The PathOlogist: Pathway Metrics Calculator and Visualization Tool

---

Features:

- Novel probe-level normalization
    - calculates probability of probes in an array being in the ‘up’ or ‘down’ state for each sample using sample-wide distribution of values
- Quantitative pathway-level metrics
    - calculates activity and consistency scores for any/all pathway(s) in a database of canonical pathways using expression data from any number of samples
- Quantitative interaction-level metrics
    - calculates metrics and probabilities for the molecules and interactions in a single pathway for multiple samples
- *Informative graphics*
    - *generates bi-dimensionally clustered heatmaps of pathway metrics, and visual representation of single pathways on a sample-by-sample basis, showing molecule probabilities, activity, and consistency and/or copy number alterations for all molecules and interactions (version 1 only)* also generates text version of above graphic
- Integrated statistical analyses
    - calculates p-values for each pathway in tests of sample classification, correlation with a continuous value, influence on sample survival, or as a target for gene alterations (eg. copy number abnormalities or mutations).

Use:

1. To run the PathOlogist from the MATLAB command prompt:
    - Set the current directory to the directory containing the code files.
    - Type `PathOlogist` at the command prompt and press enter. The tool should pop up automatically.

2. RMA Normalization

- The PathOlogist's analytic pipeline begins with RMA-normalization of probe-set intensities from a set of .cel files. To perform the normalization:

    i. Locate the folder containing the .cel files to be analyzed.
    ii. Locate the library file (.cdf file) corresponding to the platform used to generate the .cel files. If the files do not match, you will receive an error message.

- After clicking 'Calculate RMA' the normalization process will begin. This may take up to 30 minutes, depending on the number of files and the speed of the computer. Once calculation has been completed, you will be prompted for a location to save the data as a text file. This data will also be stored in the tool until cleared (see 3. Load Data) or the tool is closed, and may be reloaded in future sessions.

3. Load Data

- As an alternative to calculating RMA-normalized probe-set intensities from .cel files, you can load your own text file of intensity readings, generated, for example, during a previous PathOlogist session or using a different normalization technique. At this screen you can also load UDP, Activity, Consistency, and Gene Hits files (see below) from previous sessions for further analysis. All input data must be in tab-delimited text file format. In general, each file should contain a matrix of values, with one column for each sample and one row for each probe/pathway/gene, along with a single line each of row and column headers. The PathOlogist is capable of integrating 5 different types of data. These are:

    i. **RMA** (robust multichip average)
    Normalized expression data. The PathOlogist can calculate RMA data from a set of .cel files and a chip-specific channel(.cdf) file.

    ii. **UDP** (up-down probabilities)
    For every probeset on a microarray chip, the probability that it is in an 'up' state (highly expressed) is calculated for each sample, based on the sample-wide distribution of normalized expression values.

    iii. **Gene Hits**: Binary account of gene status. This can represent copy number, methylation, mutation, etc., and is used in pathway visualization displays and identifying 'enriched' pathways (ie. pathways that are 'hit' more often than would be expected at random). Data should specify 0 for 'unhit' and 1 for 'hit'.

iv. **Activity**
Pathway-based metric representing how likely interactions within the pathway are to occur.

v. **Consistency**
Pathway-based metric representing overall pathway logic by comparing expected v. actual expression of interaction components

Input File Formats

	Data	Row Labels	Column Labels
RMA	Normalized expression	Probe names	Sample names
UDP	Probabilities (range 0-1)	Probe names	Sample names
Gene Hits	Unhit/hit (0/1)	Entrez Gene IDs	Sample names
Activity	Pathway score (range 0-1)	Pathway names	Sample names
Consistency	Pathway score (range 0-1)	Pathway names	Sample names

- The order of samples for data types highlighted in the same color in the above table must be identical.

- Click the button for the file type to be loaded, and locate the file in the directory browser. The selected file name should appear in the textbox. Press '**Load Data'** to read in the file. Once the data has been loaded, the associated button will turn yellow, and the file location will appear. If there is an error, an error message will appear.

