen currently with the increase in software tools development. As a necessary complement to understanding these large data sets, bioinformatics and whole genome informatics must provide software tools that allow mining of gene and putative gene data at this whole genome level [7-13]. Key software challenges include developing platform-independent, interactive and user-friendly tools that are flexible and useful for research. These are either currently being developed or are available to perform comparative genomics at different levels of biological complexity.

There are several genome analysis tools and database extant. Software such as STRING (EMBL; <http://string.embl.de/>) [10] and KEGG (Kyoto University Bioinformatics Centre; <http://www.genome.ad.jp/kegg/>) [11] are web resources and databases that offer information on proteins, as well as provide tools for the analysis of proteins at the functional level. They also offer useful genome comparison tools.

STRING is a database comprising orthologous proteins and predicted functional interactions between proteins in prokaryotes. It identifies functional interactions from various lines of evidence, both computational and experimental. However, STRING does not perform gene order analysis. KEGG is a database comprising various levels of biological information. At the lowest level, it contains information on gene annotation, orthology and whole genome alignments. At the highest levels, it has information on metabolic and functional pathways. KEGG provides pre-computing archived whole genome alignments. Unfortunately, being a pre-computed resource, KEGG is of limited use when performing comparative genomics with a novel or proprietary prokaryotic genomic sequence.

A number of whole genome sequence alignment tools have also been developed over the past few years. Noteworthy among them are MUMmer (TIGR) and AVID (LNBL) [12,13]. MUMmer <http://www.tigr.org/software/mummer/> performs a pair-wise alignment of two large and closely related nucleic acid sequences. It is based on a suffix tree alignment algorithm, which is a departure from

the traditional hashing-based 'generate-and-test' approach. Another tool in this package PROmer performs amino acid alignments of genomic sequences, after a six-frame translation. This is applicable for genome sequences that are too divergent to align at the nucleotide level. Synteny may be inferred from the aligned sequences.

AVID is a global alignment tool, designed to align large genomic regions. The VISTA suite of comparative genomics tools <http://www-gsd.lbl.gov/vista/index.shtml> uses AVID as the alignment program for a range of comparative studies. While both MUMmer and AVID display synteny at the nucleotide and/or the translated amino acid level, they do not display the arrangement of syntenic genes in the aligned genomes. There is an alignment software tool, ACT, which does generate maps (Sanger Centre; <http://www.sanger.ac.uk/Software/ACT/>). It provides alignments of multiple annotated genomic sequences, however the display graphics is complicated. This output is difficult to navigate.

GeneOrder3.0 occupies a unique niche among these whole genome computational tools. It performs a whole genome comparison of two prokaryotes at the level of the coding sequence, without any assumptions of phylogenetic relatedness. The alignment is displayed as the relative arrangement of similar coding sequences between the two species, allowing for easy identification of regions of synteny. All alignments are computed "on-the-fly," thus providing flexibility in running this software, for example, newly acquired genome sequences not yet archived or proprietary sequences not archived in public databases. Plans are in place for storing pre-computed alignments for immediate returns upon subsequent requests. Moreover, GeneOrder3.0 is easy to use for a typical molecular biologist with limited computational experience.

Very recently, reports on two other gene order and synteny identification software have been published [5,6]. These two software programs, GOV ("Gene Order Visualizer;" <http://kishore2.bii.a-star.edu.sg/Gov.html> (not currently accessible 05/17/04) and GeneSyn ftp://159.149.110.11/pub/GeneSyn_1.0), are reported to identify syntenic genes between two organisms of interest. The GOV software [5] is a web-based site that was inaccessible during the course of this evaluation process. GeneSyn [6] is a Unix-based software that is resident on the user's workstation. It compares the arrangements of genes in multiple genomes; however, it does not include a graphical display. It also requires files to be in a format that is not widely used. Furthermore it is not immediately clear how the genes are compared to one another to identify synteny. Both have deficiencies relative to GeneOrder- both current and past versions.

In light of the recent publications of whole genome comparative tools, it is clear that one increasingly important approach in comparative genomics is to identify common and related genes (orthologs, paralogs and metalogs) in the organisms, and to study their locations and physical relationships to each other within the genomes. Gene order and synteny are important components of this whole genome examination. These were topics at a workshop on "Gene order dynamics, comparative maps and multigene families" (Sainte-Adele, Quebec on Sept. 22-25, 2000).

