BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2283–2284
doi:10.1093/bioinformatics/bts431

*Systems biology*

Advance Access publication July 5, 2012

# A sequence comparison and gene expression data integration add-on for the Pathway Tools software

Peter M. Krempl[1,2], Juergen Mairhofer[1,3], Gerald Striedner[1,3] and Gerhard G. Thallinger[1,2,*]

[1]ACIB, Petersgasse 14, 8010 Graz, [2]Institute for Genomics and Bioinformatics, Graz University of Technology, Petersgasse 14, 8010 Graz and [3]Institute of Applied Microbiology, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria

Associate Editor: Martin Bishop

## ABSTRACT

**Summary:** We present a plug-in for Pathway Tools, an integrated systems biology software to create, maintain and query Pathway/Genome Databases. Fully integrated into the graphical user interface and menu, this plug-in extends the application's functionality by the ability to create multiple sequence alignments, systematically annotate insertion sequence (IS) elements and analyse their activity by cross-species comparison tools. Microarray probes can be automatically mapped to target genes, and expression data obtained with these arrays can be transformed into input formats needed to visualize them in the various omics viewers of Pathway Tools. The plug-in API itself allows developers to integrate their own functions into the Pathway Tools menu.

**Availability:** Binaries are freely available for non-commercial users at http://genome.tugraz.at/PGDBToolbox/ and can be used on all platforms supported by Pathway Tools. A user guide is freely available at: http://genome.tugraz.at/PGDBToolbox/documentation.shtml.

**Contact:** ptools@acib.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 23, 2012; revised and accepted on July 3, 2012

## 1. INTRODUCTION

Pathway/Genome Databases (PGDBs) are highly integrated model organism databases that store genome sequence and annotation as well as a reconstruction of the metabolic and regulatory network. Well-curated PGDBs like EcoCyc (Keseler *et al.*, 2011) are an important tool to gain better understanding of cellular component interaction and how changes in an organism's network affect its phenotype. PGDBs are usually maintained, queried and visualized using the Pathway Tools software developed by the Bioinformatics Research Group at SRI International (Karp *et al.*, 2010), which has extensive capabilities in the fields of operon and pathway prediction, analysis of biological networks, species comparison and analysis of omics datasets. The importance of PGDBs to biochemical research is reflected by an ever growing number of PGDB collections and software tools maintaining and analysing such databases (Caspi *et al.*, 2010; Le Fèvre *et al.*, 2007; Vellozo *et al.*, 2011; Vieira *et al.*, 2011).

Our toolbox aims at further extending the capabilities of PGDB-based systems biology research by means of plug-ins that can be added to the Pathway Tools software to address special aspects of analytical workflows not covered by the core software itself. Multiple sequence alignments of orthologous genes across related organisms, including a freely defined 5′ sequence region, allow detecting differences in gene, promoter or regulator binding regions that affect gene function or expression. Systematic annotation of insertion sequence (IS) elements, combined with cross-species comparison functions, provides insight into transposon activity and gene clusters affected thereof. Tools for mapping microarray probes to target genomes and visualizing expression profiles in the genome browser are provided; these tools also create input files for data visualization in Pathway Tools' various overview diagrams.

## 2. IMPLEMENTATION

Implemented in LISP, our toolbox can be loaded directly into the Pathway Tools application at startup. An individual LISP package separates toolbox functions from Pathway Tools internal functions. Toolbox components are grouped into modules that address a specific task and can be used widely independent from each other; a core module provides common functions and services. All toolbox functions enumerated in the following sections are fully integrated into the Pathway Tools graphical user interface (GUI) and accessible by a plug-in menu placed in the application's main menu.

### 2.1 Menu plug-in API

Although Pathway Tools provides an extensive LISP API that allows integration of high-level script or console-based database query and modification routines, it currently does not provide a plug-in API for its GUI. As a consequence, dynamic integration of user interface commands and menus at runtime requires considerable programming effort and in-depth knowledge of the internal program structure.

As part of this toolbox, we introduce an easy-to-use plug-in API that allows integration of additional submenus and functions to the application's menu by a single function call. Besides its use within the toolbox modules, this API can also be used by other developers to add their specific functions to the plug-in menu.

### 2.2 Multiple sequence alignment of orthologs

Pathway Tools offers a multi-genome browser that allows aligning genome annotations based on links between orthologous genes and is very useful for detection of large-scale differences

*To whom correspondence should be addressed.

between organisms. Analysis of point mutations—especially in the 5′ region of genes, where they might account for significant changes in expression levels (Yu *et al*., 2009)—requires repeated gene browsing, sequence export into FASTA files and alignment using external tools, as this system lacks a built-in sequence alignment feature. Therefore, we introduced an expanded CLUSTAL W alignment (Larkin *et al*., 2007) to compare orthologs of a selected gene at DNA sequence level. Organisms that should be included in the alignment are selected in the same way as for display in the multi-genome browser. A user-defined number of base pairs in 5′ direction of the gene can be included into the alignment. The initial CLUSTAL alignment is processed to include a header describing all aligned genes; additional 5′ sequences are highlighted by capitalization, and accurate position information relative to the start of the gene is provided. The final result is presented in a dialog window from where it can be saved as text file.

