Bioinformation Volume 5 open access

www.bioinformation.net Issue 9 **Software**

SeqMaT: A sequence manipulation tool for phylogenetic analysis

Pavan Kumar Attaluri[1], Mary C Christman[2], Zhengxin Chen[1], Guoqing Lu[2]*

[1]Department of Computer Science; [2]Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA; Guoqing Lu - Email: glu3@unomaha.edu; Phone: 1-402-5543195; Fax: 1-402-5543532; *Corresponding author.

Received December 23, 2010; Accepted January 01, 2011; Published February 07, 2011

Abstract:
Most bioinformatics tools require specialized input formats for sequence comparison and analysis. This is particularly true for molecular phylogeny programs, which accept only certain formats. In addition, it is often necessary to eliminate highly similar sequences among the input, especially when the dataset is large. Moreover, most programs have restrictions upon the sequence name. Here we introduce SeqMaT, a Sequence Manipulation Tool. It has the following functions: 1) data format conversion, 2) sequence name coding and decoding, 3) redundant and highly similar sequence removal, and 4) data mining utilities. SeqMaT was developed using Java with two versions, web-based and standalone. A standalone program is convenient to manipulate a large number of sequences, while the web version will guarantee wide availability of the tool for researchers and practitioners throughout the Internet.

Availability: both web-based and standalone versions are available at http://glee.ist.unomaha.edu/seqmat. The standalone version can be downloaded from the Downloads section. In addition, the website contains a tutorial and sample data.

Keywords: SeqMaT, format conversion, phylogeny, data mining

Background:
Molecular phylogeny is a fundamental approach studying species evolution and gene function. Many phylogenetic analysis programs are available, but each program often requires a particular type of input sequence format. Most tools have restrictions regarding the allowable length of sequence name and characters used. Automatically trimming the sequence names may lose important information, such as species names. Most often the trimmed names are redundant, which cannot be accepted by any analysis program. In addition, some programs do not accept sequence names with special characters. Furthermore, the input may contain identical or highly similar sequences which need to be removed to save computational resources and improve resolution of the resulting trees.

A variety of sequence formats are available for phylogenetic analysis, including FASTA **[1]** and Phylip **[2]**. The formats can be converted using tools such as ReadSeq **[3]**, Bugaco **[4]**, Format Converter **[5]** and EMBOSS **[6]**. However, ReadSeq and Bugaco cannot convert formats from or to Mega (Interleaved and Sequential). These tools do not provide tree format conversions from Newick to Nexus. Phylip sequence format allows only 10 characters in the sequence name; the output file from most format conversion tools cannot be accepted by Phylip because of this constraint. Other alternatives such as, REFGEN and TREENAMER **[7]**, are available for handling the sequence name problem. However, those tools do not accept FASTA files other than GenBank and DOE JGI formats; they extract the total or a part of the accession number to rename the sequence. Additionally, they don't support conversion back to the original names once the tree is created. It is thus extremely inconvenient for the user to analyze the phylogenetic result.

The computer program libraries such as BioJava **[8]** and BioPerl **[9]** provide functions to convert between different formats but require computer skills to use them. On the other hand, sequence manipulation tools focus mainly on utility functions such as removing and sorting sequences. SeqMaT combines format conversion and essential functions of sequence manipulation under a single platform, with the expectation of addressing the issues with currently available tools.

Methodology:
Based upon our practical experience, we designed a pipeline to deal with the various data formatting and conversion issues described above. Figure 1 shows data flow and major functional modules of SeqMaT. SeqMaT accepts and converts a number of sequence formats, including FASTA, Clustal, Mega, Phylip, Nexus, EMBL, GenBank, PIR, Table, Hennig86, and plain text. The functional modules include 1) finding and removing redundant sequences, 2) format conversion, 3) removing special characters in sequence names, 4) sequence name encoding and 5) sequence name decoding.

Removing duplicate sequences:
In large sequence files, it is often difficult to check for duplicate sequences and the tools may generate error messages because of this redundancy. SeqMaT provides two ways to identify the redundant sequences from the given dataset for all the major sequence formats. One way is removing identical sequences; the other way is to collapse sequences above a user provided threshold (with one representative sequence used for subsequent analysis). The latter requires aligned sequences as input.

Data format conversion:
Data format conversion can be used for converting sequences from one format to another for future analysis. SeqMaT allows conversions among all major sequence and tree formats. We have provided both sequential and interleaved options for Phylip and Mega formats.

