Chapter 8

Application of MATLAB in *-Omics* and Systems Biology

Arsen Arakelyan, Lilit Nersisyan and Anna Hakobyan

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/62847

Abstract

Biological data analysis has dramatically changed since the introduction of high-throughput *-omics* technologies, such as microarrays and next-generation sequencing. The key advantage of obtaining thousands of measurements from a single sample soon became a bottleneck limiting transformation of generated data into knowledge. It has become apparent that traditional statistical approaches are not suited to solve problems in the new reality of "big biological data." From the other side, traditional computing languages such as C/C++ and Java, are not flexible enough to allow for quick development and testing of new algorithms, while MATLAB provides a powerful computing environment and a variety of sophisticated toolboxes for performing complex bioinformatics calculations.

We have used MATLAB to develop the pathway signal flow (PSF) algorithm for assessment of pathway activity changes based on high-throughput gene expression and pathway topologies. Additionally, we have created a KEGGParser tool for parsing, editing, and visualizing Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps. We have used these tools to obtain a collection of KEGG pathways and evaluate their activity changes in different clinical forms of pulmonary sarcoidosis (PS). The application of PSF provided an extended systems view on pathway deregulation states and implicated several new pathways in sarcoidosis that had not been identified using other analysis approaches.

Keywords: biological data mining/visualization, *-omics* data analysis, pathway visualization, pathway flow analysis, KEGG pathway database

1. Introduction

Biology and biomedicine have always been quantitative scientific disciplines and data collection has been an important part of biological knowledge inference. Biological data types

are very diverse and their nature has dramatically changed since introduction of new measurement technologies starting from the mid-twentieth century. DNA/RNA/protein sequencing [1, 2], X-ray crystallography [3], antibody-based assays [4, 5], and various modifications of polymerase chain reaction (PCR) [6, 7], allowed for collection of data about various aspects of cell function in normal and diseased conditions. A few of the most common biodata types include but are not limited to:

- Sequences—the one-dimensional orderings of monomers (DNA/RNA/proteins).
- Graphs—representation of a set of objects and pairwise interactions between them (pathway maps, protein–protein interaction nets).
- High-dimensional data—each sample is defined by hundreds or thousands of measurements, usually concurrently obtained (e.g., high-throughput gene expression data; see below).
- Geometric information—information about 3D structure of proteins, lipids, and nucleic acids.
- Patterns—regularities that characterize biological entities (transcription factor binding sites, network motifs, etc.).
- Models—mathematical or visual representations of dynamic or static behavior of biological entities.
- Literature—biological literature itself can be regarded as data to be exploited via data/text mining for revealing findings that would otherwise go undiscovered.

Traditionally, statistics has played significant role in biological data analysis [8]. It is used in biomarker identification [9–11], testing new drugs [12], analysis of genetic associations [13, 14], understanding associations between the levels of different biomolecules, experiment planning, etc. This has been made possible due to several key factors brought together. First of all, introduction of the "data analysis" concept was made by Tukey [15], implying that there is no need to be a statistician to be able to analyze and interpret the data. Next, the active penetration of computers into various research fields, as well as development of computer software and data analysis packages that can be used by "non-statisticians," has greatly enforced the progress in biomedical data analysis.

A new era of biomedical research has started in the twenty-first century with the advent of massively parallel measurement technologies. Traditional data collection approaches (low-throughput) are mainly focused on measurement of a few carefully selected parameters from a large number of samples, while high-throughput methods, such as microarrays, next-generation sequencing, and proteomics approaches, allow for acquiring dozens or hundred thousand observations from a single sample. Low-throughput techniques still remain an important tool in biomedical research; however, only high-throughput approaches provide global outlook on complex biological processes occurring at cellular, tissue, or even organismal level [16, 17]. This breakthrough has also changed the main paradigm in biological data representation, shifting from datasets containing large number of observations with few

http://dx.doi.org/10.5772/62847

variables (i.e., attributes or entries in the data vector) to datasets containing more variables than observations (high-dimension, low-sample size data (HDLSS)). HDLSS data is generated via various technologies measuring protein activity/abundance levels, gene expression levels (the amount of transcripts generated from each gene), ribonucleic acids (RNA) abundances, etc. For example, a typical high-throughput measurement of gene expression is performed for about 20,000 genes per subject, while the number of subjects rarely exceeds 10–20.

With this new data type, classical statistical approaches frequently fail to produce meaningful results because they are not designed to cope with this growth of dimensionality in datasets [18–20]. Thus, a demand for new algorithms and software for data analysis and visualization has emerged. From the other side, traditional computing languages, such as C/C++ and Java, are not flexible enough to allow for quick development and testing of new algorithms. This is because in contrast to scripting languages they require compilation of the whole code before execution, definition of variable types, etc. The limitations of compiled languages pose the challenge of having professional software engineers (or dedicated persons with programming skills) for writing software. In contrast, scripting languages allow for execution of separate lines of the script, without caring for variable type definitions and memory allocation beforehand. They are commonly at higher language level and additionally are supported by a wide range of packages for specific scientific purposes.

In this sense, MATLAB provides a powerful computing environment and a variety of sophisticated toolboxes for performing complex bioinformatics calculations. In this chapter, we discuss the examples of MATLAB application for high-throughput gene expression data analysis and visualization based on the algorithms and software developed by our research group.

