J. Appl. Environ. Biol. Sci., 4(9S)458-462, 2014

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

E-SICT: An Efficient Similarity and Identity Matrix Calculating Tool

Muhammad Tariq Pervez[1], Masroor Ellahi Babar[2], Asif Nadeem3, Naeem Aslam[4] and Tanveer Hussain[5]

[1]Department of Computer Science, Virtual University of Pakistan
[2]Department of Bioinformatics, Virtual University of Pakistan
[3]Institute of Biochemistry and Biotechnology, University of Veterinary and Animal Sciences, Lahore, Pakistan
[4]Department of Computer Science, NFC Institute of Engineering & Technological Training, Multan, Pakistan
[5]Department of Live Stock Production, University of Veterinary and Animal Sciences, Lahore, Pakistan

Received: September 12, 2014
Accepted: November 23, 2014

ABSTRACT

The latest generation sequencing, microarrays and protein shotgun experiments are generating terabytes of data on daily basis. To process large data efficiently is one of the highly scrutinized areas in bioinformatics. Secondly, the use of phylogenetic in the analysis of genome has increased and pairwise determination of similarity and identity between protein or DNA sequences is one of the important analyses in the domain of phylogenetics. To construct similarity/identity matrix of very large alignments efficiently, E-SICT, which is a graphical user interface based tool, was developed using java programming language. E-SICT has implemented divide and conquer approach using Java threads to make this tool very efficient. Results showed that E-SICT can calculate identity matrix of an alignment comprising 11298 sequences (each sequence was 2696 base pairs long) in less than 100 seconds. It calculated similarity matrix of an alignment having 1614 sequences in less than 90 seconds. It also allows the user to save and print the identity or similarity matrix.

KEYWORDS*:* Similarity; Identity; MSA

1 INTRODUCTION

The technology of next-generation sequencing (NGS) is generating huge amount of data on daily basis. Previous experiments which were conducted using microarrays generated data in megabytes whereas experiments being conducted in the present era are producing data in gigabytes. Managing and analysis of so much big data is demanding momentous investment in computational infrastructure. The scientists are devising new tools and technologies to cope with the immense increase in raw data {Jones et al. 2012; Sankar et al. 2002; Campanella et al. 2003}. Finding similarity between two or more sequences is one of the important parts of sequence analysis. Searching nucleotide and protein databases is one of the highly securitized activity in the domain of bioinformatics {Sabo et al. 2005}. Searching similarity and identity help in identifying structural and functional information of the sequences {Arnold et al. 2005}. Determining pairwise identity or similarity between protein or DNA sequences plays a primary role in the field of phylogenetics. Comparison of the sequences at residue-to-residue or base-to-base level results in computation of the percent identity (PID). Computation of percent similarity is more complex. Sequence gaps and mismatches are considered in the evaluation and PID is computed using various substitution matrices {Arnold et al. 2005; Needleman and Wunch, 1970; Pearson and Lipman, 1988; Shpaer et al. 1996}.

Several tools are available for similarity search. Most of them search similarity of a given sequence with the sequences in the selected database. Secondly they find similarity of two sequences at one time. Many tools search similarity across the local alignments. Pairwise BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) searches similarity of two sequences at one time based on the local alignment {Tatusova and Madden, 1999}. MegAlig, which is that part of DNASTAR package (DNASTAR, Inc.) may also be used to compute similarity but it is not available free of cost and as an independent product. MatGAT (Matrix Global Alignment Tool) {Campanella et al. 2003} is tool to generate similarity/identity but it reconstructs an alignment first which is itself a time consuming task.

*** Corresponding Author:** Muhammad Tariq Pervez, Department of Computer Science, Virtual University of Pakistan. tariq_cp@hotmail.com

Secondly it does not support more than 500 sequences. SIAS (Sequence Identity and Similarity) {Reche et al. 2008} can also compute similarity and identity but being a web based application it does not support big alignments.

E-SICT (Efficient Similarity and Identity Matrix Calculating Tool) is tool which is written in Java programming language. It calculates similarity/identity matrix of large alignment efficiently. Results showed that E-SICT calculated identity matrix of 11298 sequences in less than one hundred seconds. Length of this alignment was 2696 base pairs. Since similarity calculation is more complex job and involves several types of formulae and protein substitution matrices. Results that E-SICT calculated similarity matrix of an alignment having 1696 sequences and 2696 base pairs length only in 90 seconds. E-SICT also provides the features to save or print similarity/identity matrix

2. RESULTS AND DISCUSSION

E-SICT is an efficient tool to calculate similarity and identity matrix of protein and DNA sequences. Its unique feature is that it is very efficient as compared to the similar tools.

