Sig2Lead User Manual

Contents

Installation/Configuration

A. Installation of R and RStudio

The latest version of R is required for configuration of Sig2Lead. At the time of writing this manual, that was version 4.0.3. This version can be downloaded at:

http://www.r-project.org/

Additionally, RStudio is needed and can be downloaded at:

https://www.rstudio.com/products/rstudio/download/

B. Download Sig2Lead from Github

Sig2Lead and associated files can be downloaded from:

https://github.com/sig2lead/sig2lead_dev/
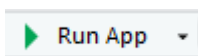
C. Installation of Dependencies/Libraries

Once R and RStudio are installed, shiny must be installed. This can be completed by typing into the R console:

install.packages("shiny")

All other dependencies and libraries will be installed upon the first time running the application.

This step may not be handled properly by MacOS.

To run the application, click the ▶ Run App ▾ button at the top middle of the RStudio interface.

Operational Workflows

A. Define Target Gene

This is the standard workflow for identifying small molecule inhibitors or activators of a target of interest. Within this workflow, a gene of interest is required and optionally, a user-defined list of compounds for scoring can be provided in SDF or SMILES format. Sig2Lead rapidly collects all data from genetic knockdowns within LINCS and identifies compounds that generate highly concordant transcriptional signatures to these genetic knockdowns (putative pathway inhibitors). This method allows scoring of small molecules to be tested for the purpose of library reduction. Results of added compounds will be ranked in descending order of similarity to LINCS analogs, with ties broken by concordance scores in the table titled "My Candidates Ranked" and compounds identified from within the LINCS library will be scored only by their concordance values to a knockdown of the target gene of interest in the table titled "LINCS Compounds Ranked" (Figure 1. Define Target Gene

workflow.). These tables can be downloaded using the download buttons just below the "Go!" button.

B. Upload a Signature

In the event that the user would like to upload a signature of their own to search for potential inhibitors to a target gene unavailable on LINCS, or molecules that generate a similar signature to some other system perturbation that is otherwise undefined, this is the workflow to choose. In this workflow, users define a signature using one of the formats defined at (www.ilincs.org/ilincs/signatures/main/). This pipeline otherwise follows the same pipeline as the "Define Target Gene" workflow.

C. Find Analogs in LINCS

This workflow simply identifies analogs to user-defined compounds. This requires added compounds and is useful in determining the baseline similarity within a compound library. LINCS compounds are all drug or drug-like, so this may be a simple filter to remove compounds that do not contain normal drug-like structures. Additionally, it was used in benchmarking as a baseline similarity of various compound libraries to the LINCS small molecules.

Tab Functionality

Figure 1. Define Target Gene workflow.

In this standard workflow, the user defines a gene of interest to identify putative pathway inhibitors. If the user defines a set of compounds from some other analyses, those compounds will be compared to the LINCS library for analogous compounds and scored by similarity and concordance in the upper table (My Candidates Ranked), which will only appear in the event of added compounds. The lower table

(LINCS Candidates Ranked) will always appear when running this workflow and scores all LINCS compounds with a concordance above the threshold in descending order of concordance.

After searching LINCS for analogs, compounds can be further analyzed through chemical similarity using the "Chemical Similarity Analysis" tab. This would be done as a standard if only Sig2Lead is run for compound identification but is not necessary to run a ceSAR analysis as described in the linked publication. To run chemical similarity clustering click the "Run SAR" button. This will initiate a somewhat slower step which captures all compounds from LINCS for the target and added compounds and clusters them by chemical similarity to one another, but will only handle up to 5,000 compounds currently to avoid a major slow down (this can be changed under advanced options if an extended time is acceptable). The output will be in the form of two figures, a heatmap and an MDS plot, and a table of centroids for each cluster (Figure 2. Chemical Similarity Analysis Tab.). The heatmap is generated through hierarchical clustering and shows a distance matrix comparing each compound identified through LINCS (Green) or added by the user (Blue). The MDS plot is an alternative view of the hierarchical clustering showing relative distances between clusters of compounds. The radius of each pie chart corresponds to the size of the cluster.

Sig2Lead

Signature Connectivity Analysis   Chemical Similarity Analysis   Network Analysis