Due to the initial expenses and limitations of the emerging sequencing technology, small genomes were necessarily the first to be compiled and analyzed. These genomes enabled studies which included comparing gene order among fully sequenced genomes of viruses, mitochondria and chloroplasts [14]. Early versions of GeneOrder and CoreGenes were developed to examine these genomes, with input limitations of about 350,000 bases [2-4]. It was understood that larger genomes, particularly the bacterial genomes, are of more interest to the general research community. In response to this, GeneOrder3.0 has been developed to allow researchers to compare the order of genes between two small bacterial genomes up to 2 Mb. Technical difficulties have precluded the application to larger genomes. These are being addressed to allow GeneOrder analyses of 4+Mb genomes. GeneOrder3.0 is a "stepping-stone" that has been tested on full genome sequence data for available small bacterial genomes that are obtainable from the NCBI GenBank database. Its continued application will generate data and results, and elicit comments and suggestions for enhancements from the genome and microbial research communities as part of the evolving nature of software and research tools.

Implementation

GeneOrder3.0 relies on the BLAST sequence alignment algorithm [12]. The client-server architecture has been implemented in J2EE technologies (Servlets and Applets), JavaScript and HTML. It uses the 'setdb' and 'BLASTP' programs from WU-BLAST <http://BLAST.wustl.edu>. This software is platform-independent; however, up-to-date versions of web browsers are required in order to run the JavaScript, especially to visualize the graph applet (Ptolemy II). Running times may vary; as an example, a Macintosh Powerbook G4 running Netscape v7.2 analyzed a pair of *Mycoplasma* genomes in 8 minutes and a pair of *Hemophilus* genomes in 37 minutes. This is also a function of genome downloading from GenBank. One important caveat: GenBank accession numbers MUST start with "NC_"

Results and discussion

GeneOrder3.0 presents "on-the-fly" genome comparisons between two organisms, both as a two-dimensional dot blot graph and as a table, listing the order of similar genes. It is based on BLAST [15]. Hyperlinks are provided from the table, leading to its corresponding "Protein Coding Region" entry in the NCBI database. BLAST scores are collected and parsed into either default or user specified ranges, available in three ranges designated 'highest', 'high' and 'low' values. These high score parameters have been previously discussed in detail [7]. Briefly, these scores are derived from the BLAST comparisons; as an example, gene number 5 of ovine adenovirus OAV287 (U40839) aligns with gene number 5 of human adenovirus HAV12 (X73487) with a BLAST score of 737 (probability of matching by chance of 4.3×10^{-77}). The default scores are set at [200+] for 'highest,' [100-200] for 'high' and [75-100] for 'low.' The three levels allow the user to identify immediately the best matches as well as probable and possible matches. The low scores matches may not be true matches, but could warrant additional analysis.

GeneOrder3.0 allows for the identification and visualization of gene and genomic re-arrangements in small bacterial genomes. Gene shuffling has been detected by DNA comparison methods on whole genomes. However, GeneOrder allows this comparison to be performed on a protein level. This allows the detection of evolutionary more distant relationships, since protein sequences are more conserved than DNA sequences. In essence, it allows for the alignment and comparison of genomes for which the nucleotide sequences are too divergent to detect similarity. This is a major advantage of GeneOrder.

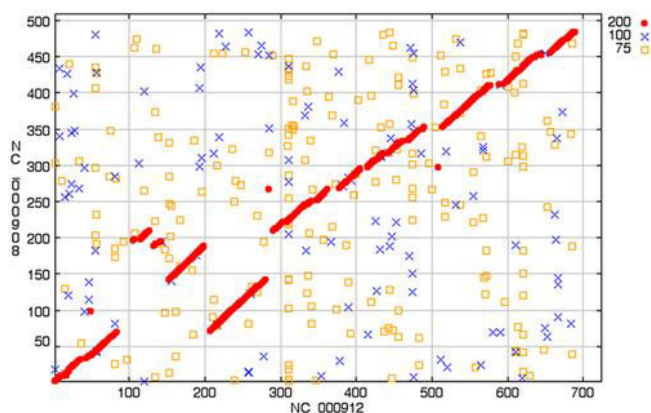


Figure 1
GeneOrder3.0 analysis of *Mycoplasma genitalium* (NC_000908) versus *Mycoplasma pneumoniae* (NC_000912). BLAST scores are set at the defaults of 200, 100 and 75 as represented by the dots, crosses and open squares, respectively.