2.1 Interface to Select MSA and Set Various Parameters

It provides a user friendly interface to select an alignment and set various parameters (Figure 1). It allows the select various protein substitution matrices, data type (DNA or protein), the action to be performed (Identity or similarity).

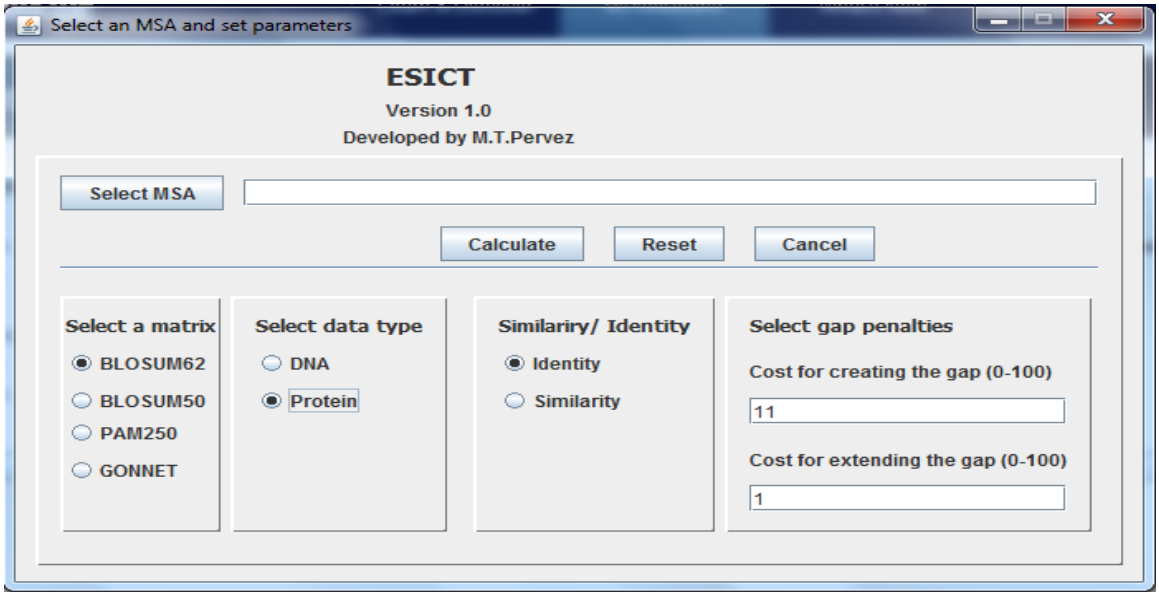

Figure 1. An interface to select an alignment and set various parameters

2.2. Calculation of Identity/Similarity Matrix

E-SICT allows the user to calculate many times big alignments as compared to the similar tools. MatGAT, which is also written in java programming, does not support more than 1000 sequences. It generates alignment prior to compute similarity/identity matrix, therefore, it consumes huge amount of time while processing small alignments (having 200 to 300 sequences). SIAS is a web based application for calculating similarity/identity. It, being an online tool, can not process big alignments. Results showed that, as compared to MatGAT and SIAS, E-SICT processed 2159% big alignment and calculated identity matrix (Figure 2).

