Computational Tools for Parsimony Phylogenetic Analysis of Omics Data

Jose Salazar,¹ Hakima Amri,² David Noursi,¹ and Mones Abu-Asab¹

Abstract

High-throughput assays from genomics, proteomics, metabolomics, and next generation sequencing produce massive omics datasets that are challenging to analyze in biological or clinical contexts. Thus far, there is no publicly available program for converting quantitative omics data into input formats to be used in off-the-shelf robust phylogenetic programs. To the best of our knowledge, this is the first report on creation of two Windows-based programs, *OmicsTract* and *SynpExtractor*, to address this gap. We note, as a way of introduction and development of these programs, that one particularly useful bioinformatics inferential modeling is the phylogenetic cladogram. Cladograms are multidimensional tools that show the relatedness between subgroups of healthy and diseased individuals and the latter's shared aberrations; they also reveal some characteristics of a disease that would not otherwise be apparent by other analytical methods. The *OmicsTract* and *SynpExtractor* were written for the respective tasks of (1) accommodating advanced phylogenetic parsimony analysis (through standard programs of MIX [from PHYLIP] and TNT), and (2) extracting shared aberrations at the cladogram nodes. *OmicsTract* converts comma-delimited data tables through assigning each data point into a binary value ("0" for normal states and "1" for abnormal states) then outputs the converted data tables into the proper input file formats for MIX or with embedded commands for TNT. *SynpExtractor* uses outfiles from MIX and TNT to extract the shared aberrations of each node of the cladogram, matching them with identifying labels from the dataset and exporting them into a comma-delimited file. Labels may be gene identifiers in gene-expression datasets or m/z values in mass spectrometry datasets. By automating these steps, *OmicsTract* and *SynpExtractor* offer a veritable opportunity for rapid and standardized phylogenetic analyses of omics data; their model can also be extended to next generation sequencing (NGS) data. We make *OmicsTract* and *SynpExtractor* publicly and freely available for non-commercial use in order to strengthen and build capacity for the phylogenetic paradigm of omics analysis.

Introduction

HIGH-THROUGHPUT ASSAYS FROM genomics, proteomics, metabolomics, and Next Generation Sequencing produce massive sets of data that are challenging to analyze in biological or clinical contexts. Two characteristics of omics data render their quantitative datasets difficult to analyze: the large number of variables (high dimensionality, usually in the tens of thousands) and their heterogeneous nature (Abu-Asab et al., 2008b; Abu-Asab and DeLeo, 2014; Amir el et al., 2013; Braun, 2014). Heterogeneity here refers to the asynchronicity (incongruence) of variables among specimens of the same class or subtype (e.g., subgroups of diseased specimens). Such heterogeneous complexity is evolutionary in nature and prevalent in developmental pathways (Joshi and

Gottgens, 2011) and the pathogenesis of diseases such as cancer or degenerative diseases (Abu-Asab et al., 2012; 2013b; Hiley and Swanton, 2014; West et al., 2010). Thus far, statistical methods used for the analysis of traditional clinical data have been proven to be substantially less effective on these high-dimensional heterogeneous datasets because parametric tests tend to homogenize, or average out, the variation within the data or use arbitrary discretions (Abu-Asab et al., 2008b; Fletcher, 2007; Gorlov et al., 2014).

Alternatively, systems biology methods provide a comprehensive analytical paradigm that is most suitable for omics data analysis, especially when using recent software such as TNT (Abu-Asab et al., 2013a; Braun, 2014; Goloboff, 1999). Some of the important goals of medical bioinformatics include modeling and subtyping heterogeneous diseases,

¹Section of Immunopathology, Laboratory of Immunology, National Eye Institute, Bethesda, Maryland.