2. Analysis of gene expression data

Gene expression is the realization of the information stored in genes through synthesis of ribonucleic acids (RNA) and proteins that perform functional and structural activities in a cell and an organism [21]. Genes are DNA fragments that code for products such as proteins and functional RNA, including ribosomal RNA (rRNA), transfer RNA (tRNA), or small nuclear RNA (snRNA). According to the central dogma of molecular biology, protein coding genes are expressed in a two-stage process involving synthesis of a messenger RNA (mRNA) (transcription) and synthesis of a protein on the mRNA template (translation) (**Figure 1**) [21]. However, in biomedical research the term "gene expression" is often used to refer only to the transcription stage, and the gene expression value usually indicates the amount of mRNA produced from each gene.

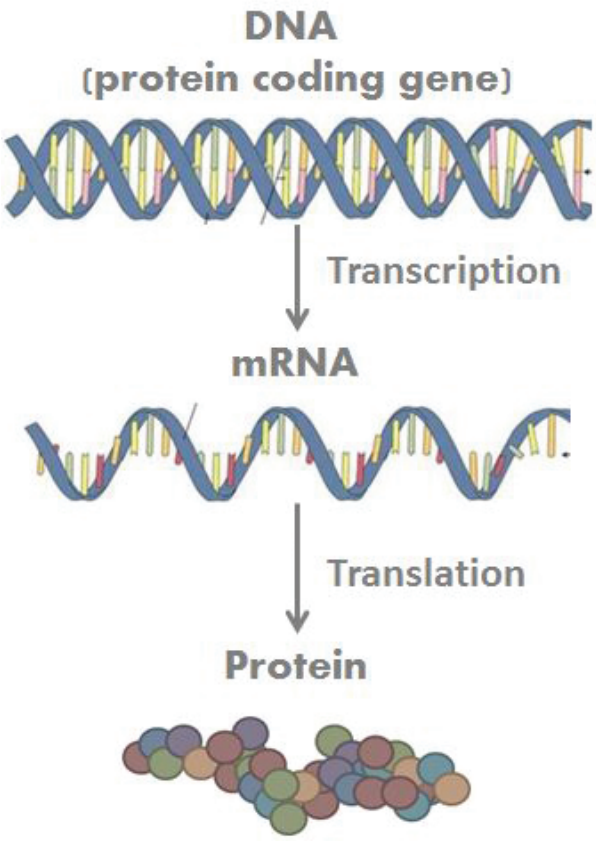

Figure 1. The synthesis of proteins from protein coding genes, according to the central dogma of molecular biology.

High-throughput analysis of gene expression is one of the cornerstones of systems biology [16]. The study of gene expression signatures has largely contributed to better understanding of molecular pathology of lung diseases [22, 23] and to identification of new disease subclasses/entities [24]. It has also provided new approaches to diagnostics [25] and helped to suggest novel therapeutic compounds [26].

There are two main techniques that allow for massively parallel measurement of gene expression in cells and tissues: microarrays and next-generation RNA sequencing (RNA-seq). The technology details of these approaches are summarized in **Figure 2** and are described in a number of publications elsewhere [2, 27–29]. The raw gene expression data for microarray and RNA-seq gene experiments are usually presented in a form of expression matrix. Each column represents all the gene expression levels for a single sample, and each row represents the expression of a gene across all the samples. This matrix serves as the source for subsequent analysis steps.

