Proceedings of the International Conference on Computer Science and its Applications (ICCSA),
p 315-319, ISBN 0-9742-4480-5, July 01-02, 2003, San Diego, CA

SeqComp – A SEQUENCE ALIGNMENT PROGRAM FOR COMPARATIVE GENOMICS

Arun K. Datta, John Fischer, Jose Paz, and Yumiko Iwai
Department of Computer Science and Telecommunication
National University, San Diego, CA 92130
and
The Scripps Research Institute, La Jolla, CA 92037
Tel: (619) 563-7122; e-mail: *adatta@nu.edu*

Abstract

Large scale sequencing effort has yielded the genomic sequence data of a variety of microbes including the pathogenic bacteria and viruses. Such information is valuable for various biological research programs including vaccine development program, which often needs the information on the unique gene and its protein. Software programs are now available to compare the sequences of two or more genomes. However, the output often needs further processing to be useful for such purpose. Here we report a sequence alignment program developed using visual basic that compares two genomic sequences and outputs the result in an excel sheet tabulating the genes with the similarity score. This program also produces an output in text showing the de/similarities between the two genes/proteins at the nucleotide and amino acid level. This program is primarily designed for comparative genomics of bacteria and viruses. However, this can be used for larger genomes.

Keywords: sequence, genes, proteins, vaccine, genome comparison, genome alignment, alignment program

Introduction

The effort of sequencing various organisms has already started producing valuable information that may be further processed to obtain information for academic and commercial purposes. As for example, more than hundred microbial genomes are now sequenced (http://www.ncbi.nlm.nih.gov/PMGifs/genomes/micr.html) including pathogenic virus and bacteria. Such information once annotated is very much useful for vaccine research. For example, the genome sequence information of *Variola virus* (Smallpox) is already known (NC_001611) and such sequence information is also available for similar organism, such as, *Vaccinia virus* (NC_001559). It is, therefore, possible to develop an effective vaccination program for this pathogenic virus (Smallpox) once the unique proteins representing this virus could be identified. Although the annotated sequence information yields information on genes, comparison between the genes of two similar genomes is still a problem.

A good software program for comparative genomics is needed for various purposes. Comparisons of two genomic sequences of sufficiently similar organisms are expected to produce data that are useful

for vaccine program. Comparison between two or more dissimilar organisms is expected to produce information valuable for the evolutionary biologists (for a review, Frazer et al., 2003). Now with the advent of sequence information on human, mouse and other vertebrates, such comparison is yielding useful information to predict previously unknown genes and even its function (Thomas and Touchman, 2002). Thousands of potential genes might be eventually identified by computational methods for comparative genomics once adequate sequence information is available. As for example, most of the findings reported by the mouse sequencing consortium (Asif et al., 2002)were the result of the computational analysis using comparative genomics approach. Genome analysts have applied this approach at many levels, from multi-megabase rearrangements reflected in chromosome structure down to single nucleotide changes between orthologous genes (Boguski, 2002). Sequence comparison of various similar and dissimilar species will undoubtedly provide valuable information that can lead to identify the coding regions, characterize the regulatory elements, compare metabolic pathways, and even the history of the evolution of those genomes (Hood et al, 1995).

Although a number of software tools are available for comparative genomics (For a review, Frazer et al., 2003; Hohl, 2002; Miller, 2001), none is known to automate the process of tabulating the genes with the similarity score. Moreover, availability of source code for further development of such tools is often restricted (Jamison, 2003).

We have created a software tool, ***SeqComp***, that can efficiently compare two genomes and produces an output that is easy to understand. With a view towards the end user, the interface provided is clear and easy to use (see Figure 1). The program has a database with hyperlinked URLs that can automatically download sequence data from the NCBI for the chosen organism. The interface is also provided for manual download with the input of known accession numbers. The output of the computer processing is in MS Excel sheet tabulating a list of all the genes with the % score of similarity between two genes. Such information is also produced for proteins. A text file is also automatically generated showing the difference between the nucleotide sequences as well as amino acid sequences. Our methods are fast and the resulting alignments exhibit a high degree of sensitivity. Information extracted by using this method can be used for uniquely identifying an organism and for vaccine research. The work is in progress to include a visualization tool for representing larger genomes.

Experimental Results

Performance evaluation

Several experiments were done to check the performance of this software program ***SeqComp***. When we compared the genomes of two viruses- the *Variola virus* (Small pox; NC_001611; 185578 bp; Shchelkunov et al, 1994; 1996) vs. *Vaccinia virus* (NC_001559; 191737 bp; Goebel, et al, 1990), the analysis took about 3 minutes using a Pentium III computer with a processor of 450 MHz. The output of this result generated a table with 558 rows with 9 columns in an Excel sheet to represent all the genes and proteins with similarity score (%). A partial list is shown in Table 1. The text generated for the output was 595 pages long in MS Word. We also compared the genomes of *M.tuberculosis* H37Rv (NC_000962; 4411529 bp for H37Rv; Cole et al 1998) vs. *M.leprae* (3268203 bp; Cole et al, 2001). Results of this and the results obtained with manual inputs will be demonstrated.