²Department of Biochemistry and Cellular and Molecular Biology, Division of Integrative Physiology, Medical Center, Georgetown University, Washington, District of Columbia.

providing methods for early detection, assessing treatment and prognosis, and uncovering other useful insights, such as detection of susceptibility to develop disease. Over the past 10 years, we have tested a number of phylogenetic methods such as maximum likelihood, Bayesian, neighbor-joining, and parsimony; we have selected parsimony as the standard analytical tool for biomedical omics data. Parsimony has been demonstrated to be an effective algorithm for analyzing high-dimensional heterogeneous data (Abu-Asab et al., 2008a; Kolaczowski and Thornton, 2004); additionally, parsimony traces change from normal to abnormal on the cladogram and produces a list of the shared aberration (the synapomorphies) of a number of related specimens (a clade) (Abu-Asab et al., 2011). Unlike the clustering that is specimen-based, parsimony is data-based and its cladogram presents the complex data in the most parsimonious distribution of the data matrix (Abu-Asab et al., 2008a).

Cladogram construction (Fig. 1), however, is a computationally complex NP-hard task (Rice and Warnow, 1997). In general, the cladograms that best represent the data cannot be constructed deterministically, nor can they be exhaustively searched for in a reasonable amount of time; good cladograms must be found heuristically. Programs that are designed to produce accurate cladograms efficiently use complex search algorithms that employ concepts such as maximum parsimony, which relies on the presumption that the simplest cladogram (the one that plots the data using the fewest state transitions) is also the most accurate (Goloboff and Pol, 2007; Kolaczowski and Thornton, 2004). This presumption is sometimes referred to as Occam's razor.

FIG. 1. A summary cladogram of prostate tissue specimens showing the relationship between four prostate tissue types: 1) normal prostate (N); 2) normal prostate adjacent to tumor (Ad.); 3) primary prostate tumor (P); and 4) metastatic prostate tumor (M). The cladogram shows one clade of normal specimens (N), two clades of adjacent specimens (Ad.), three mixed clades of adjacent and primary specimens (Ad.+P), one mixed clade of primary and metastatic specimens (P&M), and three clades of metastatic specimens (M). Notice the gradient that the cladogram shows from normal → Adjacent → Adjacent+Primary → Primary+Metastatic → Metastatic. Dataset GDS2545 was downloaded from: <http://www.ncbi.nlm.nih.gov/geo/>; the dataset originally generated by Yu et al. (2004). Data were prepared with OmicsTract and processed through TNT and the accuracy verified with manual calculations (See Supplementary Data 1 and 2).

However, the developers of search software cannot accommodate every possible raw data format. Their programs require specific data formats that are often different from the original raw data. Converting between data formats is a common practice when using multiple data-handling programs, and it is a time-consuming task when it is not automated. Additionally, a uniform data input allows seamless integration of datasets [for an example, see Abu-Asab et al., (2008b)]

Thus far, there is not any publicly available program for converting quantitative omics data into input formats to be used in off-the-shelf robust phylogenetic programs such as MIX (Mixed parsimony algorithm: <http://evolution.genetics.washington.edu/phylip/doc/mix.html>) and TNT (Tree analysis using New Technology: <http://www.cladistics.com/aboutTNT.html>) (Felsenstein, 1989; Goloboff, 1999). We are the first to create two Windows-based programs, OmicsTract and SynpExtractor, to address this issue.

OmicsTract automatically extracts, or prepares, gene-expression microarray or mass spectrometry data into the input file format of MIX or TNT to be used to generate cladograms. To do this, OmicsTract requires a set of control specimens as a reference of normal range in order to sort out the abnormal values in the experimental specimens. The program was designed to be used on data from pathology and treatment studies, and for the stratification of a patient population. SynpExtractor uses the *outfiles* of MIX and TNT to match the synapomorphies of each node of the cladogram with the gene identifier in gene-expression datasets or the m/z values in mass spectrometry datasets.

The authors are making OmicsTract and SynpExtractor publicly and freely available for non-commercial use in order to strengthen the phylogenetic paradigm of omics analysis.

Materials and Methods

OmicsTract and SynpExtractor are Windows (XP and newer versions) applications that were designed and developed in Microsoft Visual Studio's Visual C# language (<http://www.microsoft.com/en-us/download/details.aspx?id=30681>). The accuracy of OmicsTract was tested by comparing its output file with one generated manually on a spreadsheet (see Supplementary Data S1 and S2; supplementary material is available online at www.liebertpub.com/omi).