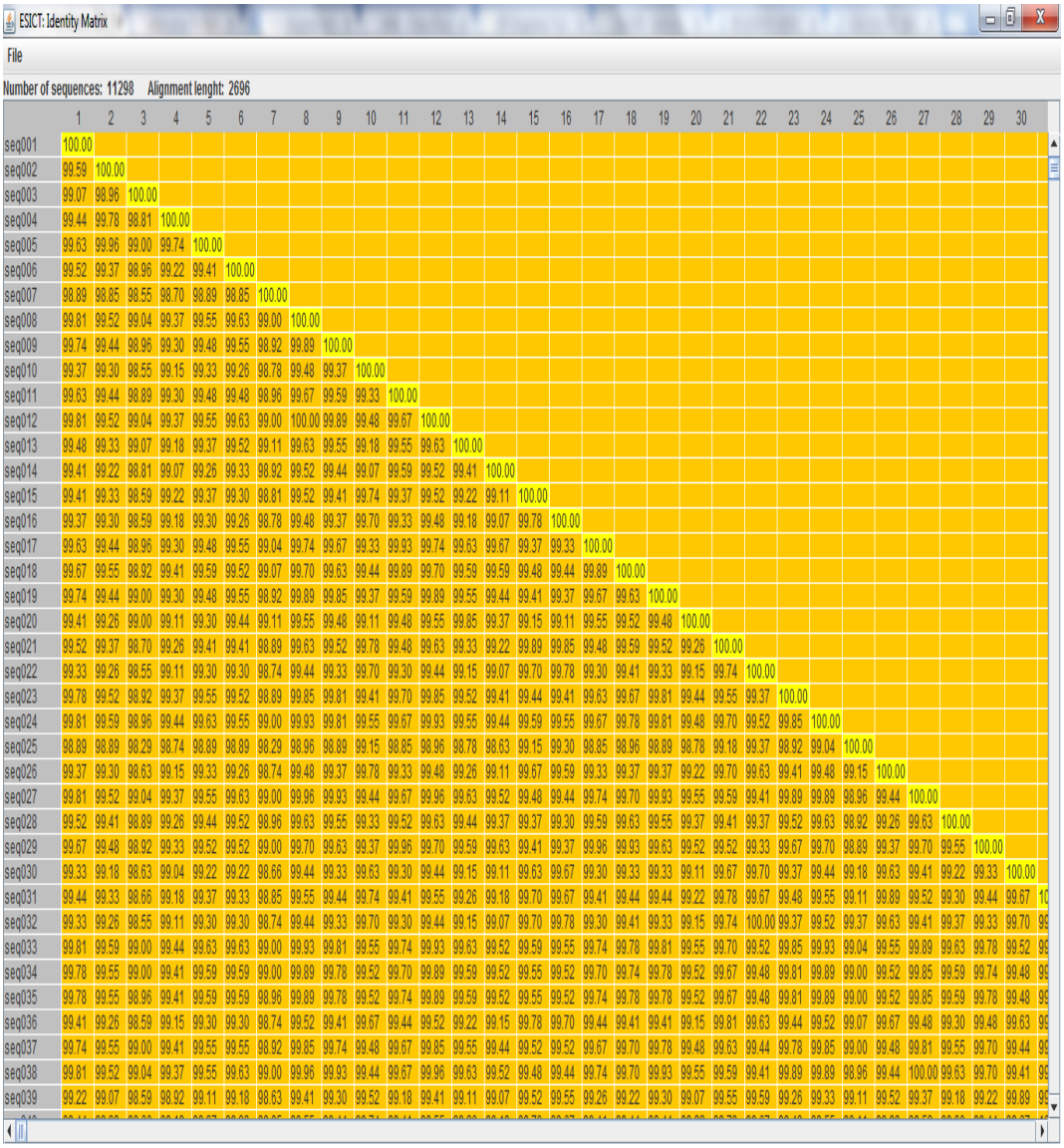

Figure 2. Identity matrix calculated by E-SICT. It calculated identity matrix of 11298 sequences in less than 100 seconds.

Calculation of similarity matrix is complex task. It involves several formulae and a protein substitution matrix. The similar tool consumes huge amount of time while E-SICT calculated similarity matrix of 1614 sequences (sequence length was 2696 base pairs) only in 90 seconds (Figure 3).