http://dx.doi.org/10.5772/62847

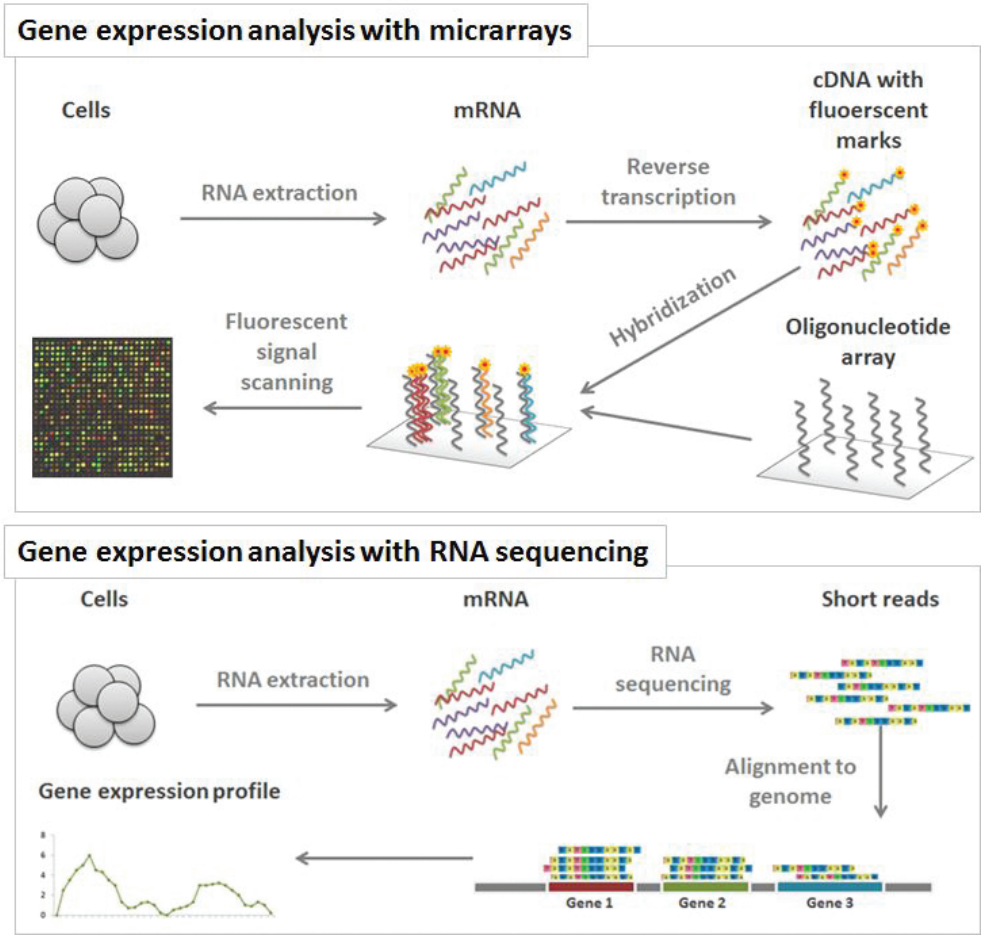

Figure 2. A general summary of two main techniques for gene expression assessment. Microarray-based techniques (at the top) are based on hybridization of complementary DNAs, obtained from RNA extracted from the cells, to specially designed oligonucleotide arrays and subsequent capturing of fluorescent signals coming from hybridized probes. The more the signal, the more RNA was produced from the gene corresponding to the respective oligonucleotide. RNA sequencing (at the bottom) utilizes next-generation sequencing technologies to obtain short sequencing reads from RNA extracted from the cells. These short reads are then aligned to the reference genome to determine the corresponding genes and to compute RNA abundance for each gene.

There are many different algorithms, also implemented in MATLAB, that are aimed at analysis of global gene expression [30–32]. MATLAB Bioinformatics toolbox demos are excellent start points to become acquainted with microarray and RNA-seq gene expression analysis. Most of these algorithms exploit a common gene-centered analysis pipeline, which identifies genes differentially expressed between studied conditions, and further annotates the gene lists by assessment of their relative abundance in predefined functional categories available

in biological databases [33], such as Gene Ontology (GO) [34], Kyoto Encyclopedia of Genes and Genomes (KEGG) [35], and others. A shortcoming of these approaches is that they overlook the interactions that exist between gene products in a cell. These interactions define the actual behavior of biological systems, along with expression values of the interacting agents. The information on interactions occurring between gene products is depicted in topologies of biological pathways, and thus, gene expression analysis that also accounts for topology information would be more informative about the state of pathways and activities of associated biological processes. Biological pathways are directed and spatially defined sequences of biomolecular physical and regulatory interactions that represent molecular signal propagations leading to functional realizations of biological processes. The behavior of a given pathway highly depends on both gene expression and its topology [36]. It is known that a significant portion of genes may be involved in more than one pathway, while perturbation of a single pathway may be conditioned by differential expression of multiple member genes. Moreover, a single disease may be characterized by orchestrated deregulation of several pathways. Finally, profiles of pathway deregulations may be common for different diseases. Thus, based on gene expression and pathway topology data, it is possible to identify common and specific pathways among diseases.

3. KEGGParser

KEGGParser is a MATLAB graph-based graphical user interface tool for parsing, editing, visualization, and analysis of biological pathway maps from Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. It is based solely on functions contained in MATLAB, MATLAB Bioinformatics toolbox version 3.x, and Image Processing Toolbox 2.x. KEGGParser is freely available at http://www.mathworks.com/MATLABcentral/fileexchange/37561.

KEGG pathways are stored in a collection of manually curated pathway maps. Each pathway is represented by an image and accompanying xml (KGML) file, which stores an xml tree structure containing information about nodes and edges. The KGML files are created from the map images with “KegSketch” program.

KEGG pathways can be accessed from MATLAB using KEGG REST interface implemented in MATLAB Bioinformatics toolbox. However, KEGG REST functions are very limited and intended for very basic operations, such as retrieval of pathway images, node information, and mapping of gene expression data via coloring of nodes on top of the map images, and are not suited for much wider range of pathway analysis needed. In contrast, KEGGParser uses information stored in pathway KGML files to convert them directly to MATLAB biograph objects (a graph representation of the pathways), which then allows for performing advanced pathway analysis. Biograph is a MATLAB data structure for implementation of directed graphs. Graph nodes can represent diverse biological agents, such as genes and proteins, and edges depict directed interactions between the agents, which can represent physical, regulatory interactions, dependencies, etc. The biograph object also stores graphical properties, such as color, labels, and sizes, used for 2D visualization.

Although the biograph object is ideally suited for storing and manipulating biological pathways, its severe limitation is the absence of generic methods for adding or deleting edges on already created graphs. This can seriously impede downstream analyses, because KGML files do not contain all the information depicted in the pathway map (Figures 3 and 4).

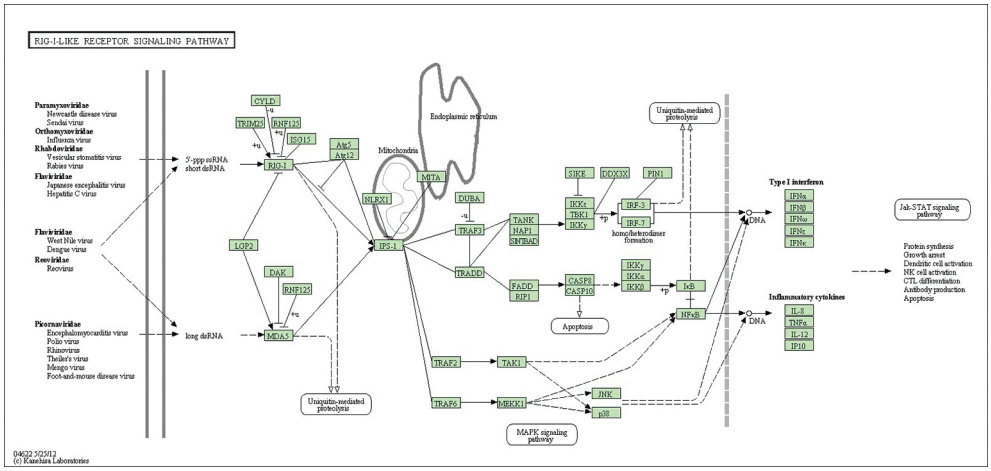

Figure 3. The RIG-I-like signaling pathway map image obtained from KEGG pathway database.

