BMC Bioinformatics

Methodology article

MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences

James J Campanella*, Ledion Bitincka and John Smalley

Address: Montclair State University, Department of Biology and Molecular Biology, 1 Normal Avenue, Montclair, New Jersey 07043 USA

Email: James J Campanella* - james.campanella@montclair.edu; Ledion Bitincka - bitinckal1@mail.montclair.edu; John Smalley - smalley@mail.montclair.edu

* Corresponding author

Published: 10 July 2003

BMC Bioinformatics 2003, **4**:29

Received: 09 May 2003
Accepted: 10 July 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/29

Abstract

Background: The rapid increase in the amount of protein and DNA sequence information available has become almost overwhelming to researchers. So much information is now accessible that high-quality, functional gene analysis and categorization has become a major goal for many laboratories. To aid in this categorization, there is a need for non-commercial software that is able to both align sequences and also calculate pairwise levels of similarity/identity.

Results: We have developed MatGAT (Matrix Global Alignment Tool), a simple, easy to use computer application that generates similarity/identity matrices for DNA or protein sequences without needing pre-alignment of the data.

Conclusions: The advantages of this program over other software are that it is open-source freeware, can analyze a large number of sequences simultaneously, can visualize both sequence alignment and similarity/identity values concurrently, employs global alignment in calculations, and has been formatted to run under both the Unix and the Microsoft Windows Operating Systems. We are presently completing the Macintosh-based version of the program.

Introduction

The application of phylogenetics in the examination of a genome has been dubbed "phylogenomics" [1–3]. The analytic process of phylogenomics is taking on more importance as additional DNA and protein sequences from a multitude of species become available.

GenBank has approximately 28 million DNA sequences in its database http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html. The number of sequences in GenBank has increased by five orders of magnitude since its founding in 1982. The Institute for Genomic Research (TIGR), by the end of 1998, had completed sequencing seven microbial genomes, half of the world total at the time. Today, TIGR is in the process of sequencing and characterizing the genomes of many major organisms of the world, including 20 animal, 19 plant, 14 protist, 8 fungal and over 100 bacterial species http://www.tigr.org.

All this new information is obviously a great asset to scientists, since there is constantly new supplementary data to be employed in genomic, physiologic and genetic research. The drawback with all of this new information is that the sheer amount of it has become overwhelming. So

much information is now becoming available that high-quality, functional gene analysis and categorization is becoming a paramount goal.

One of the most important analyses that can be employed in phylogenomics or phylogenetics is the pairwise determination of similarity or identity between DNA or protein sequences. The percent identity is the calculated percentage of how two sequences compare at a base-to-base or residue-to-residue level. The percent similarity is a more strict calculation where sequence gaps and mismatches are included in the evaluation and scored using a more complex formula and a comparison look-up table [4–6].

We have noted that there is a lack of non-commercial software available that is able to both align a series of DNA or protein sequences and also calculate pairwise levels of similarity/identity. Timothy Carver's DISTMAT program http://bioinfo.pbi.nrc.ca:8090/cgi-bin/emboss.pl?_action=input&_app=distmat calculates pairwise divergence, but not similarity, and it only functions if the sequences have already been aligned by some other computer program. Pairwise BLAST http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html may also be used to calculate similarity, but its limitations are that only two sequences may be analyzed at one time and percent similarity/identity are based on local alignment – not global alignment [7]. MegAlign, which comes with the DNASTAR package (DNASTAR, Inc.), also generates similarity matrices, but it is quite expensive and not available as a stand-alone product.

MatGAT (Matrix Global Alignment Tool) is a simple, easy to use similarity/identity matrix generator that calculates the similarity and identity between every pair of sequences in a given data set without requiring pre-alignment of the data. The program performs a series of pairwise alignments using the Myers and Miller global alignment algorithm [8], calculates similarity and identity, and then places the results in a distance matrix. In order to increase alignment speed, they are computed in the C++ language while the "front-end" of the MatGAT program is encoded in Java.

