

## Gene expression

## KEGGanim: pathway animations for high-throughput data

Priit Adler<sup>1,†</sup>, Jüri Reimand<sup>2,†</sup>, Jürgen Jänes<sup>2</sup>, Raivo Kolde<sup>3</sup>, Hedi Peterson<sup>1,3</sup> and Jaak Vilo<sup>1,2,3,\*</sup><sup>1</sup>Estonian Biocentre, Riia 23b, <sup>2</sup>University of Tartu, Institute of Computer Science, Liivi 2 and <sup>3</sup>QureTec Inc. Ülikooli 6a, Tartu, Estonia

Received on August 5, 2007; revised on October 22, 2007; accepted on November 19, 2007

Advance Access publication December 1, 2007

Associate Editor: Olga Troyanskaya

## ABSTRACT

**Motivation:** Gene expression analysis with microarrays has become one of the most widely used high-throughput methods for gathering genome-wide functional data. Emerging *-omics* fields such as proteomics and interactomics introduce new information sources. With the rise of systems biology, researchers need to concentrate on entire complex pathways that guide individual genes and related processes. Bioinformatics methods are needed to link the existing knowledge about pathways with the growing amounts of experimental data.

**Results:** We present KEGGanim, a novel web-based tool for visualizing experimental data in biological pathways. KEGGanim produces animations and images of KEGG pathways using public or user uploaded high-throughput data. Pathway members are coloured according to experimental measurements, and animated over experimental conditions. KEGGanim visualization highlights dynamic changes over conditions and allows the user to observe important modules and key genes that influence the pathway. The simple user interface of KEGGanim provides options for filtering genes and experimental conditions. KEGGanim may be used with public or private data for 14 organisms with a large collection of public microarray data readily available. Most common gene and protein identifiers and microarray probesets are accepted for visualization input.

**Availability:** <http://biit.cs.ut.ee/KEGGanim/>

**Contact:** vilo@ut.ee

## 1 INTRODUCTION

High-throughput methods such as microarrays have changed the research pace in molecular biology. Thousands of genes and proteins are now routinely studied under experimental conditions (Buck *et al.*, 2004; Eads *et al.*, 2000; Schena *et al.*, 1995), with results stored in public microarray databases like GEO (Barrett *et al.*, 2007) and ArrayExpress (Parkinson *et al.*, 2007), and protein databases like Pride (Jones *et al.*, 2006).

As genomic and proteomic data accumulate, researchers envisage complex systems behind biological processes and

functions. Genes and proteins rarely operate alone in the cell, but are regulated by elaborate mechanisms and bound into networks (Alon, 2007). Systems biology approaches are applied to view these networks in detail. Well-studied parts of networks called pathways have roles in cell signaling, gene regulation and metabolism as well as human disease. Pathways are described in databases like KEGG (Kanehisa *et al.*, 2006) and Reactome (Vastrik *et al.*, 2007).

Knowledge of pathways stored in databases is often far from complete. Bioinformatics methods are needed that combine various experimental data to verify existing knowledge and propose new hypotheses. Visualization has a key role in understanding complex and dynamical phenomena of pathways, proteomics and gene expression. Several efforts have been made in this area, but there is still a need for interactive web-based pathway resources. For example, KEGG allows the user to colour genes on the pathway. Reactome Skypainter and PathwayExpress (Khatri *et al.*, 2007) link genes to pathways using overrepresentation analysis. However, these tools have no means for directly incorporating experimental data. Also, fixed images fail to deliver the temporal and spatial dynamics behind pathways and gene expression. GenMAPP (Dahlquist *et al.*, 2002) and BioCyc Pathway Tools (Paley *et al.*, 2006) produce user-defined pathways and cellular wiring diagrams, and allow inclusion of expression data with some animation capabilities. No such systematic visualization functions are available for the comprehensive KEGG resource.

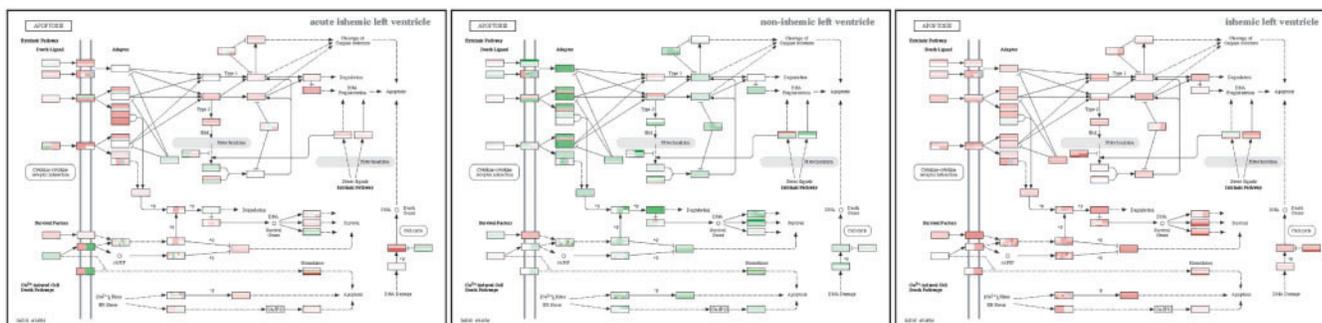
## 2 APPROACH

KEGGanim is a novel web-based visualization tool that links manually curated pathway maps from KEGG with experimental data from sources like gene expression and proteomics. KEGGanim shows animated figures of pathways with genes and proteins depicted as coloured rectangles. Pathway members are painted red or green according to their experimental values in the given dataset. Animation changes the colour values of these rectangles while looping over experimental conditions in the dataset, for instance moments in a timeseries (Spellman *et al.*, 1998), healthy and diseased samples (Alon *et al.*, 1999), or samples of healthy tissues (Ge *et al.*, 2005).

