BMC Bioinformatics

Methodology article

CoreGenes: A computational tool for identifying and cataloging "core" genes in a set of small genomes

Nikhat Zafar[1], Raja Mazumder[1] and Donald Seto*[1,2]

Address: [1]School of Computational Sciences, George Mason University, 10900 University Boulevard, MSN 4E3, Manassas, VA 20110 USA and [2]Center for Biomedical Genomics and Informatics, College of Arts and Sciences, George Mason University, 10900 University Boulevard, MSN 4E3, Manassas, VA 20110 USA

E-mail: Nikhat Zafar - nikhatzafar@yahoo.com; Raja Mazumder - rmazumde@yahoo.com; Donald Seto* - dseto@gmu.edu

*Corresponding author

Published: 24 April 2002

BMC Bioinformatics 2002, **3**:12

Received: 27 November 2001
Accepted: 24 April 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/12

Abstract

Background: Improvements in DNA sequencing technology and methodology have led to the rapid expansion of databases comprising DNA sequence, gene and genome data. Lower operational costs and heightened interest resulting from initial intriguing novel discoveries from genomics are also contributing to the accumulation of these data sets. A major challenge is to analyze and to mine data from these databases, especially whole genomes. There is a need for computational tools that look globally at genomes for data mining.

Results: CoreGenes is a global JAVA-based interactive data mining tool that identifies and catalogs a "core" set of genes from two to five small whole genomes simultaneously. CoreGenes performs hierarchical and iterative BLASTP analyses using one genome as a reference and another as a query. Subsequent query genomes are compared against each newly generated "consensus." These iterations lead to a matrix comprising related genes from this set of genomes, e. g., viruses, mitochondria and chloroplasts. Currently the software is limited to small genomes on the order of 330 kilobases or less.

Conclusion: A computational tool CoreGenes has been developed to analyze small whole genomes globally. BLAST score-related and putatively essential "core" gene data are displayed as a table with links to GenBank for further data on the genes of interest. This web resource is available at [http://pumpkins.ib3.gmu.edu:8080/CoreGenes] or [http://www.bif.atcc.org/CoreGenes] .

Background

The development of genomics instrumentation, technology and methodology, as well as their integration and deployment in many fields of research, has evolved from producing manageable small streams of DNA sequence data to generating an inundating amount of DNA sequence and whole genome data. This massive amount of raw DNA sequence data can be described simply and aptly as a "tsunami" – a tremendous and unexpected wave. An unprecedented wave can be either overwhelming or overwhelmed, depending upon the preparedness of investigators. Preparations include having available or developing appropriate computational tools. One particular area of continuing concern is the ability to separate interesting and relevant data from "noise." This process is known as data mining and is enhanced by the development of effec-

tive and "user-friendly" bioinformatics tools and computational methods [1–10].

Many researchers have been interested in studying individual proteins, identifying single genes and characterizing putative genes, *i.e.*, "open reading frames." There are several sets of bioinformatics tools that allow researchers to characterize DNA sequences, in particular, to locate and define the above single or few gene entries. Examples include the GCG Wisconsin package and the Staden package [9,10]. However, there is also a growing subset of researchers who are interested in "whole genome" analyses [3,7,8]. With respect to these researchers, there are still gaps in the data mining tool set available for the global study of genomes [11]. This paper addresses the critical need for more "whole genome" tools by describing CoreGenes which is intended to expedite the analyses of genomes globally in order to locate, identify and catalog related genes that may constitute a "core" of essential genes common to these related organisms. A matrix of closely related and putatively homologous genes allows a better understanding of the relationships and organization of genomes and their host organisms.

As noted earlier, JAVA-based software allows distributed programming over the Web, providing an interactive graphical user interface (GUI) [2]. Earlier, we have taken advantage of this resource to develop GeneOrder2.0 [8] a tool that allows the determination of gene co-linearity. We have again used this public domain JAVA software resource (Apache Software Foundation) to develop another computational tool. CoreGenes allows for "user-friendly" visualizations of putative gene identity in sets of genomes, related or unrelated, globally from sets of two to five small whole genomes in a GUI environment. The URL address for this software is [http://pumpkins.ib3.gmu.edu:8080/CoreGenes] or [http://www.bif.atcc.org/CoreGenes] .

Computational methods for whole genome studies

Comparative genomics and more specialized fields such as comparative virology, etc., involve the comparison of DNA sequences, genes and genomes [12–14]. Recent rapid data acquisition is allowing the analyses of whole genome sequences, especially the smaller genomes such as mitochondria and chloroplasts [15–17], as well as the larger bacterial genomes [18,19] and large tracts of eukaryotic chromosomes, especially from related organisms [12–14,20–23]. These studies include the determination of the order of genes, *i.e.*, co-linearity [24,25], the location of synteny [26–28] and the identification of clusters of orthologous genes [cog] between two genomes [21–23]. Along similar lines of thought, it should be extremely useful to locate, identify and catalog the sets of "core" genes common to these genomes-genomes which otherwise may be related or semi-related or unrelated in other respects. These global views allow for a deeper understanding of one organism in the context of another, especially in regards to their genomic contents. In addition, the comparison of multiple genomes and the identification of related genes and "core" genes can lead to insight into the structure and function of genes and genomes [4]. This is very useful in genome annotations and also in the identification and characterization of functions for "newly found" putative genes.