We developed MatGAT because of a perceived need. MatGAT runs under both the Unix and Microsoft Windows Operating Systems. We are presently completing the Macintosh OS X-based version of the program. The program operates as a native application and makes use of graphical interfaces, allowing the user to employ standard fonts installed on their machine and printer. Data may be input into MatGAT by cutting and pasting or using a browse function for larger files. Files must be plain .txt in the standard FASTA format. In multiple sequence analysis, each field must have a FASTA title line starting with a ">"

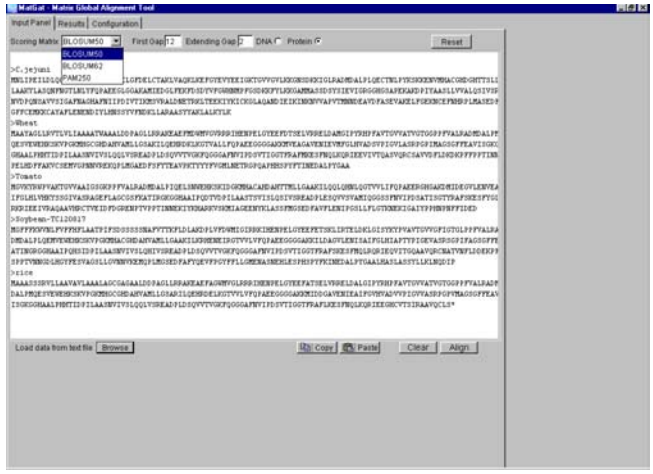

Figure 1
Screen shot of the data input screen of MatGAT v2.0. Protein or DNA sequences in the FASTA format can be entered by hand into the data box, uploaded from a text file, or pasted into place. Several scoring matrices are available for analyses: BLOSUM50, BLOSUM62, and PAM250. Additionally, "First Gap" and "Gap Extension" conditions may be altered for optimal alignment. The "Clear" button will delete the input sequence data and alignments, while leaving the matrix output unaffected until new data are analyzed.

(Fig. 1). Numerals and spaces are allowed during data entry of the comment line after the ">", while numbers and spaces are automatically removed by MatGAT in the sequence data. Test data files in the FASTA format are included with the MatGAT archive. The user may specify which type of alignment matrix (BLOSUM50, BLOSUM62, and PAM250) to employ with their protein sequence examination.

Data files of up to 200 DNA or protein sequences have been analyzed successfully using MatGAT. The DNA sequences analyzed were 1000–2000 basepairs in size and took ~90 min to finish a run using a Pentium 3 Processor on a standard PC. The protein sequences ranged from 300 to 600 amino acids in length and took 12 min to complete an analysis using the same machine.

The output for MatGAT may be viewed on the computer screen or printed directly. The results may also be saved as a text file, or Microsoft Excel delimited file, to be used for further statistical and phylogenetic analyses. Moreover, when first booted up MatGAT searches for the presence of Excel on the user's hard drive. If detected, this information is saved and output matrices may then be directly transferred to Excel by the click of a single screen button.

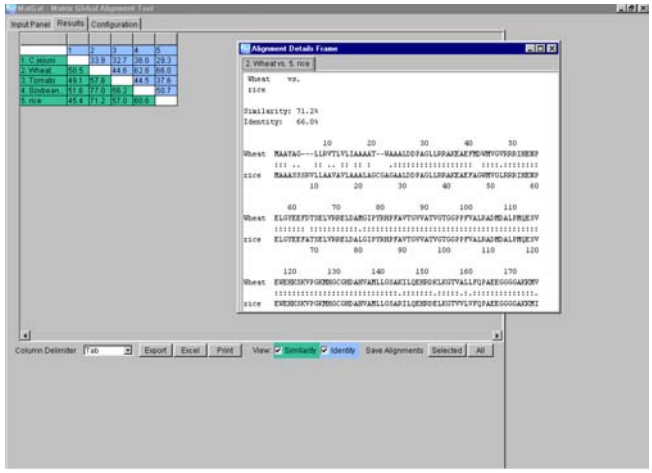

Figure 2
Screen shot of MatGAT v2.0 output running under Windows XP. Aprotein data set is analyzed by MatGAT. The upper matrix contains the identity of the data set and the lower is the similarity. The inset screen contains a pop-up window generated by clicking on the sequence pair of interest; this window displays the pairwise alignment of the tomato and soybean protein sequences. The "Save All" button saves all the alignments into a text file, while the "Save Selected" button is enabled once you select one of the alignments for display. This button will selectively save all the alignments that you have chosen to view.

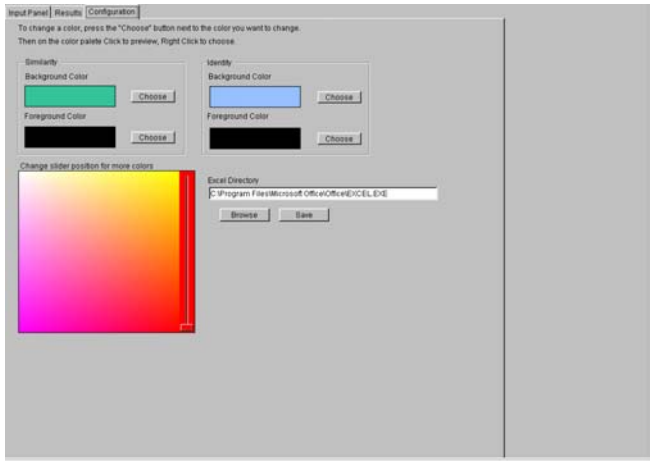

Figure 3
Screen shot of the Configuration Window of MatGAT v2.0. The colors of data output may be controlled from this screen, as well as configuration of Excel recognition by MatGAT.

