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

Peter M. Kreempl
Juergen Mairhofer, Gerald Striedner, Gerhard G. Thallinger

Conference on Predicting Cell Metabolism and Phenotype
Menlo Park, March 4-6, 2013
Toolbox Features

- Multiple sequence alignment of orthologs
- Mapping of gene expression data
- Annotation and cross-species comparison of IS Elements
- Plug-In API

- Seamless integration into Pathway Tools user interface
- Easy to install: Automated installer
Expanded CLUSTAL W alignment of orthologous genes
  - Complements Pathway Tools’ sequence viewers and Multiple Genome Browser
User defined additional 5’ region to compare
  - Promoters
  - Regulator binding sites
  - ...
Highlighting of
  - 5’ region – lower case letters
  - Coding region – upper case letters
Additional legend and position information
• Select gene
  • Gene or protein page
  • Ortholog links must be present

• Select organisms to compare
  • Like for Pathway Tools Multiple Genome Browser
  • …or using organism selection dialog of the toolbox

• Select from menu:
  ➢ ACI B-Tool box
  ➢ Species Comparison
  ➢ Sequence alignment of orthologous genes

• Enter length of 5’ region
Multiple Sequence Alignment

Example Results

CLUSTALW alignment of genes

EG10998 = thrA in Escherichia coli K-12 substr. MG1655 [ECOLI]
GCQ2-181 = thrA in Escherichia coli K-12 substr. DH10B [ECOL316385]
GD40-362609 = thrA in Escherichia coli 0157:H7 str. Sakai [ECOL336585]

Additional 5' region (180 bp) is shown in lowercase letters.

CLUSTAL 2.1 multiple sequence alignment

EG10998
- agcagataaaaaattacagatcaacaccatacctcgaagcgttacagcacaccatattac -180..-122
GCQ2-181
- agcagataaaaaattacagatcaacaccatacctcgaagcgttacagcacaccatattac -180..-122
GD40-362609
agcagataaaaaattacagatcaacaccatacctcgaagcgttacagcacaccatattac -180..-121
** * * ** * * * * * * * * **

EG10998
cacaccatccacaccatcaccacaggaacgtgcggctcgacgcgtacaggaacacagaa -121..-62
GCQ2-181
cacaccatccacaccatcaccacaggaacgtgcggctcgacgcgtacaggaacacagaa -121..-62
GD40-362609
cacaccatccacaccatcaccacaggaacgtgcggctcgacgcgtacaggaacacagaa -120..-61
** * * ********* **********

EG10998
aaaagccggaactgcagtcggtgcgttttttttttctgacccaaggttaacgagtaacaa -61..-2
GCQ2-181
aaaagccggaactgcagtcggtgcgttttttttttctgacccaaggttaacgagtaacaa -61..-2
GD40-362609
aaaagccggaactgcagtcggtgcgtttttttttctgacccaaggttaacgagtaacaa -60..-2
**********************************

EG10998
cATCGCAGTTGGAGAGTTCCGCCGCTACACTGCGCAAAATGCAAGAAGCTTTCTGCTGT -1..59
GCQ2-181
cATCGCAGTTGGAGAGTTCCGCCGCTACACTGCGCAAAATGCAAGAAGCTTTCTGCTGT -1..59

[Image]
Gene Expression Data Integration
Feature Description

• Import of probe libraries or array designs from different file formats:
  • Tab-delimited text (with header row)
  • MAGE-TAB Array Design Format (ADF)
  • GenePix Array List (GAL)
• Sequence-based mapping of microarray probes to target PGDB genomes
• Conversion of expression data into input for
  • GFF tracks for genome browser
  • Overlay of gene expression data in
    • Cellular overview
    • Genome overview
    • Regulatory overview
Gene Expression Data Integration
Why Sequence-Based Probe Mapping?