Identification of "core" genes from small whole genomes is useful and complements other data derived from these genomes. Small genomes include those from viruses [3], mitochondria [14,15] and chloroplasts [16]. The increasing importance of the large amount of DNA sequence data recently collected from these small genomes is reflected in the better understanding of their biology [3,4,12–14] and in the upsurge of publications analyzing these genomes and the organisms to which they belong [15–28]. Genome co-linearity, gene clustering and homolog identification are three global genome analyses which are important in many fields of research, including resolving phylogenetic and evolutionary relationships [15–17].

Results and Discussion

Description of CoreGenes

CoreGenes is written in JAVA-based programming incorporating the 'setdb' and 'BLASTP' programs from the WU-BLAST package of Washington University, [http://BLAST.wustl.edu] . The basis of this iterative comparison rests on the BLASTP algorithm [29]. A flowchart of the processes is illustrated in Figure 1. This software allows for the identification, characterization, catalog and visualization of putatively essential "core" genes in sets of two to five genomes in a user-friendly GUI environment. A table with additional content information is generated from the analyses. CoreGenes has been validated with representative genomes from several families of viruses, as well as mitochondrion and chloroplast genomes. In these examples, it locates and identifies putatively related genes directly and gene clustering indirectly. In light of the similarities of certain genes generated by CoreGenes, one may ponder their relationships upon further and closer inspection, given that the high BLAST scores between two genes do not always imply an orthologous relationship [30]. In other words, the complexity of these BLAST scores suggests that the user should perform rigorous phylogenetic analysis of each set of homologous genes to determine true orthology. Though if the user uses a high threshold value while using GeneCore, s/he will increase the chances to retrieve orthologous genes.

One obvious application is to use this tool as a step in the characterization of an "alphabet" of putatively essential