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 9 **Software**

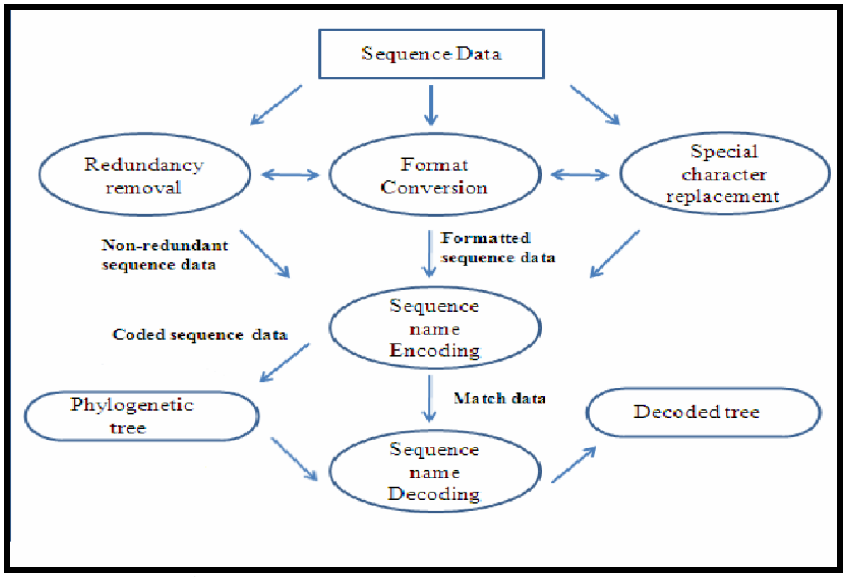

Figure 1: Flow chart of SeqMaT

Special characters replacement:
Most of the tools do not allow special characters such as '/' or '|' in sequence descriptions. SeqMaT provides an option for replacing the special characters and white space with an underscore character '_' in the sequence description. Currently this option works for Nexus and FASTA sequence formats.

Sequence name coding and encoding:
Automatically trimmed sequence names created by other format conversion tools may be redundant and cannot be accepted by analysis programs. SeqMaT efficiently addresses the problem of long sequence names. This includes a two-step process, sequence encoding and tree decoding.

The coding part takes a set of input sequences in FASTA, Clustal or Mega formats and for each sequence the whole description is replaced with a short alpha-numerical name. These temporary names and original descriptions are stored in a table. Two files, the match table with original and new names and a sequence file in which each sequence has a short unique name are given as output from this step. These converted sequences are easy to use in a tree analysis.

The decoding part takes the match table generated in the coding step and a tree file as input. It replaces the temporary names in the tree file with the original name (or description) from the match table. The resulting tree file has the original description for each sequence without any redundancies in it, which can be opened in any tree viewing program.

Other applications:
SeqMaT also provides additional functions to manipulate data for data analysis and mining. First, the sequence data in FASTA format can be converted to Attribute Relational File Format (ARFF), a native format of the data mining tool - WEKA. Secondly, it calculates individual residue frequencies and k-mer frequencies for a given set of sequences, with both protein and nucleotide sequences accepted. Thirdly, a user can select a particular number of representative sequences that will be randomly picked from a large dataset. Lastly, a user can extract sequences within certain dates and geographical locations. This application takes the starting and ending years from which the user needs to extract sequences and the number of sequences for each year. The program extracts the sequence data per the user choice. This function is particularly useful when collecting sequences for calculating evolutionary rates and the time of most recent common ancestry.

Software platform:
The background programs were written in Java and the web integration was done using JSP/Servlets. The standalone version is provided as a jar file and runs on any systems with the Java Runtime Environment installed.

Applications:
SeqMaT is currently used by the Bioinformatics lab at the University of Nebraska at Omaha and elsewhere.

Acknowledgements:
We are grateful to Mohammad Shafiullah for his assistance in computer systems administration. This publication was made possible by the grants from National Institutes of Health R01 LM009985-01A1 and National Science Foundation DEB-0732969. The authors also acknowledge the UCRCA, the University of Nebraska at Omaha, for continuous funding support to this research.

References:
[1] DJ Lipman & WR Pearson. *Science* **227**: 1435 (1985) [PMID: 2983426]
[2] J Felsenstein. *Cladistics* **5**: 164 (1989)
[3] D Gilbert. *Curr. Protoc. Bioinform.* (2003) [PMID: 18428689]
[4] http://www.bugaco.com/converter/ biology/sequences/
[5] http://hcv.lanl.gov/content/sequence/ FORMAT_CONVERSION/FormatExplain.html
[6] P Rice *et al. Trends in Genetics* **16**: 276 (2000) [PMID: 10827456]
[7] G Leonard *et al. Evolutionary Bioinformatics* 5:1 (2009) [PMID: 19812722]
[8] RC Holland *et al. Bioinformatics* **24**:397 (2008) [PMID: 18689808]
[9] JE Stajich *et al. Genome Res* **12**: 1611 (2002) [PMID: 12368254]

Edited by P Kangueane
Citation: Attaluri *et al.* Bioinformation 5(9): 400-401 (2011)