OmicsTract: User Procedure

Selecting data files. The OmicsTract program has a graphical interface dialogue window (Fig. 2) that allows the user to select a comma-separated-values (CSV) data file, which then appears in the leftmost pane. Clicking on a file name in that pane displays the data table in the center pane. The user marks the columns of the control data and the columns of experimental data (in the context of biomedicine, these are respectively healthy specimens' data and diseased specimens' data). After clicking the button labeled "Define control data," for example, the user would select the column headers for each control specimen before hitting "Done." The program will highlight identified control specimens in green and experimental specimens in red.

FIG. 2. OmicsTract program interface with an example dataset opened. After selecting a CSV file from which to read data, the user must specify the columns containing the data. The program needs separate control specimens and diseased specimens in order to carry out polarity assessment of data values properly. The user also specifies which parsimony program(s) should be used (MIX or TNT), and can modify an additional rule (“Polarization type”) for TNT. The program does not open the file or begin processing until the lower-right “Process selected datasets” button is pressed.

Selecting output format. The user also specifies for which of the two programs the output should be formatted. The first, MIX (Mixed Method Parsimony), is a program from PHYLIP package (Phylogeny Inference Package) (Felsenstein, 1989). The second, TNT (Tree analysis using New Technology) (Goloboff, 1999), is recommended because it is newer and faster than MIX, but either or both programs can be selected.

Selecting polarity format. There are two additional options, one of which is the choice between the default “1, 0, 1” pattern or “1, 0, 2.” The latter is trivalent, meaning that abnormalities are further differentiated by whether they are above or below the controls’ range for every variable. The other option enables the user to polarize both the control data as well as experimental data; usually only the experimental data is polarized.

Saving the output file. Upon clicking the “Process selected datasets” button, the user is prompted by a save dialog, after which the program begins to process the selected datasets. Example input files for MIX and TNT are shown in Figure 3.

OmicsTract: Program procedure

OmicsTract algorithmic procedure is shown in Table 1; the program reads and processes the selected CSV files sequentially, polarizing each row in the file to the end of the table. The algorithm iterates over the user-specified control values in a row to find the highest and lowest values and creates the range of healthy values for that row. Then, for the experimental values of the same row, a zero (0) is recorded if the value is within the healthy range (i.e., between the minimum and maximum); otherwise, a one (1) is recorded (i.e., if the value is outside of the healthy range). This process is termed

polarity assessment, which converts the continuous data into binary data of 0s and 1s. If the “1 | 0 | 2” option was selected, abnormal lows are represented by 1 and abnormal highs are represented by 2. These values are saved in the program’s memory until they are ready to be written to an output file. The original file is not modified during this process.

After it completes polarity assessment, OmicsTract writes the input file for MIX, TNT, or both. The two file formats are text-based; MIX files are saved with the “.txt” extension, while TNT files are saved as “.tnt.” Double-clicking a file with a “.tnt” extension runs TNT if it is installed on the computer. OmicsTract writes these input files by creating a new text file and writing the polarized values to it in the necessary format. The files are saved with the file name previously specified by the user in the save prompt. Once the output files have been saved, the program offers to remove the finished datasets from the list in the left pane.

OmicsTract-generated input file for MIX and TNT

The proper input file format is needed to successfully use MIX or TNT. This section describes the input files of MIX and TNT.

MIX input files adhere to a format common to all PHYLIP programs (Felsenstein, 1989). The first line contains the number of specimens and number of characters, where the latter is the same for all specimens of the dataset; and the two numbers are separated by five spaces. The remaining lines contain the data; the first 10 characters are used for the name of the specimen, followed by the bivalent values for that specimen. A new line is started only for a new specimen. After the last specimen, the file ends.