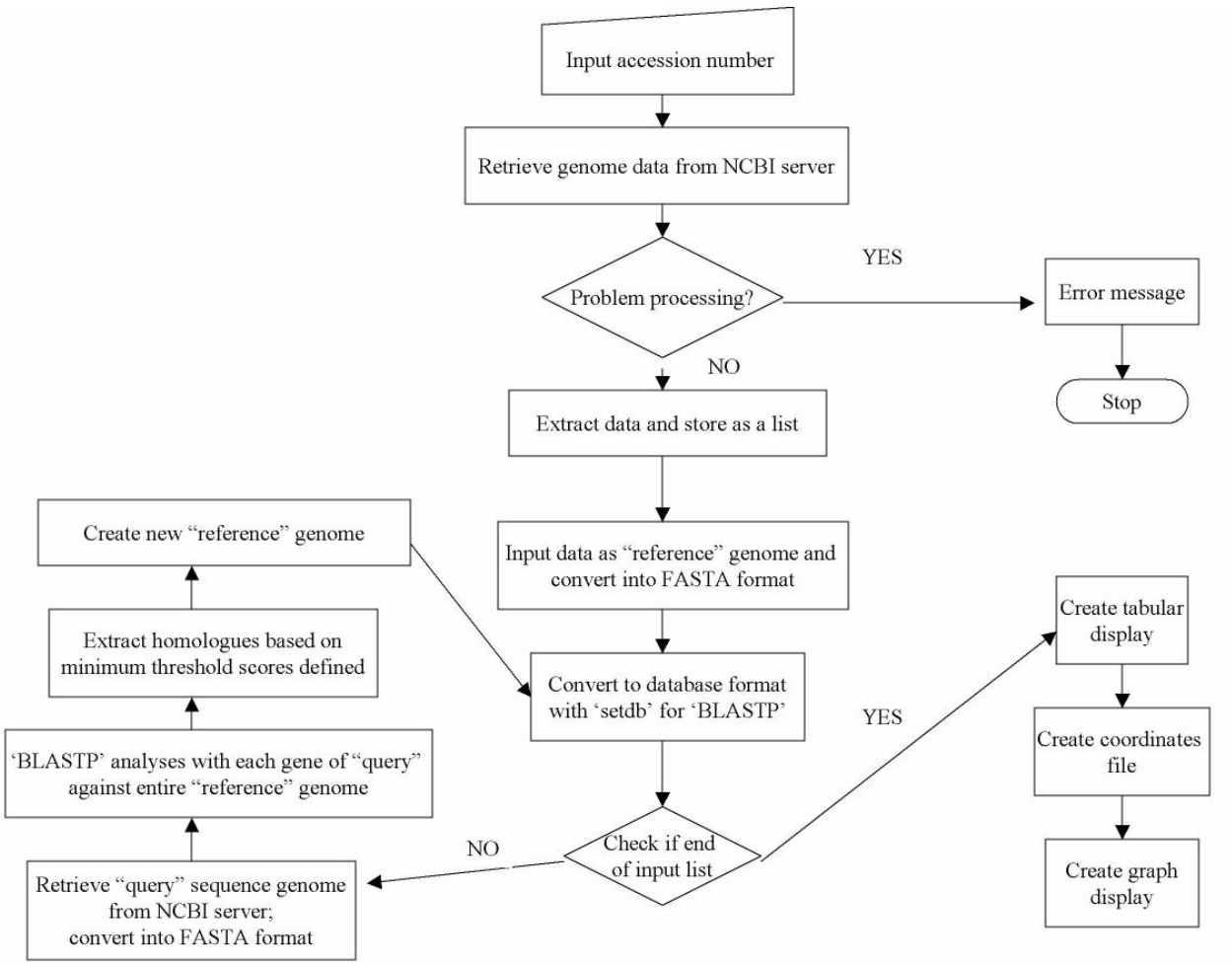


Figure 1
Flowchart of CoreGenes analysis Up to five genomes can be entered into the GUI and analyzed per session.

"core" genes in a set of closely related genomes such as from a collection of poxvirus genomes (*unpublished data*).

CoreGenes graphical user interface

The CoreGenes GUI contains three levels of data input/output, starting with an interface for the entry of two to five genomes via GenBank accession numbers (Figure 2) and ending with a display of the corresponding protein of interest as archived in the NCBI database. Contained within the top-level GUI (Figure 2) is an entry field for up to five genome sequences. *Nota Bene,* entering GenBank accession numbers with dyslexic renderings will result in "error messages." It is preferable to use the recent versions of GenBank accession numbers, i.e., prefixed with "NC_..."

Once the program is initiated, the respective genome data are downloaded from the GenBank database (Figure 1). These genome sequence data are subsequently parsed into protein-coding sequences [as annotated in the GenBank database] and are converted by CoreGenes into "GeneOrder2.0-FASTA" format [7,8,29,31]. Comparisons are performed and the results are presented in a tabular format in the subsequent GUI. Each gene has a hyperlink to its entry in the NCBI database.

Data mining algorithm

BLASTP protein similarity analyses [29] between the reference sequence and the first query sequence are performed sequentially, with each query protein compared individually to the entire protein database of the reference genome. This is similar to the algorithm for the GeneOrder analyses [7,8]. If the alignment score between the reference protein and a query protein meets or exceeds a defined similarity threshold number, then the proteins are paired and their accession numbers stored. A consensus