Applications

The smallest genomes in the bacterial world belong to the genus *Mycoplasma*. These bacteria are distinguished by their lack of a cell wall and by their small genome sizes. Sequenced *Mycoplasma* genomes range in size from ca. 580 kb to 1.35 Mb. Their genomes encode far fewer genes than larger "free-living" bacteria (e.g., *Escherichia coli*). This and other features of the genus make these smaller genomes an intriguing model for early comparative genomic studies and for testing whole genome comparison software, e.g., as a 'stepping stone' to the development of software to analyze much larger genomes. Therefore, for validation, GeneOrder3.0 has been used to analyze pairs of the following sequenced genomes: *Mycoplasma genitalium* (580,074 bases, NC_000908), *M. pneumoniae* (816,394 bases, NC_000912), *M. pulmonis* (963,879 bases, NC_002771), *M. penetrans* (1,358,633 bases, NC_004432) and *M. gallisepticum* (996,442 bases, NC_004829). Additionally two other small genome bacteria were also examined: *Ureaplasma Urelyticum* (751,719 bases, NC 002162) and *Tropheryma whippelii str. Twist* (927,303 bases, NC_004572). GeneOrder3.0 has been used to compare *H. influenzae* (1,830,136 bases, NC_000907) to *H. ducreyi* (1,698,955 bases, NC_002940).

Figure 1 shows an analysis of *M. genitalium* with *M. pneumoniae*. From this display of the relative arrangements of coding sequences we can easily see that gene order and synteny have been largely maintained between the two organisms. Portions of the genomes (specifically genes from '100–300' region of the organisms) have been translocated. There are two large gaps in the diagonal at this region; a perfect diagonal represents perfect synteny. These gaps appear to be a result of translocations of segments of the chromosome. Individual genes that are highly conserved but nevertheless have moved are also seen as red dots (BLAST score 200+) away from the diagonal. In addition, there are several crosses and squares (representing BLAST score ranges of '100–200' and '75–100,' respectively) that putatively represent fast evolving genes, and therefore may be involved in specialized or adaptive functions such as pathogenesis. In addition, genes aligned horizontally or perpendicularly may represent potentially, but not necessarily, related genes. A careful and detailed analysis of the graph generated by GeneOrder3.0 reveals several genes that can be investigated further to understand better the species-specific differences of the two organisms. This allows the researcher to home in on potential reasons of specialization and adaptations.

As with all software and research tools, further work remains to allow more flexibility and usefulness. For example, plans include the important aspect of re-developing the algorithm and software coding to allow efficient

and rapid analysis of larger genomes (ca. 4+Mb), currently limited by technical non-trivial difficulties. A database of pre-computed data will be made available so that subsequent identical requests are filled immediately. An alternative to the graph applet will allow users (1) to link each point to a GenBank entry, current available as links from the table, (2) to print the graph directly and (3) to provide "mouse whispers" that allow a brief notation of the gene under the cursor on screen.

Conclusions

GeneOrder3.0 is a web-accessible whole genome comparative tool that provides gene order and synteny visualization for a pair of genomes of sizes up to 2 Mb "on-the-fly." The current GeneOrder3.0 input limitation of ca. 2 Mb genomes is a "stepping stone" in developing the algorithm to accommodate the analyses of more interesting and complicated larger prokaryotic genomes, ca. 4+Mb. Mounting interest in the larger bacterial genomes and their analyses is illustrated in the recent bacterial genome comparisons publications [16-18] and by the accumulation of larger numbers of these genomes. The growing whole genomes database requires the development of whole genome analyses tools, including alternative versions of similar tools for versatility and availability. GeneOrder3.0 is such a tool.