• Name matching depends on:
  • Complete and accurate target gene annotation of probe library/array design
  • Extensive annotation of gene names, synonyms and accessions in target PGDBs

• Common errors using name matching:
  • Incomplete annotation
    ➔ low mapping efficiency
  • Annotation of probes and PGDB not compatible
    ➔ low mapping efficiency
  • Ambiguous or misleading synonyms in either annotation
    ➔ false positive assignments
Sequence based probe mapping

- Circumvents drawbacks of name matching
- Provides unambiguous matching of probes to their exact target genes
  - No false-positive matches due to ambiguous gene synonyms
- Allows detection of cross-talking probe matches
- Provides stable mapping efficiency regardless of annotation of probes and target PGDBs
Gene Expression Data Integration
Example: GFF Tracks in Genome Browser

*Escherichia coli* K-12 substr. MG1655 Chromosome: *gadE*

Experimental data: Reference Design time-course, *E.coli* (Takahashi et al., 2011); ArrayExpress Accession E-GEOD-6033
Gene Expression Data Integration
Example: Visualization in Cellular Overview

Experimental data: Reference Design time-course, *E.coli* (Takahashi et al., 2011); ArrayExpress Accession E-GEOD-6033
Insertion Sequence (IS) Elements are:
• short mobile genetic elements
• a major cause of genomic modifications

ACIB PGDB Toolbox provides:
• Systematic annotation and browsing of the IS Element taxonomy
  • IS Element types as Paralogous Gene Groups
• Species comparison features
  • Check for correct ortholog links between strains
  • Detection of shared and singleton IS Element loci
Annotation of IS Elements

Annotation

...annotate genes

create new IS Element families, groups and types...
Annotation of IS Elements
Browsing: Root of Taxonomy Tree

Parent Classes: Paralogous-Gene-Groups

Child Classes:
- IS Element family: IS1 (1),
- IS Element family: IS3 (3),
- IS Element family: IS4 (2),
- IS Element family: IS5 (1),
- IS Element family: IS30 (1)
Annotation of IS Elements
Browsing: IS Element Type Overview Page

*Escherichia coli K-12 substr. MG1655 Group: IS186B

Superclasses: Insertion Sequence Elements -> IS Element family: IS4 -> IS Element group: IS231

Group Members: insL-1 (IS186/IS421 transposase),
insL-2 (IS186/IS421 transposase),
insL-3 (predicted IS186/IS421 transposase)

Locations of Mapped Genes:
• Export of IS Element annotation in GFF v2 format
• Cross-species comparison
  • Selection of organisms: like for multiple sequence alignment feature
  • Tabular output of orthologous and singleton IS Element loci
• For each strain: Tabular output of
  • IS Element genes
  • Adjacent genes
  • % matching identity of +/- 5000 bp region in other strains
    • Detection of missing ortholog links
  • Orthologous IS Element genes in other strains
Plug-In API
Feature Description

- Easy-to-use API to integrate user-defined tools into the Pathway Tools graphical user interface
- Direct integration of own commands into the application’s menu
- Support for creation of GUI dialog windows
- Useful for
  - Pathway Tools users - to integrate their own queries and tools
  - Third-party developers - to provide user-friendly access to their Pathway Tools enhancements
- Detailed documentation in toolbox user guide
• Write your analysis function:
  (defun my-function ...)

• Wrap your function in a CLIM command:
  (clim:define-command com-my-command-name () (my-function))

• Use the plug-in API to place it in the menu:
  (acib::register-plugin-menu-command
   "Menu Command Name" 'com-my-command-name
   :submenu '("First Level Submenu" "Second Level Submenu"))

• That’s all!
Plug-In API
Example: Menu integration
Toolbox Availability

- [http://genome.tugraz.at/PGDBToolbox](http://genome.tugraz.at/PGDBToolbox)
- Contact: ptools@acib.at

- Free of charge for academic/non-commercial use
  - Registration required

- User guide (freely available):
  [http://genome.tugraz.at/PGDBToolbox/documentation.shtml](http://genome.tugraz.at/PGDBToolbox/documentation.shtml)

- Reference – please cite:
Acknowledgements

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 - for valuable discussion and testing of toolbox features
Funded by:

Austrian Research Promotion Agency

Federal Ministry of traffic, innovation and technology

Federal Ministry of economy, family and youth

Styrian Promotion Agency

Government of Styria

Tyrolean Location Agency for Business and Science

Technology promotion agency of the City of Vienna
A Company of

Graz University of Technology

University of Natural Resources and Life Sciences Vienna

University of Graz

University of Innsbruck

Joanneum Research Graz
Contact:

Prof. Dr. Anton Glieder
CEO & CSO
anton.glieder@acib.at

Dr. Mathias Drexler
CEO & CFO
mathias.drexler@acib.at