For TNT files, text commands can prepare the program settings in advance. The TNT input files produced by OmicsTract automatically set basic memory and data parameters

FIG. 4. SynpExtractor program interface. Two files are required to run SynpExtractor: the first is an outfile from either MIX or TNT, and the second is a list of identifiers of the data variables (gene-list, m/z values, etc.). SynpExtractor will produce a comma-separated file containing the synapomorphies of every node of the cladogram as listed in the outfile.

Creating the output file. Clicking the “Create table” button will execute the program, and the user is prompted to name the output file and set its save location. When opening the output CSV file with Excel, the user will find the nodes and their synapomorphies arranged into columns.

SynpExtractor: Program procedure

The SynpExtractor program procedure is summarized in Table 2. The program reads and extracts the chosen nodes of the cladogram from the *outfile* of MIX or TNT, identifying

synapomorphies and matching them with identifiers from the provided list of gene or m/z labels. The program compiles this information into a CSV file, where each of the processed nodes has a column.

Downloading the programs

OmicsTract and SynpExtractor are freely available to the public for academic research and teaching only, and not for commercial use or resale. To download the programs, visit: [http:// software.phylomics.com/](http://software.phylomics.com/).

TABLE 2. SYNPEXTRACTOR PROGRAM PROCEDURE

<p>$P(N, i)$ is the polarity value of the synapomorphy at character i at the given node N. This value may be 1, 0, or “?” (undetermined).</p> <p>$L(i)$ is the text label for character number i.</p> <p>Since the outputs from MIX and TNT do not contain character labels and only identify characters by their numbered position in the dataset, a separate file containing the labels is necessary for SynpExtractor to define $L(i)$.</p> <p>MIX</p> <p>MIX’s output files contain, for each node N, a list of values M_N such that $M_{N,x}=P(N, x)$.</p> <p>For each node N selected by the user, SynpExtractor reads M_N and constructs an alphabetized set A_N:</p> <p>$A_N=\{ L(x) \mid P(N, x)=1 \}$.</p> <p>SynpExtractor then writes the name of N, followed by each entry of A_N, to a column in its output CSV file.</p> <p>TNT</p> <p>TNT’s output files contain, for each node N, a list of values T_N:</p> <p>$T_N=\{ (x, P(N, x)) \mid P(N, x)=0 \vee P(N, x)=1 \}$</p> <p>For each node N selected by the user, SynpExtractor reads T_N and constructs A_N:</p> <p>$A_N=\{ L(x) \mid P(N, x)=1 \}$.</p> <p>SynpExtractor then writes the name of N, followed by each entry of A_N, to a column in its output CSV file.</p>
--

Results

We are illustrating the use of OmicsTract and SynpExtractor with a gene-expression dataset (GDS2545) downloaded from NCBI's GEO (<http://www.ncbi.nlm.nih.gov/geo/>). MIX and TNT infiles were created from GDS2545 dataset with OmicsTract. The first specimen/subject in the infiles corresponds to the first column in the raw data table. The full lengths of the input files are not shown (Fig. 3); however, a full length data of a specimen can be viewed in the supplementary data (see Supplementary Data S1). The sequence of specimens follows that of the original raw data table. A summary of the generated cladogram from the *infiles* is shown in Figure 1.

The synapomorphies of representative nodes of the cladogram as produced by SynpExtractor can be viewed in the supplementary data (see Supplementary Data S3).

Discussion

Cladograms have proven to be useful for modeling complex diseases, such as cancers and degenerative diseases including age-related macular degeneration (AMD), and developmental pathways (Abu-Asab et al., 2011; 2013b; Brim et al., 2012; 2014; Joshi and Gottgens, 2011). The cladograms group individuals based on their shared derived modification, which facilitates disease subtyping as well as the placement of individuals within groups according to their health status. If a person is asymptomatic but is grouped near diseased specimens on a cladogram, then that person may be genetically progressing towards the disease (Abu-Asab et al., 2013a). The potential use of cladograms in early detection of cancers and degenerative diseases is of great importance; early detection can be the difference between treatable and untreatable conditions for a patient (Glorikian, 2014).