Screen output includes clear labels for the names of analyzed sequences (Fig. 2) with data columns of adjustable width. The identity is encoded in the upper matrix and similarity in the lower matrix. Clicking the cursor over a similarity or identity value on the on-screen matrix will create a pop-up window containing the global alignment actually employed to obtain that score (Fig. 2). MatGAT has the ability to output either selected pairwise alignments or all alignments generated into a text file allowing the researcher to see the basis of the similarity/identity matrix. MatGAT's Configuration Screen (Fig. 3) allows alteration of the matrix background and text colors, permitting clear differentiation between values.

Error detection routines include detection and stripping of numbers in DNA and protein data; detection of inappropriate DNA bases other than G, T, C, A, N, and * for wildcards; discrimination between DNA and amino acid sequences and indication of the appropriate type of analysis; automatic stripping of spaces from pasted datasets and prohibition of spaces in data during manual entry; and, finally evaluation of the number of sequence entries and error flagging if this value is not greater than one.

Availability

MatGAT v2.0 can be obtained as a compacted Zip-file from the following World Wide Web sites: http://www.angelfire.com/nj2/arabidopsis/MatGAT.html or http://www.bitincka.com/ledion/matgat. Additionally, the software has been submitted for public distribution to the Indiana University Biology Archive (IUBIO Archive) http://iubio.bio.indiana.edu/soft/molbio/evolve/. The PC version of the program requires the presence of a JAVA run-time environment under the following MS Windows interfaces: Windows 98, 2000, NT, or XP. The Unix version of the program must also run on a JAVA-enabled machine. Additionally, the PC version of the program will run effectively under Windows emulation on Macintosh Computers running under OS X.

The Java runtime environment is available on all PC computers installed with Windows 98, or later, and Netscape. If the user does not have Java installed on their PC, then they may obtain it from http://java.sun.com/j2se/1.3/download.html. Macintosh users may download Java from http://devworld.apple.com/java/download.html. Users of the Unix Operating system may download Java from http://www.sco.com/developers/java/download/index.html.

Additional Files

A link for downloading the MatGAT v2.0 program for Windows is included with this article [see Additional file: 1 1]. The archive is formatted as a Microsoft Zip file, entitled "MatGAT 2.0.zip", and can be opened by any

Windows unpacking program such as WinZip. Included in the archive are all files that are needed to run MatGAT 2.0, including two test data files entitled "Test Data DNA.txt" and "Test Data Protein.txt". Additionally in the archive, there is a README.txt file that acts as a help and bug repair update file. Once the program files are unzipped from their archive and into their own folder, the user starts the program by double-clicking on the MS-DOS batch "Run" file in the directory.

Authors' Contributions

LB coded the Windows and Unix MatGAT programs. JC conceived of the program, guided the overall design/debugging process, and drafted the manuscript. JS participated in design/debugging of MatGAT and is programming the Macintosh version of the package.

Additional material

Additional file 1

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-29-S1.zip]

Acknowledgements

We thank Vinela Bakllamaja for her interface suggestions, and Lisa Campanella for her help in editing this article.

References

1. Eisen JA: **Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis** *Genome Research* 1998, **8:**163-167.
2. Eisen JA and Hanawalt PC: **A phylogenomic study of DNA repair genes, proteins, and processes** *Mutat Res* 1999, **435:**171-213.
3. Eisen JA and Wu M: **Phylogenetic analysis and gene functional predictions: phylogenomics in action** *Theoretical Population Biology* 2002, **61:**481-487.
4. Needleman SB and Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins** *J Mol Biol* 1970, **48:**443-453.
5. Pearson WR and Lipman DJ: **Improved Tools for Biological Sequence Comparison** *Proc Natl Acad Sci* 1988, **85:**2444-2448.
6. Shpaer EG, Robinson M, Yee D, Candlin JD, Mines R and Hunkapiller T: **Sensitivity and Selectivity in Protein Similarity Searches** *Genomics* 1996, **38:**179-191.
7. Tatusova TA and Madden TL: **Blast 2 sequences – a new tool for comparing protein and nucleotide sequences** *FEMS Microbiol Lett* 1999, **174:**247-250.
8. Myers EW and Miller W: **Optimal alignments in linear space** *Comp Applic Biosci* 1988, **4:**11-17.