KEGGanim allows a researcher to observe expression and protein production dynamics in the context of pathway

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



**Fig. 1.** KEGG animated pathways example: cinefilm snapshots display expression of genes involved in the apoptosis pathway during myocardial remodelling. The snapshots show pathway behaviour in three samples of left ventricle; non-ischaemic (middle), ischaemic (right) and acute ischaemic (left) conditions are observed. Ischaemic tissue, a common reason for heart disease, causes decrease in blood supply that leads to apoptosis and eventually to tissue death.

dependencies. Animating a pathway over consecutive timepoints reflects the behaviour of master regulatory genes, propagation of signals in the pathway over time, and the avalanche of up- and downregulation caused by the master regulator. When analysing a set of conditions or tissues on microarray, KEGGanim allows to reason about tissue specificity or the influence of global conditions on the pathway and its components. Figure 1 shows an example of KEGGanim output.

### 3 METHODS

KEGGanim combines KEGG pathway data with a matrix of experimental values of genes and proteins. First, the user needs to select a pathway of interest from a dropdown menu in the web interface, which corresponds to a graphical map downloaded from the KEGG database.

The second input is a matrix containing experimental values for genes and proteins. A number of gene expression datasets from GEO and ArrayExpress are available in KEGGanim for immediate analysis. KEGGanim automatically fetches all associations to the genes in the pathway from the g:Profiler software (Reimand *et al.*, 2007), and creates an animation of the related experimental values over different conditions. If several probesets or proteins match a pathway member, the corresponding node is split into smaller coloured areas to reflect different experimental values. Users can upload their own data for analysis and visualization. Most common gene and protein IDs and microarray probesets are accepted as input, for instance standard names, RefSeq, Entrez, Affymetrix, UniProt, EnsEMBL as well as species-specific IDs. Uploaded data is optionally centred and normalized, and missing values may be replaced with fixed values or via the kNN method (Troyanskaya *et al.*, 2001) implementation in GEPAS (Montaner *et al.*, 2006).

Additional options help to interpret the animations and concentrate on specific conditions or components. KEGGanim tooltips display names and descriptions of genes and proteins when the user hovers over corresponding pathway members. Lineplots display the amounts of proteins or the gene expression levels. The user can narrow down the study by selecting a subset of conditions to view. Experimental values for irrelevant pathway members and related probesets may also be excluded from the animation. The *cinefilm* feature overcomes the technical difficulty of including animations in printed materials by allowing the user to extract pathway snapshots of timepoints or conditions into a separate image (Fig. 1). These features are especially useful in visualizing timeseries data. The GIF animations

produced by KEGGanim do not require special software packages for viewing, and can easily be inserted into presentation slides, web pages, tutorials, etc.

### 4 DISCUSSION

Advancing high-throughput technologies allow researchers to gather information about organizational, functional and physical layers of the cell. There is an increasing need for ideas that successfully integrate layers of data and explain the elaborate mechanisms responsible for creating the observed measurements. With the development of KEGGanim, we wish to contribute to the data integration goal and provide methods that take advantage of the powerful human visual analysis skill.

### 5 CONCLUSION

KEGGanim is a simple web-based visualization tool that links manually curated KEGG pathway maps with high-throughput data. The tool creates animations that allow intuitive visual analysis of condition or tissue-specific changes in gene expression or protein levels within the selected pathway. KEGGanim is already actively used in several research initiatives, e.g. in functional profiling of mouse embryonic stem cell development.

### ACKNOWLEDGEMENTS

This research has been supported by the EU FP6 grants ENFIN LSHG-CT-2005-518254, FunGenES LSHG-CT-2003-503494 and Estonian Science Foundation ETF5724. The authors would like to thank Dr N. Billon, M. Kull, J. Hansen and the reviewers of this manuscript.

*Conflict of Interest:* none declared.

### REFERENCES

- Alon, U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Gen.*, **8**, 450–61.
- Alon, U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–50.
- Barrett, T. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.*, **D35**, D760–D765.

- Buck,M.J. *et al.* (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.
- Dahlquist,K.D. *et al.* (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, **31**, 19–20.
- Eads,C.A. *et al.* (2000) MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res.*, **28**, E32.
- Ge,X. *et al.* (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86**, 127–141.
- Jones,P. *et al.* (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.*, **D34**, D659–D663.
- Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **D34**, D354–D357.
- Khatri,P. *et al.* (2007) Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res.*, **W35**, W206–W211.
- Montaner,D. *et al.* (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, **W34**, W486–W491.
- Paley,S. *et al.* (2006) The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Res.*, **34**, 3771–3778.
- Parkinson,H. *et al.* (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **D35**, D747–D750.
- Reimand,J. *et al.* (2007) g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **W35**, W193–W200.
- Schena,M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
- Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
- Vastrik,I. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.