Run SAR   Show Advanced Options

⬇ Representatives
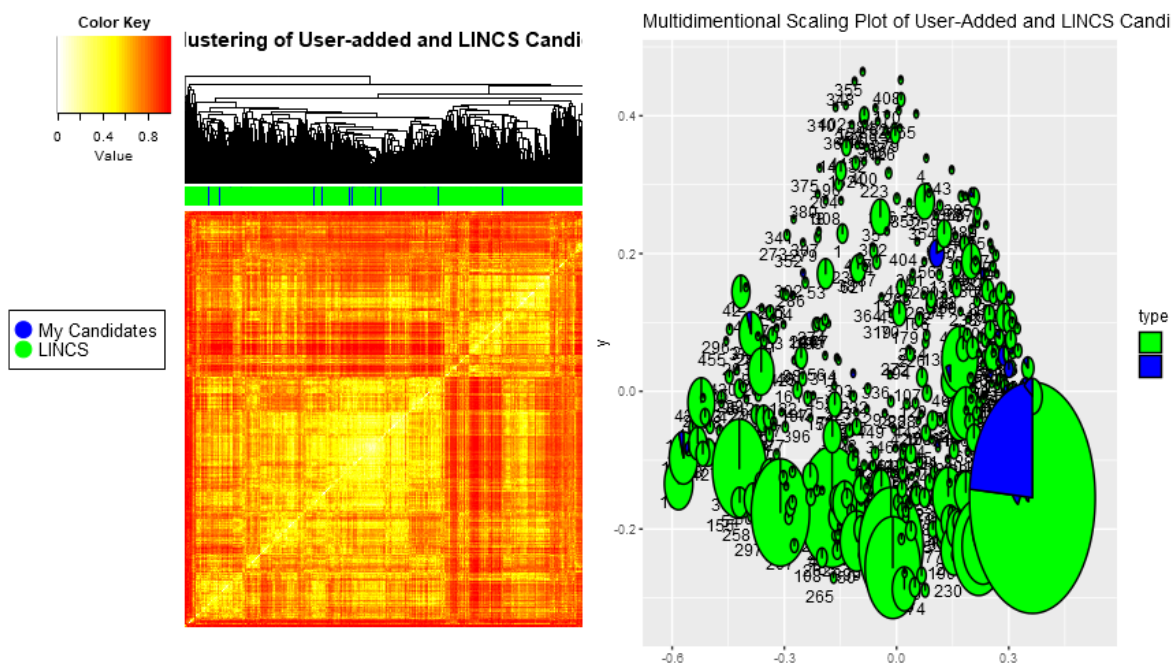
Figure 2. Chemical Similarity Analysis Tab.
Chemical similarity analyses can be performed if Sig2Lead is solely being used to identify putative inhibitors as a means to test a broader chemical space. By clustering compounds, users can ensure that each cluster has representatives tested, thus including any important chemical moieties.

Finally, Network connectivity analyses can be performed after chemical similarity analyses using STITCH on the "Network Analysis" tab. This analysis is intended to scrape any known information about identified compounds and interactions with members of the pathway. This step can be performed either through a global view (all identified compounds) or on a cluster by cluster basis (much faster). When running Global STITCH analysis, only the compounds found in clusters of sufficient size, as determined on the "Chemical Similarity Analysis" tab's advanced options will be included unless the "Shows all

clusters" box is checked. The Global STITCH analysis is a very slow process depending on the number of compounds added.

Figure 3. Cluster Network STITCH.
Under the STITCH Network analysis, known interactions will be mapped between the target gene and putative compounds submitted via Sig2Lead. The target gene is mapped in yellow, other STITCH derived genes within the pathway are mapped in gray, and compounds identified via Sig2Lead are mapped in red. This figure demonstrates that many compounds identified via Sig2Lead serve as pathway inhibitors and may even be known to inhibit other members of the pathway of interest, as opposed to a direct inhibitor.

Advanced Options
For those so inclined, many of the defaults within Sig2Lead can be adjusted.

Signature Connectivity Analysis Options
   A.   Users can select to identify activators instead of inhibitors. Functionally, this is searching for molecules that are discordant to the gene knockdown of interest instead of concordant.

B. Users can change the method of similarity search. The default, minSim is an exact chemical similarity search utilized by Sig2Lead to provide ultra-fast measurements of distance. This is a novel contribution within Sig2Lead that is currently unpublished as its own method. For comparison, a more well-known distance metric fpSim is an option, but will run about 100x slower than minSim, while generating the same results.

C. Finally, users can change the concordance threshold. The default of 0.2 is the minimum allowed threshold and corresponds to the threshold used in benchmarking for the ceSAR publication. For some targets, a more stringent cutoff may be necessary. The range of these concordances can theoretically be from -1 to 1, but realistically below 0.2 have no significant concordance and very few generate a concordance above 0.5 or 0.6.

Chemical Similarity Analysis Options

A. For clustering, different Tanimoto Similarities can be selected. The default is 0.75, which groups compounds of modest difference together, but can be adjusted to make more (closer to 1) or less stringent clusters (closer to 0). This will only be reflected in the MDS plot and representatives table.

B. The minimum cluster size for inclusion in the MDS plot can also be changed. By default, compounds that are not structurally related to at least two others at the specified threshold are considered to be independent and a representative of that group cannot be retrieved. The user can instead cluster all compounds regardless of cluster size, or increase this threshold to only consider those from large clusters of related compounds.