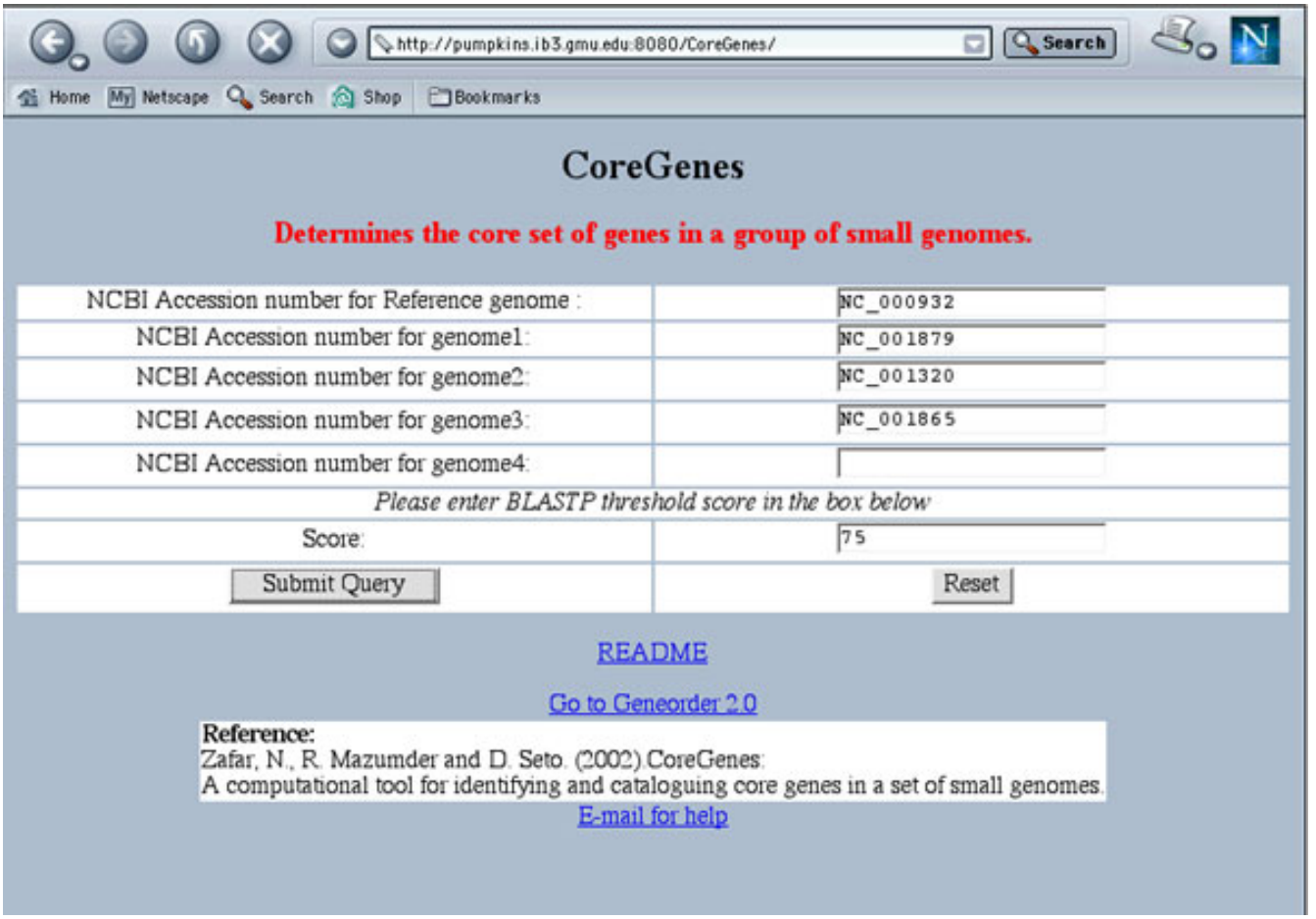
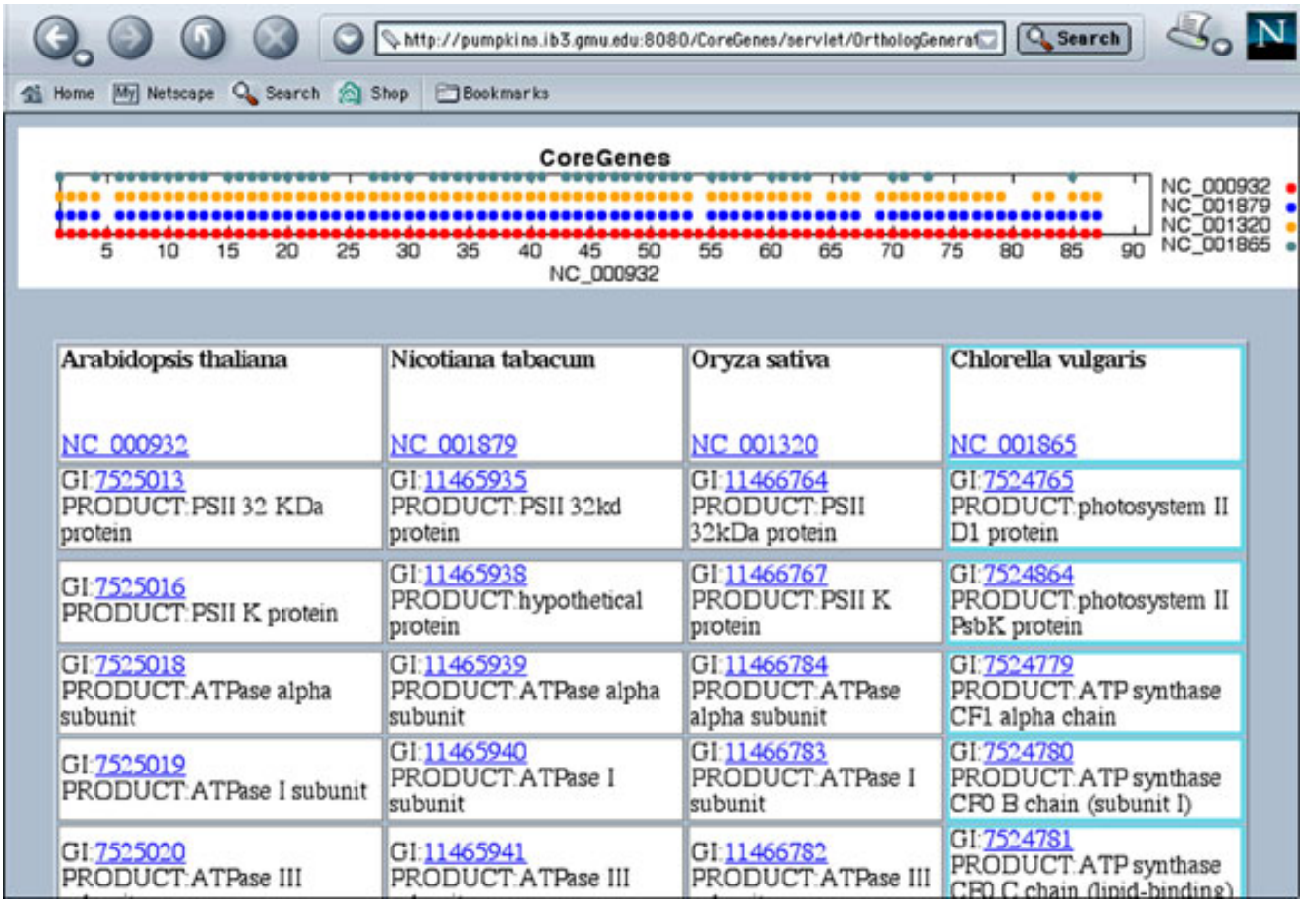