- If UDP data has been loaded, be sure to specify the microarray chip used to generate the expression data. Currently, the PathOlogist supports data from Affymetrix Human Genome U133A, U133B, U133_plus2, U133AAofAv2, U95 chips, and Mouse Genome 430A, 430B, and 430plus2 chips.

- If your platform is not listed, choose 'add new platform' from the dropdown list. You will be prompted for a label for the new platform and a probe linking file. This file should be a textfile with one column of probeset names and a corresponding column of Entrez Gene IDs.

4. Calculate Probabilities

- Once RMA data has been loaded, UDP (up-down probability) data can be calculated. This process involves fitting each sample-wide set of probe intensities to a model representing 'up' and 'down' distributions and assigning each individual intensity value a probability of being in the 'up' state. The calculations may take approximately an hour or more, depending on the number of samples.

- Before pressing 'Calculate', you have the option of assigning each sample a more informative label, to be used as column headers in the results file.

- If you have not already done so, select the platform used to generate the expression data here.

- Once the probabilities have been calculated, you will be asked to specify a location to save the data as a text file, and it will be automatically loaded into the tool's Load Data page.

5. Choose a Pathway

    - Here you can select pathways for inclusion in calculation of activity and consistency metrics or pathway visualization. The PathOlogist is designed to provide analysis for any pathway in the Pathway Interaction Database (http://pid.nci.nih.gov). Currently, this database comprises over 500 canonical molecular pathways. Pathways are listed in alphabetical order, followed by its reference source (NCI/Nature, Kegg, or BioCarta)

    - Multiple pathways can be chosen by holding down the Shift or Ctrl button (for metrics calculations only).

    - Alternatively, all the pathways in the database can be analyzed in turn by checking the '**all pathways'** option.

    - Note that for pathway visualization (See *8.Display a Single Pathway),* only one pathway may be selected at a time.

6. Choose Samples

    - Once UDP data has been loaded, the sample names from the file should appear in the list box.

    - Choose samples one at a time, choose multiple samples by holding down the Shift or Ctrl button, or select '**all samples'**.

7. Calculate Metrics

    - Activity and/or consistency scores for every pathway/sample selected are calculated and written to two new text files in the location of your choice. This process may take up to an hour, depending on the number of pathways/samples selected.

8. Display a Single Pathway

- Pressing '**Draw Pathway'** before loading UDP data will link to the selected pathway in a PID web browser. This currently only available for NCI/Nature and BioCarta pathways.

- Press '**Draw Pathway**' with UDP data loaded to investigate the details of a single pathway. Three summary text files are generated and saved to a folder in the directory of your choice. These files list molecule probabilities, interaction activity scores, and interaction consistency scores for all samples selected. This option is available only if a single pathway has been selected.

- Pressing '**Draw Pathway'** with the '**show biograph'** option checked generates a biograph object containing a visual representation of the pathway chosen *(version1 only).* The pathway drawing will display probabilities, activity, consistency, and/or gene hits.

    i. *In the graphic, molecules are represented by boxes, interactions by circles. Inputs to interactions are connected by a gray arrow, while interaction promoters are connected by a green arrow and inhibitors by a red arrow.*

    ii. *Molecule probabilities are represented by the color of molecule nodes (white = 0, bright aqua = 1). Molecules with unknown probabilities are gray.*

    iii. *Activity is represented by the size of interaction nodes (smaller = lower activity, larger = higher activity)*

    iv. *Consistency is represented by the color of the arrows connecting each interaction to an output (lowest = black, highest = bright blue)*

    v. *Gene hits are represented by molecule node outlines. Molecules with a positive hit are outlined in green, those with a negative hit are outlined in yellow, and those with normal measurements are outlined in thick black. Molecules without copy number information retain their thin black outline.*

    vi. *Scroll over molecules or interactions to see probabilities or activity scores.*

    vii. *Double clicking on a molecule in the drawing will open a web browser displaying information about that molecule from the CGAP website (http://cgap.nci.nih.gov). If the selected molecule is a complex, a new drawing will open first, showing the components of the complex.*

- Checking the '**text file**' option will generate a separate text file for each sample, detailing the graphic structure. This file contains a list of nodes with node label, color, shape, size, outline color, and outline width for each, as well as a list of edges with color and width for each.