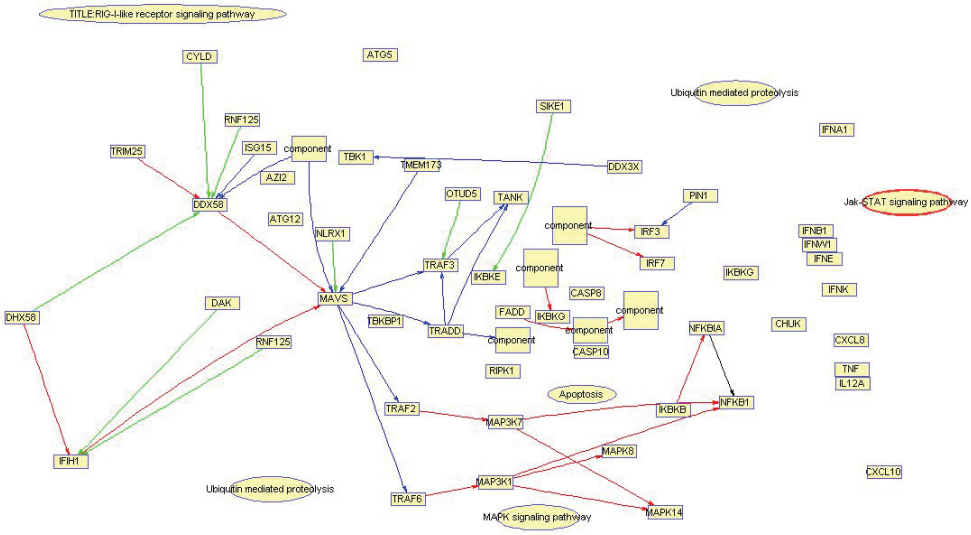

Figure 4. The RIG-I-like pathway map obtained by parsing the pathway KGML file.

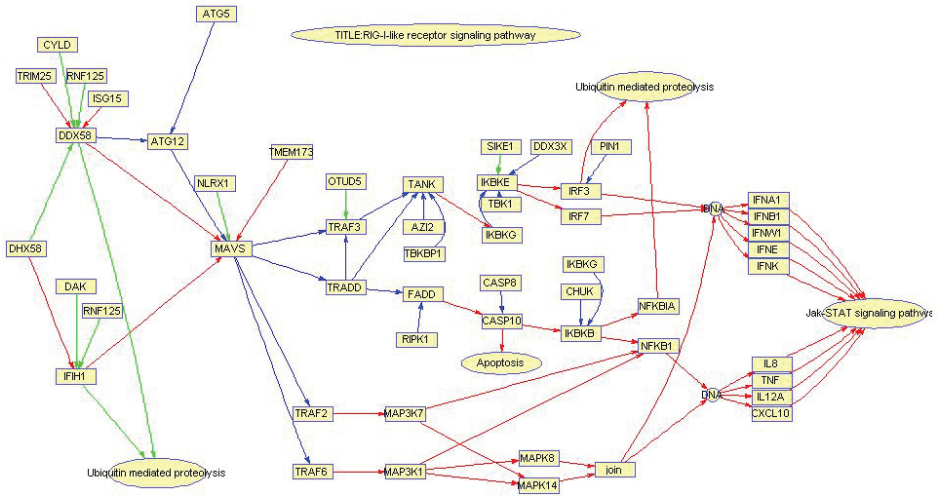

Figure 5. Manipulation of a sample biograph object with graph manipulation tools. The process of node and edge addition and deletion is depicted. See the main text for the respective example code.

Moreover, it is known that pathways can change their topology (i.e., interactions between nodes in a pathway) due to mutations, regulation of gene expression, , etc. In order to overcome these limitations, we have developed several functions that allow for graph manipulations on pre-calculated graphs. They are accessible from the link http://www.mathworks.com/MATLABcentral/fileexchange/37475. The following example and **Figure 5** demonstrate the biograph object editing process using graph manipulation tools.

```
% Create an adjacency matrix
cm = [0,0,1; 1,0,0; 0,1,0];
%convert adjacency matrix to biograph object
bg = biograph(cm);
%plot graph object
view(bg);
%add new node to existing graph
bg = node_add(bg);
%change node label
bg.Nodes(4).ID = 'Node 4';
view(bg);
%add new edge from node 4 to node 1
bg = edge_add(bg, 4, 1);
view(bg);
% delete created edge
bg = edge_del(bg, 4,1);
view(bg);
% delete node 4
bg = node_del(bg, 4);
view(bg);
```

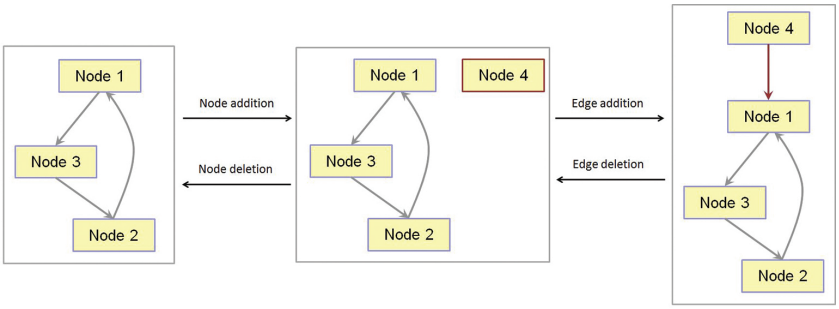

Figure 6. The RIG-I-like pathway map obtained by KGML file parsing and further corrections and recovery of missing information using KEGGParser.