Figure 2
Screenshot of a CoreGenes session GenBank accession numbers are entered into each "sequence" field. Two to five genomes may be entered to extract the consensus set of "core" genes.

map of related genes is generated and stored. Hierarchical comparisons with additionally entered query genomes, up to four in total, are performed in each session.

In detail, the process continues as query genome number 2 data are retrieved from GenBank and treated as described above, *i.e.*, this set of proteins is compared against the first consensus set of paired genes formed between the reference genome and query genome number 1. A second consensus set of related genes is generated and stored. Query genome numbers 3 and 4 are iteratively and separately analyzed in an analogous manner. A caveat is that if query genome number 1 does not have a match to the reference genome, then a subsequent query genome number 2 match to the original reference genome (*i.e.*, possible true related gene) will be discarded. In other words, hierarchical matches must occur between the reference genome, query genome number 1 and query genome number 2 in order for CoreGenes to identify BLAST matches between the reference genome and the query genome number 2. A visual presentation of this is shown in Figure 3 (top panel), where the genomes are aligned with the reference genome serving as the "x-axis." Genes from query genomes that have the desired BLAST matches are arrayed vertically above the reference genome. This, despite its shortcoming of terminating a further analysis should there be no match between the two immediate genomes, is useful as a simple map of the order of genes contained in the reference genome. It also serves as a quick simple survey of the set of genomes in terms of BLAST matches.

However, permutations of the five genomes must be analyzed in order to collect the comprehensive set of putatively related core genes. Given the five genomes to be queried, this task is daunting manually. Of course it would be useful to generate a table of genes that bin across only 2, 3 or 4 genomes. This is being addressed actively. It is anticipated that this comprehensive table of genes including rows with matches across only two, three or four



Arabidopsis thaliana NC_000932	Nicotiana tabacum NC_001879	Oryza sativa NC_001320	Chlorella vulgaris NC_001865
GI:7525013 PRODUCT:PSII 32 KDa protein	GI:11465935 PRODUCT:PSII 32kd protein	GI:11466764 PRODUCT:PSII 32kDa protein	GI:7524765 PRODUCT:photosystem II D1 protein
GI:7525016 PRODUCT:PSII K protein	GI:11465938 PRODUCT:hypothetical protein	GI:11466767 PRODUCT:PSII K protein	GI:7524864 PRODUCT:photosystem II PsbK protein
GI:7525018 PRODUCT:ATPase alpha subunit	GI:11465939 PRODUCT:ATPase alpha subunit	GI:11466784 PRODUCT:ATPase alpha subunit	GI:7524779 PRODUCT:ATP synthase CF1 alpha chain
GI:7525019 PRODUCT:ATPase I subunit	GI:11465940 PRODUCT:ATPase I subunit	GI:11466783 PRODUCT:ATPase I subunit	GI:7524780 PRODUCT:ATP synthase CF0 B chain (subunit I)
GI:7525020 PRODUCT:ATPase III	GI:11465941 PRODUCT:ATPase III	GI:11466782 PRODUCT:ATPase III	GI:7524781 PRODUCT:ATP synthase CF0 C chain (lipid-binding)

Figure 3
Screenshot of a CoreGenes analysis The analysis generates a two-dimensional color-coded plot (top panel) displaying the core genes contained in a set of chloroplast genomes: *A. thaliana*, *N. tabacum*, *O. sativa* and *C. vulgaris*. The reference genome is the x-axis. Each genome is represented vertically above the reference by a different colored dot, indicated independently at the side of the graph. This data is also presented as a table (bottom panel) displaying the "core" genes contained in a set of chloroplast genomes: *A. thaliana*, *N. tabacum*, *O. sativa* and *C. vulgaris*. This data include hyperlinks to the NCBI database. A BLASTP threshold score is set at the default of "75" for this session.

genomes will be made available in the near future. Meanwhile, upon the completion of the above algorithm, a table containing the extracted GenBank data and summarizing the "core" genes within the queried genomes is generated (Figure 3 bottom panel). The columns of this table can also be exported via "cut and paste" into Microsoft Excel and Word programs to generate publication quality figures.