Cladogram generating programs also identify synapomorphies, or information common to all individuals in the cladogram or one of its subgroups, termed clades in phylogenetic terminology (Joshi and Gottgens, 2011). If all diseased groups on a cladogram share some gene modifications that are not present in the healthy individuals, then those modifications may serve as biomarkers of the disease or one or more of its subtypes. Genetic modifications associated with a disease are important not only for understanding disease pathogenesis, but also for early detection and diagnosis of the disease (Abu-Asab et al., 2011; Ferreira et al., 2012).

Furthermore, the present work has shown that the cladogram can model the dynamic progression of disease because it plots the specimens into hierarchical progressive arrangement where the specimens at the top of the cladograms have the highest number of synapomorphies such as gene expression aberrations (Abu-Asab et al., 2011). Thus, the cladogram represents the disease on a continuous spectrum from healthy to the most diseased, and consequently can provide opportunities for early detection, subtyping, and individualized assessment of treatment, as well as susceptibility to develop the disease (Abu-Asab et al., 2013a).

OmicsTract and SynpExtractor can be applied in both research and clinical settings. They make the process of generating cladograms more accessible as an analytical paradigm for an array of high-throughput omics data; this is because both programs handle omics data preparation automatically and effortlessly. In research, they assist in model-

ing complex diseases and shed light on a disease's subtypes and genetic clonal aberrations. In clinical applications, these two programs will contribute to identifying an individual's health status by placing them within the disease spectrum on a cladogram; thus detecting their susceptibility and early disease transformations prior to the development of clinical symptoms and manifestations.

Even though OmicsTract supports only the polarity assessment, the software can be extended in the future to include more output formats. An example of a process that would require a different format is the analysis of continuous characters in TNT, which avoids transforming each datum into one of two values. This method thus avoids simplifying data, and it seems like an appropriate next step for OmicsTract to incorporate.

OmicsTract saves users from the arduous task of re-searching input formats and it automates the time-consuming data-conversion step of phylogenetic analysis, while SynpExtractor produces all the potential biomarkers of the disease and its subtypes. Although the usefulness of cladograms has been demonstrated before, the availability and convenience of having these two programs will encourage researchers to explore and carry out phylogenetic analysis, which we think is very appropriate to handling big data in biomedicine.

Conclusions

The new software tools, OmicsTract and SynpExtractor, facilitate phylogenetic analysis of omics data by automatically preparing the data for two parsimony phylogenetic programs, MIX and TNT. The phylogenetic cladograms generated by these programs are useful in modeling complex diseases, with the capacity for early detection and diagnosis, as well as disease subtyping. Cladograms are a way to analyze highly dimensional omics data while still accounting for their heterogeneity, and visualize a summary of the results on a phylogenetic tree.

Taken together, OmicsTract bridges the gap between raw omics data (such as mass spectrometry metabolomics and proteomics, as well as gene-expression microarray) and parsimony analysis by automatically extracting and formatting the data for parsimony phylogenetic analysis, and saving users time and effort. SynpExtractor lists synapomorphies of all the clades and nodes of the generated cladogram. In the future, OmicsTract may be extended to work with more data types than microarray and mass spectrometry, or more friendly analysis programs, which would increase its functionality and usefulness to researchers and clinicians alike.

Acknowledgments

The research presented here was partially supported by the intramural program of the National Institutes of Health.

Author Disclosure Statement

The authors declare they have no conflicting financial interests.

References

- Abu-Asab M, Chaouchi M, and Amri H. (2008a). Evolutionary medicine: A meaningful connection between omics, disease, and treatment. *Proteomics Clin Appl* 2, 122–134.