The graph manipulation tools are used by KEGGParser for parsing KGML files, creating and editing biograph objects that represent KEGG pathway graphs. Along with creation of pathway graphs, KEGGParser automatically handles and respectively edits part of inconsistencies between KGML files and map images. The overall KEGGParser workflow and usage examples are described in detail in the original publication [37]. In this chapter we present two different use cases for this software.

The first example refers to retrieval, editing, and visualization of KEGG pathways using KEGGParser. As an example we have chosen RIG-I (retinoic acid-inducible gene 1) - like receptor signaling pathway. RIG-I-like receptors and downstream signaling are key elements in sensing viral pathogens and generating innate immune responses [38, 39]. Activation of this pathway is essential for production of various cytokines, mediators of inflammation, and proliferation of immune system cells. Deregulation of RIG-I-like pathway activity is implicated in many autoimmune disorders, such as systemic lupus erythematosus and Aicardi–Goutières syndrome [40]. In order to access the basic characteristics of the pathway topology, we used KEGGParser for retrieving and editing the corresponding KGML file. The pathway map from KEGG pathway database, as well as the native graph object parsed using KEGGParser, is presented in Figures 3 and 4 (static image and parsed without automatic corrections). As can be noticed, the parsed graph in MATLAB has many missing edges and unconnected nodes, making subsequent analysis improper. This is because KGML file contains information only about protein–protein interactions and the information about other types of interactions present in the map images is lost after KGML parsing. Using KEGGParser, we manually edited the pathway graph object and restored missing edges (**Figure 6**).

In order to perform graph theory-based analyses, we stored unedited and edited pathway graphs in rig_like.mat and rig_corr.mat files, respectively (available at [41]). Using graph theory functions implemented in MATLAB Bioinformatics toolbox, we performed some basic comparisons of unedited and manually edited pathway graphs. First we compared connectivities of nodes in the graphs:

```
%Comparison of edited and unedited KEGG graphs using MATLAB
%graph theory functions: node degrees
%
% Load the unedited graph, use load rig_cor for the edited one
load rig_like.mat
bgtemp = bg.deepCopy;
n = length(bgtemp.nodes);
mat = zeros(n,3);
for i = 1:n
        % all nodes connected to the given node
tot = length(getrelatives(bgtemp.nodes(i)))-1;
up = length(getancestors(bgtemp.nodes(i)))-1;
down = length(getdescendants(bgtemp.nodes(i)))-1;
mat(i,:) = [tot,up, down];
end

% Plot the histogram of node degrees
hist(mat(:,1))
clear all
```

Analysis of node degree histograms in the unedited graph shows significant skew toward 0-degree nodes, compared to the edited graph (**Figure 7A** and **B**).

```
%Comparison of edited and unedited KEGG graphs using MATLAB
%graph theory functions: analysis of connected components
%
% Load the unedited graph, use load rig_cor for the edited one
load rig_like.mat
bgtemp = bg.deepCopy;
[S, C] = conncomp(bgtemp, 'Directed', false)

% Plot the histogram of nodes in connected components
hist(C, unique(C))
clear all
```

The results showed that there are five strongly connected multi-node components (with four nodes on average) in the unedited pathway graph, while the edited graph identifies six components containing on average seven nodes (**Figure 7C** and **D**). These components correspond to nodes that form separate pathway branches and lead to different functional outcomes associated with pathway activation.

Finally, the distributions of all lengths of shortest paths between pairs in the unedited and edited graphs were significantly different, showing longer average path in the edited graph consistent with the static pathway map image (**Figure 7E** and **F**). Thus, KEGGParser can facilitate the better representation of biological pathways in MATLAB and contribute to the adequate analyses of pathway topologies. Manual/automatic editing options help to restore correct topologies of the pathways. Using KEGGParser, we have created a collection of

http://dx.doi.org/10.5772/62847

signaling biological pathways that were further used for analysis of pathway activity perturbations in various diseases (see pathway signal flow (PSF) section).

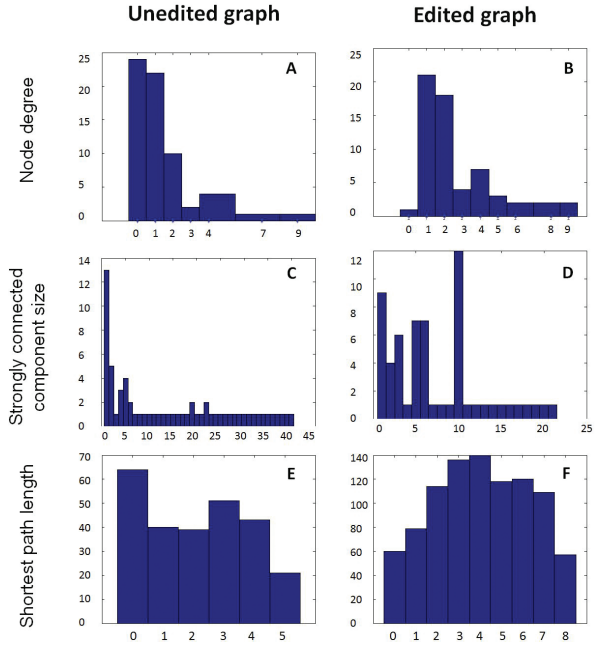

Figure 7. Graph characteristics of RIG-I-like signaling pathway after initial KGML parsing (unedited graph) and further editing to correct for missing nodes and edges (edited graph). (A and B) Node degree distributions; (C and D) the distributions of sizes of strongly connected components; and (E and F) the distribution of shortest path lengths.