Accession numbers of each gene and very brief descriptions are presented in each individual block within this matrix, as extracted directly from the GenBank database. Each individual gene is hyperlinked from this table to the NCBI website to allow the investigator an opportunity to view the unique GenBank file for the gene of interest.

Using CoreGenes

CoreGenes generates a matrix of "core" genes for two to five genomes analyzed simultaneously. Currently, it is limited to small whole genomes, *ca*. 330 kb or smaller, as long as the data have been annotated in GenBank. Genomes of 35 kb, 150 kb and 250 kb have also been analyzed successfully (data not shown). The upper limit has not been explored in great detail. One main drawback is the time it takes to do each analysis, as the GenBank server needs to be accessed for each paired genome analysis. An alternative being developed is an option to run CoreGenes in a "batch processing" mode where the analyzed data are e-mailed back to the user after a request submission.

Similarity ranges

Contained in the top-level window is a field to define the minimum protein similarity score (*i. e.*, "BLASTP" threshold score). These can be either the default ("75") or a user-defined value. Score ranges are related to the similarities of the proteins being queried [7,8]. For reference, the three similarity ranges that can be defined for running GeneOrder2.0 are highest ("A"), high ("B") and low ("C") [7,8]. The BLASTP threshold score ranges for each are as follows: "A" is defined from [200-∞), "B" is defined from [100–200) and "C" is defined from [75–100). Genes with matches in the "A" range are true homologs, while those in the "B" range are likely related and those in the "C" range require visual validation of the level of identity in order to ensure a true match. Related gene matching values for CoreGenes are also defined in this manner. Caveat: it is always recommended that the results between two BLAST matches be scrutinized as reports have suggested that the closest BLAST match is often not the nearest neighbor [30].

Examples of CoreGenes analyses

This tool has been validated with analyses of several diverse virus, chloroplast and mitochondrion genomes. For example, a set of four chloroplast genomes (Figures 2 and 3) and a set of five mitochondrion genomes (data not shown) from evolutionary divergent sets of organisms were run independently to demonstrate the power and capabilities of CoreGenes. Shown in Figure 3 is an output from one of these analyses. With the BLASTP threshold score set at "75," the "core" genes are cataloged and displayed with brief identifying information from the GenBank database. Sixty-one "core" genes were cataloged from the set of chloroplast genomes (data not shown). The genomes are as follows: *Arabidopsis thaliana*, NC_000932; *Nicotiana tabacum*, NC_001879; *Oryza sativa*, NC_001320; and *Chlorella vulgaris*, NC_001865. Mitochondrion genomes are as follows: *Homo sapiens* (NC_001807), *Gallus gallus* (NC_001323), *Caenorhabditis elegans* (NC_001328), *Drosophila melanogaster* (NC_001709) and *Schizosaccharomyces pombe* (NC_001326). An analysis was also performed with a mixture of mitochondrion and chloroplast genomes. Interestingly, several putatively related genes were detected in this particular analysis (data not shown).

Additional validations

In addition to the aforementioned chloroplast and mitochondrion genomes, and of more interest to our research group, CoreGenes has been validated with virus genomes ranging in size from 35 kb to 330 kb (data not shown). Specifically, it has been run with combinations and permutations of adenovirus genomes, *ca.* 35 kb (NC_001405, NC_001406, NC_002067, NC_001454, NC_001460, NC_000942, NC_001813 and NC_002501, poxvirus genomes, ca. 250 kb NC_001559, NC_001266, NC_001266, NC_003027, NC_001132, NC_001731 and NC_002642), and other viruses of varying sizes: *ca.* 150 kb (*e.g.*, baculoviruses: *Heliocoverpa armigera nucleopolyhedrovirus G4*, NC_002654 and *Lymantria dispar nucleopolyhedrovirus* NC_001973) and *ca.* 330 kb (*Paramecium bursaria Chlorella virus 1*, NC_000852).

A group of three chordopox viruses (vaccinia NC_001559, *Molluscum contagiosum* virus NC_001731, and fowlpox virus NC_001266) and two entomopox viruses (*Melanoplus sanguinipes entomopoxvirus* NC_001993 and *Amsacta moorei entomopoxvirus* NC_002520) was analyzed with CoreGenes. With related genomes such as these, the data can also be used as a predictive tool for the elucidation of an "alphabet" of essential genes especially in collaboration with "wet bench" analyses such as the characterization of temperature sensitive mutants, for example, poxviruses (data not shown).