9. Enter Sample Data

- The PathOlogist is capable of identifying important pathways based on information you enter about the samples. Different types of sample data allow for different statistical analyses of pathway scores.

- You may enter:

  i. Sample Labels: A more informative label for each sample, to be used in visual displays. These may or may not be unique for each sample.

  ii. Class: A grouping of samples for classification analysis. You may have as many classes as you wish, although ideally each class would contain more than one sample. Examples of classes might be Case/Control, or Cancer type (Breast/Lung/Renal etc.)

  iii. Value: A numerical value associated with each sample used for correlation analysis. Ideally the values will be on a continuous scale, and might represent features such as GI50 response to treatment, or patient's age.

  iv. Survival: A numerical value associated with each sample representing the time until some designated endpoint (eg. death, or relapse).

  v. Censor: A logical (0/1) value associated with survival, representing whether the sample reached the endpoint (or not) before observation terminated.

- To enter data, select the circle above the correct data type, and click '**Enter Data**'. A window should pop up, into which you can type/paste sample data. Data for each sample should be on a separate line. One line of data must be entered for each sample (an error message will appear otherwise), in the same order as the samples in the activity and consistency data. Click '**Submit**' to return to the Sample Data screen.

- Non-numerical entries for Value and Survival/Censor data will be converted to NaN, and those samples will not be included in correlation or survival analysis respectively. All numerical Censor data that is not '0' will be considered as logical 'true'.

10. View Heatmaps

- Once activity/consistency scores have been calculated or loaded, they can be viewed in heatmap format.

- Select pathways to include in the heatmap individually or by variance, and select samples individually or by class.

- Choose to include rows for activity and/or consistency.

- In the heatmap, similar pathways are clustered on the vertical axis, and similar samples are clustered on the horizontal axis. High activity/consistency appears as a bright red tile while low activity/consistency appears as a bright green tile. Mouse over individual tiles to see pathway/sample labels and metrics scores.

11. Find Top Pathways

- Once pathway metrics have been calculated, this feature allows identification of pathways that best reflect sample features such as class, survival, etc.

- There are 4 types of statistical analyses. Each uses a different type of sample data:

    i. Binary Classification: *Requires input of sample class in step 9.* Pathways are identified that are best able to classify samples into two groups. In the scrollbox, highlight the class(es) to comprise group 1, and click 'Set Group 1'. Then do the same for group 2 and press 'Go.' The PathOlogist will perform a 2-sample t-test for each pathway, using the scores from the 2 groups of samples, and will then return a p-value representing the probability that the 2 groups of scores were from the same distribution.

    ii. Linear Correlation: *Requires input of sample values in step 9.* Pathways are identified for which activity/consistency scores across samples have the best linear correlation with entered sample values.

    iii. Survival: *Requires input of sample survival and censor data in step 9.* For each pathway, samples are divided into a high-scoring and low-scoring group, using k-means clustering. Pathways are then identified for which survival trends in the low-scoring group are significantly different than those in the high-scoring group,

according to a Kaplan-Meier survival analysis and kruskalwallis test of significance.

iv. Gene Hits Targeting: *Requires a Gene Hits file to be loaded in step 3.* For each pathway, the PathOlogist calculates the number of genes within the pathway that were hit for each sample, and generates a p-value. A significant p-value indicates the pathway as a whole is being targeted – ie. genes within the pathway are being hit more than would be expected if hits were simply random.

- For linear correlation, survival analysis, and gene hits targeting, you may choose to include only certain classes of samples in the analysis.

- For each test, the PathOlogist will generate an ordered list of the most significant pathways along with their p-values (and rho, in the case of linear correlation, to determine the direction of the linear relationship). Once these pathways are displayed, highlight a pathway of interest and click on **'Plot Selected Pw'** to see visual confirmation of the results (not available for gene hits).

- After performing any test, click **'Write Results to File'** to generate a text file consisting of header rows detailing the samples included, and columns with the ordered list of pathways and their p-values.

- Once top pathways are identified, choose the percent of top pathways to view and return to the heatmap viewer to see sample clustering, or get a detailed look at a single pathway using the '*Draw Pathway'* feature.