4. Pathway signal flow

The PSF algorithm can be used to assess the changes in activity of a given biological pathway depending on the pathway topology and relative gene expression [42]. In contrast to the traditional gene-centered approaches for expression data analysis, PSF also takes into account the interactions between gene products (i.e., proteins, RNA, etc.) and other biological entities and, thus, provides richer outlook on actual molecular events associated with pathways.

This algorithm calculates how the activating signal is propagated from pathway input nodes, through interactions between intermediate node pairs, to the output nodes, leading to functional realization of associated biological processes. The amount of signal approaching the output nodes is called PSF. The extent of changes in the pathway flow is indicative of the likeliness of the given pathway to be involved in the biological processes underlying the phenotypic differences between the studied conditions. The detailed description of PSF is

given in a number of publications [42, 43]. Here we will bring an example of PSF usage for analysis of pathway flow perturbations in pulmonary sarcoidosis (PS) and its different clinical forms.

PS is a systemic granulomatous disease with unknown cause [44]. It is characterized by massive influx of activated T-lymphocytes and macrophages into the lungs, formation of granulomas, and lung function failure. The immune disturbances and and granulomatous deposits resolve spontaneously in 60–70% of PS patients; the rest follow a chronic course [44]. Though significant advances have been made in understanding of immunological features of this disease, the central pathomechanisms of its development are yet unknown. In this study, we aimed at identification of differentially deregulated pathways in sarcoidosis, as well as in different clinical forms of the disease, compared with healthy lung. We have used two microarray gene expression datasets from Gene Expression Omnibus public repository (dataset IDs: GDS3580 and GDS3705). The gene expression data and PSF scripts are available for download from [41].

The results of PSF analysis indicate that inflammation-related pathways are significantly upregulated in sarcoidosis compared to healthy lung, while pathways interfering with cell proliferation and fibrosis are downregulated (**Table 1**). Moreover, compared to self-limiting PS, the progressive form of the disease is characterized by more prominent deregulation of pathways associated with proinflammatory response and fibrotic conversion of the tissue (**Table 2**).

Pathway	**PSF**	***p* value**
PPAR (Peroxisome proliferator-activated receptor) signaling pathway	0.98	0.0000
Chemokine signaling pathway	33.05	0.0000
NF-kappa B signaling pathway	1.62	0.0000
Apoptosis	0.98	0.0000

Table 1. Pathway activity deregulation in pulmonary sarcoidosis compared to healthy lung.

Pathway	**PSF**	***p* value**
Fc gamma R-mediated phagocytosis	13.72	0.0000
Chemokine signaling pathway	12.65	0.0000
Fc epsilon RI signaling pathway	7.27	0.0000
Ras signaling pathway	6.55	0.0000
PI3K (Phosphoinositide 3-kinase)-Akt signaling pathway	2.98	0.0000
HIF-1 (Hypoxia-inducible factor 1) signaling pathway	2.62	0.0000
B-cell receptor signaling pathway	2.18	0.0000
NF-kappa B signaling pathway	2.17	0.0000
NOD (Nucleotide-binding oligomerization domain)-like receptor signaling pathway	1.81	0.0339

Pathway	PSF	*p* value
VEGF (Vascular-endothelial growth factor) signaling pathway	1.71	0.0000
MAPK (mitogen-activated protein kinases) signaling pathway	1.69	0.0000
p53 signaling pathway	1.33	0.0000
Apoptosis	1.26	0.0000

Table 2. Pathway activity deregulation in progressive versus self-limiting sarcoidosis.

These findings are in compliance with the existing knowledge on sarcoidosis pathogenesis. Numerous experimental studies, including our own, have implicated immune/inflammatory pathways, such as Toll-like receptor signaling, phagocytosis, and chemokine signaling in sarcoidosis [45–47]. However, application of PSF provided an extended systems view on pathway deregulation states. Moreover, PSF implicated several new pathways that were not detected using gene-centered analysis approaches in the original publications [48, 49].

5. Conclusions

This chapter demonstrates the advantages of MATLAB for performing bioinformatics research. The powerful scripting language combined with various toolboxes makes it a valuable tool for creation of complete pipelines for *-omics* data analysis and visualization. We have contributed to MATLAB Bioinformatics toolbox with the PSF algorithm for assessment of pathway activity changes, and KEGGParser for fine-tuned pathway editing and visualization. MATLAB GUI support allows for visualization of the results, making it easy to use for the general scientific audience. Combining our tools with the rest of the MATLAB Bioinformatics toolbox has the power of having various complete pipelines for high-throughput data analyses.

Finally, MATLAB scripting language allows for easy development and testing of various algorithms that can later be translated into other scripting and programming languages. It should be noted that KEGGParser and PSF, originally developed in MATLAB, were then ported to various programming and scripting environments, such as Java and R [43, 50].

Author details

Arsen Arakelyan[1,2*], Lilit Nersisyan[1,2] and Anna Hakobyan[1]

*Address all correspondence to: aarakelyan@sci.am

1 Bioinformatics Group, Institute of Molecular Biology MAS RA, Yerevan, Armenia

2 College of Science and Engineering, American University of Armenia, Yerevan, Armenia

References

[1] Moorthie S, Mattocks CJ, Wright CF. Review of massively parallel DNA sequencing technologies. Hugo J. 2011;5(1–4):1–12. DOI: 10.1007/s11568-011-9156-3.