Availability and requirements

GeneOrder3.0 is a platform-independent, web-based application. GenBank accession numbers or proprietary lists of contiguous genes in a FASTA format are required for input. The site is available at <http://binf.gmu.edu/genometools.html> and at <http://pasteur.atcc.org:8050/GeneOrder3.0>. GeneOrder3.0 software requires a web-browser that supports Javascript.

Authors' contributions

SC and SK provided software development and testing of the software, as well as technical support and maintenance of the site. AP provided supervision and genome validation; RM assisted in the development of the algorithm and interpretation of the GeneOrder plots. CB provided biological direction and validation of the tool. DS conceived of this project, and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgements

DS dedicates this work to the memory of Dante Ferrini (March 26, 1915 – January 26, 2004). JAVA-based programming is coded using "Jakarta-Tomcat," which is generously made available by the Apache Software foundation <http://www.apache.org>. Ptolemy II is available through the Regents of the University of California <http://ptolemy.eecs.berkeley.edu/ptolemyII>. We thank Dr. Jennifer Weller for critical comments on this manuscript. This work has been supported in part by a grant from ATCC and a Faculty

Development Award from the Office of the Provost (George Mason University).

References

1. Bernal A, Ear U, Kyrpides N: **Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.** *Nucleic Acids Res* 2001, **29**:126-127.
2. Mazumder R, Kolaskar A, Seto D: **GeneOrder: comparing the order of genes in small genomes.** *Bioinformatics* 2001, **17**:162-166.
3. Zafar N, Mazumder R, Seto D: **Comparisons of gene colinearity in genomes using GeneOrder2.0.** *Trends Biochem Sci* 2001, **26**:514-516.
4. Zafar N, Mazumder R, Seto D: **CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes.** *BMC Bioinformatics* 2002, **3**:12.
5. Sakharkar KR, Chow VT: **Exploring genome architecture through GOV: a www-based Gene Order Visualizer.** *Bioinformatics* 2004, **20**:984-5.
6. Pavesi G, Mauri G, Jannelli F, Gissi C, G P: **GeneSyn: a tool for detecting conserved gene order across genomes.** *Bioinformatics* in press.
7. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W: **PipMaker – a web server for aligning two genomic DNA sequences.** *Genome Res* 2000, **10**:577-586.
8. Cummings L, Riley L, Black L, Souvorov A, Resenchuk S, Dondoshansky I, Tatusova T: **Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes.** *FEMS Microbiol Lett* 2002, **216**:133-138.
9. Xie T, Hood L: **ACGT – a comparative genomics tool.** *Bioinformatics* 2003, **19**:1039-1040.
10. vonMering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** *Nuc Ac Res* 2003, **31**:258-261.
11. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nuc Ac Res* 2004, **32**Database issue:D277-280.
12. Bray N, Dubchak I, Pachter L: **AVID: a global alignment program.** *Gen Res* 2003, **13**:97-102.
13. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Gen Biol* 2004, **5**:R12. Epub Jan 30.
14. Boore JL, Brown WM: **Big trees from little genomes: mitochondrial gene order as a phylogenetic tool.** *Curr Opin Genet Dev* 1998, **8**:668-674.
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
16. Bhattacharyya A, Stilwagen S, Ivanova N, D'Souza M, Bernal A, Lykidis A, Kapatral V, Anderson I, Larsen N, Los T, Reznik G, Selkov E Jr, Walunas TL, Feil H, Feil WS, Purcell A, Lassez JL, Hawkins TL, Haselkorn R, Overbeek R, Predki PF, Kyrpides NC: **Whole-genome comparative analysis of three phytopathogenic Xylella fastidiosa strains.** *Proc Natl Acad Sci U S A* 2002, **99**:12403-12408.
17. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes.** *Science* 2002, **298**:1616-1620.
18. Ivanova N, Sorokin A, Anderson I, Galleron N, Candelon B, Kapatral V, Bhattacharyya A, Reznik G, Mikhailova N, Lapidus A, Chu L, Mazur M, Goltsman E, Larsen N, D'Souza M, Walunas T, Grechkin Y, Pusch G, Haselkorn R, Fonstein M, Ehrlich SD, Overbeek R, Kyrpides N: **Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis.** *Nature* 2003, **423**:87-91.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