Limitations

Server Connectivity

CoreGenes run time is a function of the network connections. If one party, such as the NCBI server, is experiencing heavy traffic or is down due to technical difficulties, then the application will stall and be unsuccessful. Sets of orthopoxviruses, *ca.* 250 kb, take approximately 25 minutes to run on a PowerMac G3 running Mac OS 9.0 and Netscape Communicator 6.1. Larger genomes are currently problematic due to the computational speed, the NCBI server and/or the user's connection timing out. This issue is being addressed.

Some network "firewalls" may be incompatible with this software, causing the connections to terminate prematurely. An error message "An internal error has occurred. Please try again later java lang.NullPointer Exception." will be displayed. Also, entering incorrect accession numbers may give this same message. Alternatively, CoreGenes has been run successfully on university and public library terminals with internet access. These organizations do not seem to have the "firewall" needs/concerns as other organizations.

Platform Limitations

CoreGenes has been validated with several different platforms and also with different web browsers: Macintosh (Explorer 4.5 and Netscape 6.1), PC (Explorer 5.0 and Netscape 4.08), SGI (Netscape) and SUN (Netscape) workstations. There are compatibility issues between CoreGenes and Macintosh (Netscape 4.7 and below). Using Netscape 6.1 surmounts these problems. This problem appears to lie in the JAVA applet included with the earlier version of Netscape for Macintosh. Moving an Apple-supplied "JAVA Accelerator for PowerPC" into the "ex-

tensions" folder may allow earlier versions of Netscape to run this program. Printing the CoreGenes applet-generated graph may be problematic due to an applet incompatibility; capturing the graph as a "screenshot" via the PC and the Mac platforms and printing independently circumvents this.

Run times vary from 1 minute and 21 seconds for a set of five adenovirus genomes (*ca.* 35 kb) to 40 minutes for a set of five poxvirus genomes (*ca.* 250 kb). Currently, if there are multiple requests, the computation may take much longer as the requests are queued. This inconvenience is being addressed and is due to the server hosting the software. Depending on the hardware, some local servers may time out during this period while waiting for this request to be processed, which will result in an error message stating that "The attempt to load 'servlet' failed." Adjusting "preference" settings on the local web browser may rectify this problem. Immediate goals of improvement include an option to have results e-mailed back to the user. We expect that there will be additional improvements in both speed and response issues when we upgrade our server hardware and rewrite some of the CoreGenes software to accommodate the larger megabase-sized genomes.

Software Limitations

Only the NCBI database can be searched at this time; in other words, only GenBank accession numbers can be used. If there is an operator error in entering the number correctly, then an error message will be displayed, *e.g.*, "The attempt to load 'servlet' failed." Improvements to this software will include providing an additional field to enter proprietary and non-GenBank genome data, similar to an option developed for GeneOrder2.0 [8].

Conclusions

CoreGenes fits into the niche for GUI-based interactive computational tools [1–10] that enhance the visualization of DNA sequence data, especially in the context of genome comparisons. It meets a critical need for tool sets containing global "whole genome" analyses tools. As noted earlier, small genomes are still of great interest to many researchers. This tool is a base to expand upon, for example, to build more robust, elegant and complementary "whole genome" computational tools. Although CoreGenes successfully expedites the determination of "core" genes during the comparisons of several small whole genomes simultaneously, it will likely be succeeded by improved software to compare and analyze even much larger genomes, especially in the megabase range. This feature is being pursued with urgency. One known current limitation in analyzing larger genomes is computational, *e. g.*, hardware; this will be addressed shortly. Increasingly powerful workstations to act as servers will allow the much more computationally intensive comparisons of megabase-sized genomes. However, this version of CoreGenes is very useful and fills a current unmet need in genome analyses, that of collecting related genes in a family of genomes. In addition to stimulating the development of similar tools, CoreGenes will allow continuing improvements to it. We plan to support aggressively this version of CoreGenes, updating with improvements and additional features, as well as to work on a more robust faster version.

Acknowledgements

The CoreGenes software is written in JAVA, JAVAScript and HTML. JAVA-based programming is coded using "Jakarta-Tomcat," which is public domain software distributed by the Apache Software Foundation [http://www.apache.org] , and the Ptolemy II system distributed by The Regents of the University of California. We thank Yusuf Azaz for technical suggestions and to Srikanth Celamkoti and Sashidhara Kundeti for rewriting codes. We are grateful to Dean Murray Black (SCS, GMU) for his support. This work was funded in part by a grant from the American Type Culture Collection and also by a Faculty Development Award from the Office of the Provost (GMU). This is dedicated to the memory of Ng Siu Hong Seto (1927–2002).