[2] Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12(2):87–98. DOI: 10.1038/nrg2934.

[3] Shi Y. A glimpse of structural biology through X-ray crystallography. Cell. 2014;159(5): 995–1014. DOI: 10.1016/j.cell.2014.10.051.

[4] Kahlert H, Cromwell O. Monoclonal antibodies. Methods Mol Med. 2008;138:183–96.DOI: 10.1007/978–1–59745–366–0_15.

[5] Zhu K, Dietrich R, Didier A, Doyscher D, Märtlbauer E. Recent developments in antibody-based assays for the detection of bacterial toxins. Toxins (Basel). 2014;6(4): 1325–48. DOI: 10.3390/toxins6041325.

[6] Lo YM, Chan KC. Introduction to the polymerase chain reaction. Methods Mol Biol. 2006;336:1–10.

[7] Fakruddin M, Mannan KS, Chowdhury A, Mazumdar RM, Hossain MN, Islam S, Chowdhury MA. Nucleic acid amplification: alternative methods of polymerase chain reaction. J Pharm Bioallied Sci. 2013;5(4):245–52. DOI: 10.4103/0975-7406.120066.

[8] Bang H, Davidian M. In: Bang H, Zhou XK, van Epps HL, Mazumdar M, editors. Statistical Methods in Molecular Biology. 1st ed. Humana Press; 2010. 636 p. DOI: 10.1007/978-1-60761-580-4.

[9] Wang IM, Stone DJ, Nickle D, Loboda A, Puig O, Roberts C. Systems biology approach for new target and biomarker identification. Curr Top Microbiol Immunol. 2013;363:169–99. DOI: 10.1007/82_2012_252.

[10] Rajcevic U, Niclou SP, Jimenez CR. Proteomics strategies for target identification and biomarker discovery in cancer. Front Biosci (Landmark Ed). 2009;14:3292–303.

[11] Vandenbogaert M, Li-Thiao-Té S, Kaltenbach HM, Zhang R, Aittokallio T, Schwikowski B. Alignment of LC–MS images, with applications to biomarker discovery and protein identification. Proteomics. 2008;8(4):650–72. DOI: 10.1002/pmic.200700791.

[12] Ahuja V, Sharma S. Drug safety testing paradigm, current progress and future challenges: an overview. J Appl Toxicol. 2014;34(6):576–94. DOI: 10.1002/jat.2935.

[13] Srivastava K, Srivastava A. Comprehensive review of genetic association studies and meta-analyses on miRNA polymorphisms and cancer risk. PLoS One. 2012;7(11):e50966. DOI: 10.1371/journal.pone.0050966.

[14] Lee YH. Meta-analysis of genetic association studies. Ann Lab Med. 2015;35(3):283–7. DOI: 10.3343/alm.2015.35.3.283.

[15] Tukey JW. The future of data analysis. Ann Math Stat. 1962;33(1):1–67. DOI: 10.1214/aoms/1177704711.

[16] Braun R. Systems analysis of high-throughput data. Adv Exp Med Biol. 2014;844:153–87. DOI: 10.1007/978-1-4939-2095-2_8.

[17] Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97. DOI: 10.1016/j.molcel.2015.05.004.

[18] Li L. Dimension reduction for high-dimensional data. Methods Mol Biol. 2010;620:417–34. DOI: 10.1007/978-1-60761-580-4_14.

[19] Huang H, Liu Y, Yuan M, Marron JS. Statistical significance of clustering using soft thresholding. J Comp Graph Stat. 2015;24(4):975–93.

[20] Ibrahim M, Jassim S, Cawthorne MA, Langlands K. Multidimensionality of microarrays: statistical challenges and (im)possible solutions. Mol Oncol. 2011;5(2):190–6. DOI: 10.1016/j.molonc.2011.01.002.

[21] Weinzierl ROJ. Mechanisms of Gene Expression. Imperial College Press; 1999. 436 p. DOI: 10.1142/9781848160606_fmatter.

[22] Thakur RK, Yadav VK, Kumar A, Basundra R, Kar A, Halder R, Singh A, Kumar P, Baral A, Kumar MM, Pal K, Banerjee R, Chowdhury S. Functional genomics of lung cancer progression reveals mechanism of metastasis suppressor function. Mol Cytogenet. 2014;7(Suppl 1 Proceedings of the International Conference on Human):I9. DOI: 10.1186/1755-8166-7-S1-I9.

[23] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–25. DOI: 10.1038/nature11404.

[24] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98(24):13790–5.

[25] DePianto DJ, Chandriani S, Abbas AR, Jia G, N'Diaye EN, Caplazi P, Kauder SE, Biswas S, Karnik SK, Ha C, Modrusan Z, Matthay MA, Kukreja J, Collard HR, Egen JG, Wolters PJ, Arron JR. Heterogeneous gene expression signatures correspond to distinct lung pathologies and biomarkers of disease severity in idiopathic pulmonary fibrosis. Thorax. 2015;70(1):48–56. DOI: 10.1136/thoraxjnl-2013-204596.

[26] Gerber DE, Oxnard GR, Govindan R. ALCHEMIST: bringing genomic discovery and targeted therapies to early-stage lung cancer. Clin Pharmacol Ther. 2015;97(5):447–50. DOI: 10.1002/cpt.91.

[27] Slonim DK, Yanai I. Getting started in gene expression microarray analysis. PLoS Comput Biol. 2009;5(10):e1000543. DOI: 10.1371/journal.pcbi.1000543.

[28] Roberts PC. Gene expression microarray data analysis demystified. Biotechnol Annu Rev. 2008;14:29–61. DOI: 10.1016/S1387-2656(08)00002-1.