- Abu-Asab M, Koithan M, Shaver J, and Amri H. (2012). Analyzing heterogeneous complexity in complementary and alternative medicine research: A systems biology solution via parsimony phylogenetics. *Forsch Komplementmed* 19 Suppl 1, 42–48.
- Abu-Asab MS, Abu-Asab N, Loffredo CA, Clarke R, and Amri H. (2013a). Identifying early events of gene expression in breast cancer with systems biology phylogenetics. *Cytogenet Genome Res* 139, 206–214.
- Abu-Asab MS, Chaouchi M, Alesci S, et al. (2011). Biomarkers in the age of omics: Time for a systems biology approach. *Omics* 15, 105–112.
- Abu-Asab MS, Chaouchi M, and Amri H. (2008b). Phylogenetic modeling of heterogeneous gene-expression microarray data from cancerous specimens. *Omics* 12, 183–199.
- Abu-Asab MS, and Deleo J. (2014). Phylogenetic cladograms: Tools for analyzing biomedical data. In: Kasabov, N. (Ed.), *Springer Handbook of Bio-/Neuroinformatics*. Springer Berlin Heidelberg, pp. 247–263.
- Abu-Asab MS, Salazar J, Tuo J, and Chan CC. 2013b. Systems biology profiling of AMD on the basis of gene expression. *J Ophthalmol* 2013, 453934.
- Amir El AD, Davis KL, Tadmor MD, et al. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol* 31, 545–552.
- Braun R. (2014). Systems analysis of high-throughput data. *Adv Exp Med Biol* 844, 153–187.
- Brim H, Abu-Asab MS, Nouraie M, et al. (2014). An integrative CGH, MSI and candidate genes methylation analysis of colorectal tumors. *PLoS One* 9, e82185.
- Brim H, Lee E, Abu-Asab MS, et al. (2012). Genomic aberrations in an African American colorectal cancer cohort reveals a MSI-specific profile and chromosome X amplification in male patients. *PLoS One* 7, e40392.
- Felsenstein J. (1989). PHYLIP: Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164–166.
- Ferreiro E, Baldeiras I, Ferreira IL, et al. (2012). Mitochondrial and endoplasmic reticulum-associated oxidative stress in Alzheimer's disease: From pathogenesis to biomarkers. *Int J Cell Biol* 2012, 735206.
- Fletcher J. (2007). What is heterogeneity and is it important? *BMJ* 334, 94–96.
- Glorikian H. (2014). Early detection: The real way to beat cancer. *Clin OMICS* 15, 14–16.
- Goloboff PA. (1999). Analyzing large data sets in reasonable times: Solutions for composite optima. *Cladistics* 15, 415–428.
- Goloboff PA, and Pol D. (2007). On divide-and-conquer strategies for parsimony analysis of large data sets: Rec-I-DCM3 versus TNT. *Syst Biol* 56, 485–495.
- Gorlov IP, Yang JY, Byun J, Logothetis C, Gorlova OY, Do KA, and Amos C. (2014). How to get the most from microarray data: Advice from reverse genomics. *BMC Genomics* 15, 223.
- Hiley CT, and Swanton C. (2014). Spatial and temporal cancer evolution: Causes and consequences of tumour diversity. *Clin Med* 14, s33–37.
- Joshi A, and Gottgens B. (2011). Maximum parsimony analysis of gene expression profiles permits the reconstruction of developmental cell lineage trees. *Dev Biol* 353, 440–447.
- Kolaczowski B, and Thornton JW. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431, 980–984.
- Rice K, and Warnow T. (1997). Parsimony is hard to beat! In: Jiang T, Lee DT (Eds), *Computing and Combinatorics*. Springer Berlin Heidelberg, pp. 124–133.
- West SL, Gartlehner G, Mansfield AJ, et al. (2010). *Comparative Effectiveness Review Methods: Clinical Heterogeneity [Internet]*. (Agency for Healthcare Research and Quality (US); Available from: <http://www.ncbi.nlm.nih.gov/books/NBK53310/toc/>, Rockville (MD)). Last access: 5/21/15.
- Yu YP, Landsittel D, Jing L, et al. (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J Clin Oncol* 22, 2790–2799.

Address correspondence to:
 Dr. Mones Abu-Asab
 Section of Immunopathology
 Laboratory of Immunology
 National Eye Institute
 Bethesda, MD 20892

E-mail: mones@mail.nih.gov