References

1. Helt GA, Lewis S, Loraine AE, Rubin GM: **JAVA-based tools for genomic data visualization** *Genome Res* 1998, **8:**291-305
2. Dicks J: **Graphical tools for comparative genome analysis** *Yeast* 2000, **17:**6-15
3. Upton C, Hogg D, Perrin D, Boone M, Harris NL: **Viral genome organizer: a system for analyzing complete viral genomes** *Virus Res* 2000, **70:**55-64
4. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back** *J Mol Biol* 1998, **283:**707-725
5. Jareborg N, Durbin R: **Alfresco – a workbench for comparative genomic sequence analysis** *Genome Res* 2000, **10:**1148-1157
6. Buerstedde J-M, Prill F: **FOUNTAIN: A JAVA open-source package to assist large sequencing projects** *BMC Bioinformatics* 2001, **2:**6
7. Mazumder R, Kolaskar A, Seto D: **GeneOrder: comparing the order of genes in small genomes** *Bioinformatics* 2001, **17:**162-166
8. Zafar N, Mazumder R, Seto D: **Comparisons of gene co-linearity in genomes using GeneOrder2.0** *Trends in Biochem Sci* 2001, **26:**514-516
9. Ray WC: **Software review: GCG Wisconsin package 10.1** *in HMS Beagle: The BioMedNet Magazine* 2001, **99:**2001 Mar 30 [http://news.bmn.com/hmsbeagle/99/reviews/sreview]
10. Staden R, Beal KF, Bonfield JK: **The Staden package** *Methods Mol Biol* 1998, **132:**115-130
11. Miller W: **Comparison of genomic DNA sequences: solved and unsolved problems** *Bioinformatics* 2001, **17:**391-397
12. Hannenhalli S, Chappey C, Koonin EV, Pevzner PA: **Genome sequence comparison and scenarios for gene rearrangements: a test case** *Genomics* 1995, **30:**299-311
13. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, Frazer KA: **Active conservation of noncoding sequences revealed by three-way species comparisons** *Genome Res* 2000, **10:**1304-1306
14. Wiehe T, Guigo R, Miller W: **Genome sequence comparisons: hurdles in the fast lane to functional genomics** *Brief Bioinform* 2000, **1:**381-388
15. Boore JL, Brown WM: **Big trees from little genomes: mitochondrial gene order as a phylogenetic tool** *Curr Opin Genet Dev* 1998, **8:**668-674
16. Lang BF, Seif E, Gray MW, O'Kelly CJ, Burger G: **A comparative genomics approach to the evolution of eukaryotes and their mitochondria** *J Eukaryot Microbiol* 1999, **46:**320-326
17. Turmel M, Otis C, Lemieux C: **The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes** *Proc Natl Acad Sci, USA* 1999, **96:**10248-10253

18. Bansal AK: **An automated comparative analysis of 17 complete microbial genomes** *Bioinformatics* 1999, **15:**900-908
19. Shira M, Hirakawa H, Kimoto M, Tabuchi M, Kishi F, Ouchi K, Shiba T, Ishii K, Hattori M, Kuhara S, Nakazawa T: **Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA** *Nucleic Acids Res* 2000, **28:**2311-2314
20. Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome** *Genome Res* 1997, **7:**959-966
21. Koop BF, Hood L: **Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA** *Nat Genet* 1994, **7:**48-53
22. Glusman G, Rowen L, Lee I, Boysen C, Roach JC, Smit AF, Wang K, Koop BF, Hood L: **Comparative genomics of the human and mouse T cell receptor loci** *Immunity* 2001, **15:**337-349
23. Footz TK, Brinkman-Mills P, Banting GS, Maier SA, Riazi MA, Bridgland L, Hu S, Birren B, Minoshima S, Shimizu N, Pan H, Nguyen T, Fang F, Fu Y, Ray L, Wu H, Shaull S, Phan S, Yao Z, Chen F, Huan A, Hu P, Wang Q, Loh P, Qi S, Roe BA, McDermid HE: **Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere** *Genome Res* 2001, **11:**1053-1070
24. Keogh RS, Seoighe C, Wolfe KH: **Evolution of gene order and chromosome number in *Saccharomyces, Kluyveromyces* and related fungi** *Yeast* 1998, **14:**443-457
25. Keller B, Feuillet C: **Colinearity and gene density in grass genomes** *Trends Plant Sci* 2000, **5:**246-251
26. Gilley J, Fried M: **Extensive gene order differences within regions of conserved synteny between the *Fugu* and human genomes: implications for chromosomal evolution and the cloning of disease genes** *Hum Mol Genet* 1999, **8:**1313-1320
27. Ku HM, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny** *Proc Natl Acad Sci, USA* 2000, **97:**9121-9126
28. McLysaght A, Enright AJ, Skrabanek L, Wolfe KH: **Estimation of synteny conservation and genome compaction between puffer fish [*Fugu*] and human** *Yeast* 2000, **17:**22-36
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool** *J Mol Biol* 1990, **215:**403-410
30. Koski LB, Golding BG: **The closest BLAST hit is often not the nearest neighbor** *J. Mol. Evol.* 2001, **52:**540-542
31. Pearson WR: **Rapid and sensitive sequence comparison with FASTP and FASTA** *Meth. Enzymol* 1990, **183:**63-98