[29] Vikman P, Fadista J, Oskolkov N. RNA sequencing: current and prospective uses in metabolic research. J Mol Endocrinol. 2014;53(2):R93–101. DOI: 10.1530/JME-14-0170.

[30] Ye N, Yin H, Liu J, Dai X, Yin T. GESearch: an interactive GUI tool for identifying gene expression signature. Biomed Res Int. 2015;2015:853734. DOI: 10.1155/2015/853734.

[31] Taylor A, Steinberg J, Andrews TS, Webber C. GeneNet Toolbox for MATLAB: a flexible platform for the analysis of gene connectivity in biological networks. Bioinformatics. 2015;31(3):442–4. DOI: 10.1093/bioinformatics/btu669.

[32] Ibrahim M, Jassim S, Cawthorne MA, Langlands K. A MATLAB tool for pathway enrichment using a topology-based pathway regulation score. BMC Bioinformatics. 2014;15:358. DOI: 10.1186/s12859-014-0358-2.

[33] Hung JH. Gene set/pathway enrichment analysis. Methods Mol Biol. 2013;939:201–13. DOI: 10.1007/978-1-62703-107-3_13.

[34] Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43(Database issue):D1049–56. DOI: 10.1093/nar/gku1179.

[35] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.

[36] Yuryev A. Introduction to pathway analysis. In: Yuryev A, editor. Pathway Analysis for Drug Discovery: Computational Infrastructure and Applications. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2008. DOI: 10.1002/9780470399279.ch1.

[37] Arakelyan A, Nersisyan L. KEGGParser: parsing and editing KEGG pathway maps in MATLAB. Bioinformatics. 2013;29(4):518–9. DOI: 10.1093/bioinformatics/bts730.

[38] Zeng W, Sun L, Jiang X, Chen X, Hou F, Adhikari A, Xu M, Chen ZJ. Reconstitution of the RIG-I pathway reveals a signaling role of unanchored polyubiquitin chains in innate immunity. Cell. 2010;141(2):315–30. DOI: 10.1016/j.cell.2010.03.029.

[39] Kaneda Y. The RIG-I/MAVS signaling pathway in cancer cell-selective apoptosis. Oncoimmunology. 2013;2(4):e23566.

[40] Kato H, Fujita T. RIG-I-like receptors and autoimmune diseases. Curr Opin Immunol. 2015;37:40–5. DOI: 10.1016/j.coi.2015.10.002.

[41] ArakelyanA. MATLAB chapter supplement: PSF scripts and gene expression data [Internet]. 2016 [updated: 16.01.16]. Available from: https://www.dropbox.com/sh/ni26ga5t0nn4kzy/AADY5XDcheAV2jBqY5XAh1Jua [Accessed: 16.01.16].

[42] Arakelyan A, Aslanyan L, Boyajyan A. High-throughput gene expression analysis concepts and applications. In: Sequence and Genome Analysis II – Bacteria, Viruses and Metabolic Pathways. Hong Kong: iConcept Press; 2013.

[43] Nersisyan L, Johnson G, Riel-Mehan M, Pico A, Arakelyan A. PSFC: a pathway signal flow calculator app for cytoscape. F1000Research. 2015;4:480. DOI: 10.12688/f1000research.6706.1.

[44] Valeyre D, Bernaudin JF, Jeny F, Duchemann B, Freynet O, Planès C, Kambouchner M, Nunes H. Pulmonary sarcoidosis. Clin Chest Med. 2015;36(4):631–41. DOI: 10.1016/j.ccm.2015.08.006.

[45] Kriegova E, Fillerova R, Tomankova T, Hutyrova B, Mrazek F, Tichy T, Kolek V, du Bois RM, Petrek M. T-helper cell type-1 transcription factor T-bet is upregulated in pulmonary sarcoidosis. Eur Respir J. 2011;38(5):1136–44. DOI: 10.1183/09031936.00089910.

[46] Arakelyan A, Kriegova E, Kubistova Z, Mrazek F, Kverka M, du Bois RM, Kolek V, Petrek M. Protein levels of CC chemokine ligand (CCL)15, CCL16 and macrophage stimulating protein in patients with sarcoidosis. Clin Exp Immunol. 2009;155(3):457–65. DOI: 10.1111/j.1365-2249.2008.03832.x.

[47] Pabst S, Bradler O, Gillissen A, Nickenig G, Skowasch D, Grohe C. Toll-like receptor-9 polymorphisms in sarcoidosis and chronic obstructive pulmonary disease. Adv Exp Med Biol. 2013;756:239–45. DOI: 10.1007/978-94-007-4549-0_30.

[48] Crouser ED, Culver DA, Knox KS, Julian MW, Shao G, Abraham S, Liyanarachchi S, Macre JE, Wewers MD, Gavrilin MA, Ross P, Abbas A, Eng C. Gene expression profiling identifies MMP-12 and ADAMDEC1 as potential pathogenic mediators of pulmonary sarcoidosis. Am J Respir Crit Care Med. 2009;179(10):929–38. DOI: 10.1164/rccm.200803-490OC.

[49] Lockstone HE, Sanderson S, Kulakova N, Baban D, Leonard A, Kok WL, McGowan S, McMichael AJ, Ho LP. Gene set analysis of lung samples provides insight into pathogenesis of progressive, fibrotic pulmonary sarcoidosis. Am J Respir Crit Care Med. 2010;181(12):1367–75. DOI: 10.1164/rccm.200912-1855OC.

[50] Nersisyan L, Samsonyan R, Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. F1000Res. 2014;3:145. DOI: 10.12688/f1000research.4410.2.