Figure 1. Graphical user interface for ***SeqComp***.

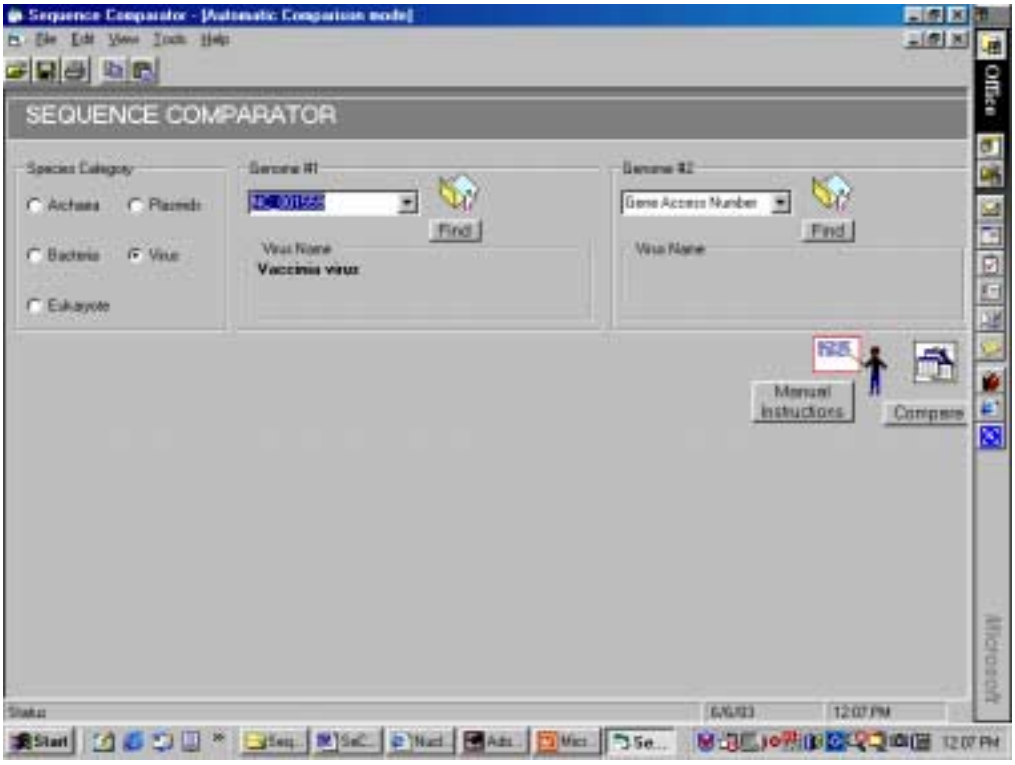
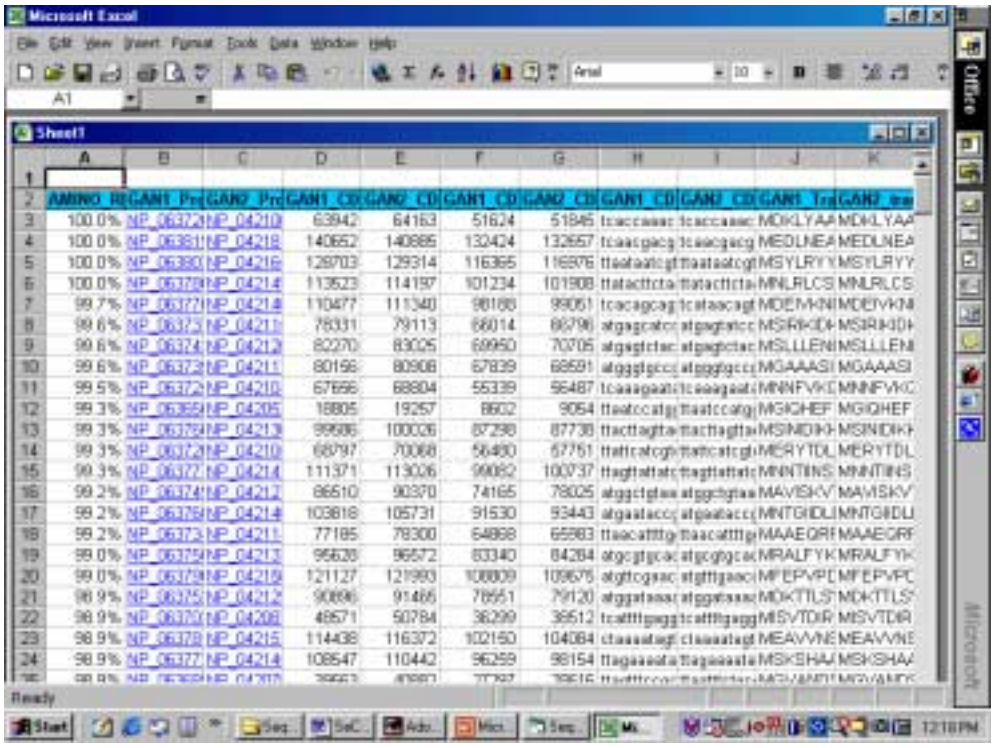