### 2.3 IS element annotation and analysis

IS elements are a major cause of genomic modifications that may change a cell's phenotype dramatically even during relatively short-termed experiments under non-mutagenic conditions, as recently demonstrated (Gaffé *et al*., 2011). They are short mobile genetic elements encoding only functions involved in their mobility. A dedicated database, IS Finder (Siguier *et al*., 2006), gives detailed information about individual ISs and their taxonomy. To analyse IS abundance and activity, we integrated a systematic IS annotation as well as suitable cross-species comparison tools.

The IS taxonomy tree of families, groups and types is placed within the 'Paralogous Gene Groups' class of each PGDB. A menu link to the root of this tree, combined with Pathway Tools' generic class tree navigation, allows easy browsing of the IS taxonomy. As IS types behave like ordinary paralogous gene groups, they profit from all display and browsing options provided for such groups, e.g. clickable links from IS types to specific transposase genes and vice versa or location maps of all genes of a certain type.

IS element annotation is facilitated by a system of dialog windows that provides selection of defined IS types as well as easy creation of new taxonomy entries. When IS element annotation is complete in all strains of interest, and orthologous genes are linked to each other across all PGDBs, a species comparison function creates tables of all IS element loci in each organism, including information about adjacent genes; for each other organism, the orthologous IS element locus (if present) and the percentage of sequence identity in the 5000 bp up- and downstream region of the IS element locus are reported. A summary table lists all groups of orthologous IS element loci and thereby allows rapid detection of shared as well as singleton loci.

### 2.3 Mapping of microarray probes and gene expression data

Pathway Tools' omics and pathway viewers offer many ways to visualize gene expression in genomic, metabolomic and regulatory context. Efficient use of these features implies accurate assignment of probes to PGDB genes, but using gene names provided in array design files often leads to unsatisfactory results (see Supplementary Material). We therefore introduce tools to map probes to PGDB genomes by probe sequences and facilitate preprocessing of gene expression data before visualization.

Microarray probes can be imported from tab-delimited text files (e.g. GAL or ADF files) using a generic parser that prompts the user to select proper probe ID and sequence columns. Probes are mapped to target genomes/PGDBs by BLAST (Altschul *et al*., 1997), followed by extraction of strong hits. Mapping results are stored in the PGDBs directory in GFF2 format, which allows displaying probe target annotation as additional track in Pathway Tools' genome browser. A tabular text file is created, containing probe IDs, their target gene IDs and names as well as the percent alignment identity, which facilitates the identification of imperfect or cross-talking matches.

To map expression data to the genome, tab-delimited data files are merged with probe-target-mappings mentioned above. A configuration file defines how expression data are organized into different tracks; time series, dye swap, clusters and optional selection/highlighting of targets on either strand are supported. Results are stored in two formats: GFF2 to be visualized in the genome browser, and tab-delimited text files that link expression levels to PGDB-internal gene IDs for visualization in Pathway Tools' cellular and pathway overview diagrams.

## ACKNOWLEDGEMENTS

We thank Peter D. Karp and the Bioinformatics Research Group at SRI International for providing the source code of the Pathway Tools software, Karoline Marisch and Theresa Scharl from the Institute of Applied Microbiology at the University of Natural Resources and Life Sciences in Vienna for testing the toolbox functions and valuable discussion, and Daniel Friedl and Samuel Schimpel for setup of the toolbox installers.

*Funding*: This work has been supported by the Austrian BMWFJ, BMVIT, SFG, Standortagentur Tirol and ZIT through the Austrian FFG-COMET Funding Program [FFG Grant 824186], and the Austrian BM.W_F through the GEN-AU project Bioinformatics Integration Network [FFG Grant 820962].

*Conflict of Interest*: none declared.

## REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Caspi,R. *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*., **38**, D473–D479.

Gaffé,J. *et al.* (2011) Insertion sequence-driven evolution of Escherichia coli in chemostats. *J. Mol. Evol*., **72**, 398–412.

Karp,P.D. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform*., **11**, 40–79.

Keseler,I.M. *et al.* (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res*., **39**, D583–D590.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Le Fèvre,F. *et al.* (2007) Cyclone: java-based querying and computing with Pathway/Genome databases. *Bioinformatics*, **23**, 1299–1300.

Siguier,P. *et al.* (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res*., **34**, D32–D36.

Vellozo,A.F. *et al.* (2011) CycADS: an annotation database system to ease the development and update of BioCyc databases. *Database*, **2011**, bar008.

Vieira,G. *et al.* (2011) Core and panmetabolism in Escherichia coli. *J. Bacteriol*., **193**, 1461–1472.

Yu,W. *et al.* (2009) AmpC promoter and attenuator mutations affect function of three Escherichia coli strains. *Curr. Microbiol*., **59**, 244–247.