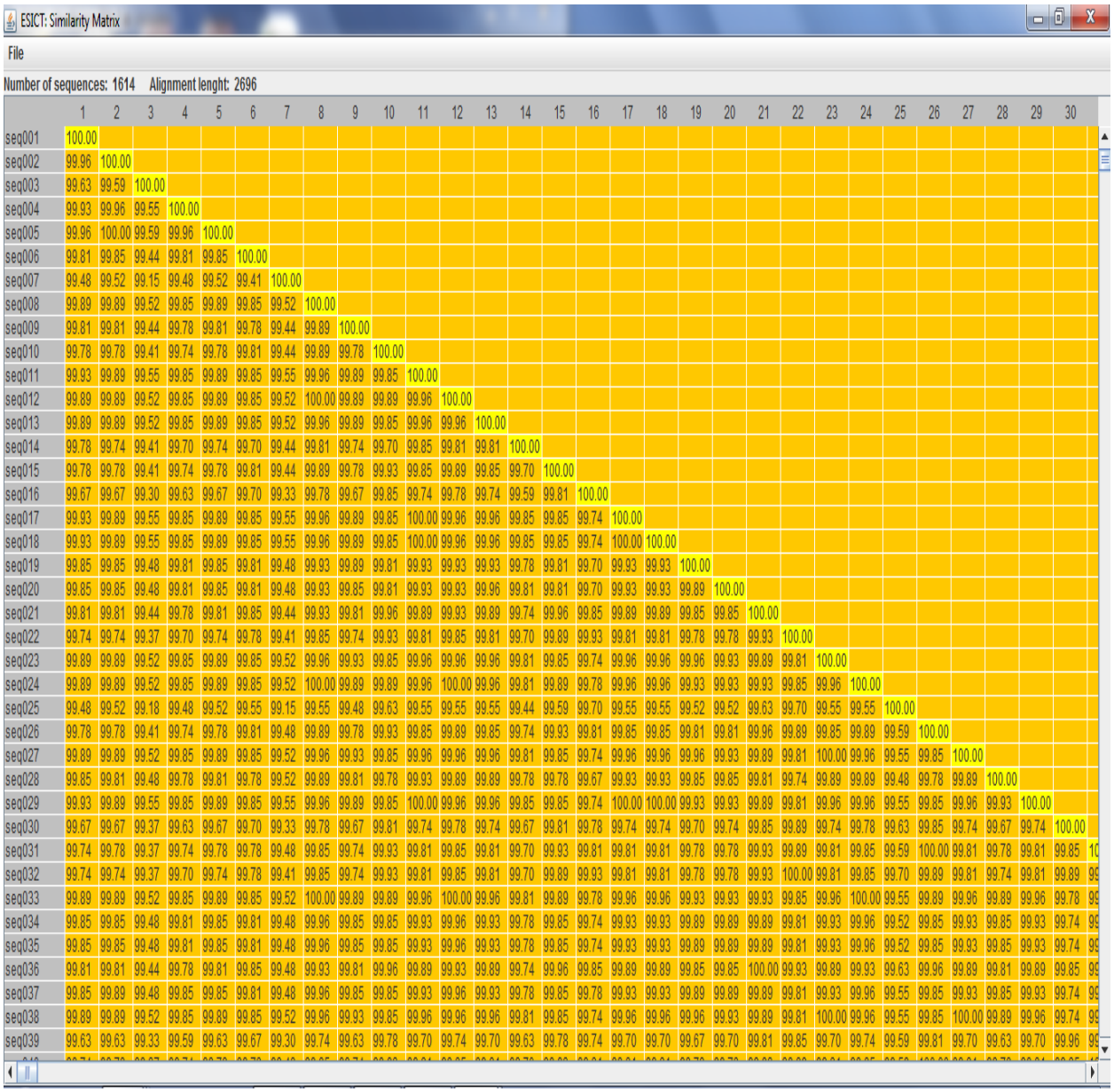

Figure 3. Similarity calculation by E-SICT. It consumes less than ninety seconds to process an alignment having 1614 sequences.

2.3. Other Features

E-SICT provides feature to print directly the results. It also allows the user to save the results in a text file to be used for further analysis and phylogenetic analysis.

3. MATRIAL AND METHODS

3.1. Computing Machine

E-SICT was developed and analyzed on a computing machine having Core i7 3.34 GHz processor and 8 GB RAM.

4. CONCLUSION

Several tools exist to calculate similarity/identity matrices. All of them either support small alignments or consume huge amount of time. E-SICT, which is written in java programming language, is very easy to use

and an efficient tool. It can calcualate similairyt/identity matrices in few seconds. It also provides features to save or print the results.

Availability

www.ivistmsa.com

REFERENCES

1. Jones, D. C., Ruzzo, W. L., Peng, X., & Katze, M. G. (2012). Compression of next-generation sequencing reads aided by highly efficient de novo assembly. Nucleic acids research, 40(22), e171-e171..
2. Sankar N, Machado J, Abdulla P, Hilliker AJ and Coe IR. (2002). Comparative genomic analysis of equilibrative nucleoside transporters suggests conserved protein structure despite limited sequence identity Nucleic Acids Res., 30, 4339–4350
3. Sæbø, P. E., Andersen, S. M., Myrseth, J., Laerdahl, J. K., & Rognes, T. (2005). PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. Nucleic acids research, 33(suppl 2), W535-W539.
4. Arnold, R., Rattei, T., Tischler, P., Truong, M. D., Stümpflen, V., & Mewes, W. (2005). SIMAP—the similarity matrix of proteins. Bioinformatics, 21(suppl 2), ii42-ii46.
5. Campanella, J. J., Bitincka, L., & Smalley, J. (2003). MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences.BMC bioinformatics, 4(1), 29.
6. 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.
7. Pearson WR and Lipman DJ: Improved Tools for Biological Sequence Comparison Proc Natl Acad Sci 1988, 85:2444-2448.
8. 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.
9. Tatusova TA and Madden TL: Blast 2 sequences – a new tool for comparing protein and nucleotide sequences FEMS Microbiol Lett 1999, 174:247-250
10. Reche et al. 2008. SIAS: Sequence identities and similarities http://imed.med.ucm.es/Tools/sias.html