Table 1. A partial list of the output generated by using ***SeqComp*** after comparing the genomes of *Vaccinia virus* and the *Variola virus*.



Discussion

With the availability of genome sequence information of several bacteria and viruses, the sequence data are gaining attention for comparative sequencing that can effectively be utilized for vaccine development programs and diagnostics. Earlier, comparative sequence information for bacterial species was reported (Delcher et al, 1999; Florea et al., 2000). However, for our purpose, we have chosen the output that was not available using those programs. The ***SeqComp*** program utilizes the information of any genome sequence that is annotated with the protein_id number, gene's name and its start and stop site in the nucleotide sequence. The visual basic program has been written to capture these informations from the NCBI's web page. A database in MS Access is accordingly designed to include all the accession numbers presently available from NCBI. These are hyperlinked for automatically downloading the sequence information once the accession number is specified in the user interface of ***SeqComp***. The output obtained using this program can easily be used for various biological experiments, such as, designing PCR primers for amplification of unique genes for vaccine research or for diagnostic purpose.

For genomes with larger annotated genes and proteins, it will be necessary to include a visualization tool to represent the volume of comparative data. Efforts are underway to develop such tool and incorporate in ***SeqComp***.

References

- Asif T. Chinwalla, Lisa L. Cook, Kimberly D. Delehaunty, Ginger A. Fewell, Lucinda A. Fulton, Robert S. Fulton, Tina A. Graves, LaDeana W. Hillier, Elaine R. Mardis, John D. McPherson, et al., (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520 - 562.
- Boguski M.S. (2002). Comparative genomics: The mouse that roared. *Nature*, **420**: 515-516.
- Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **396**:190
- Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, et al. (2001). Massive gene decay in the leprosy bacillus. *Nature*, **409**:1007-11.
- Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999). Alignment of whole genomes. *Nucleic Acids Res.* **27:** 2369-2376.
- Florea, L., Riemer, C., Schwartz, S., Zhang, Z., Stajonovic, N., Miller, W., and McClelland, M. 2000. Web-based visualization tools for bacterial genome alignments. *Nucleic Acids Res.* **28:** 3486-3496.
- Frazer K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C. (2003) Cross-species sequence comparisons: A review of methods and available resources. *Genome Research*, **13**: 1-12.
- Goebel,S.J., Johnson,G.P., Perkus,M.E., Davis,S.W., Winslow,J.P. and Paoletti,E.(1990). The complete DNA sequence of vaccinia virus. *Virology* **179**: 247-266.
- Hohl, M., Kurtz, S., and Ohlebusch, E. (2002). Efficient multiple genome alignment. *Bioinformatics*, **18**: S312-S320.

- Hood, L., Rowen, L., and Koop, B.F. (1995). Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. *Ann. NY Acad. Sci*., **758**: 390-412.
- Jamison, D.C. (2003). Open bioinformatics. *Bioinformatics*, **19**: 679-680.
- Miller, W. (2001). Comparison of genomic DNA sequences: Solved and unsolved problems. *Bioinformatics* **17:** 391-397.
- Shchelkunov,S.N., Blinov,V.M., Resenchuk,S.M., Totmenin,A.V., Olenina,L.V., Chirikova,G.B. and Sandakhchiev,L.S.(1994) Analysis of the nucleotide sequence of 53 kbp from the right terminus of the genome of variola major virus strain India-1967. *Virus Res*. **34**: 207-236.
- Shchelkunov,S.N., Totmenin,A.V. and Sandakhchiev,L.S.(1996). Analysis of the nucleotide sequence of 23.8 kbp from the left terminus of the genome of variola major virus strain India-1967. *Virus Res*. **40**: 169-183.
- Thomas, JW and Touchman, JW (2003). Vertebrate genome sequencing: building a backbone for comparative genomics. *Trends in Genetics*, **18**: 104-108.