[**scaffoldAnalyticsJCIM.tex**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/28575ec708dd8bb0f33062c15d36b8d3781e911b...5164c3886656ccfe8cde15e2d4b514559ddb59a0#57eacaa2478e9f8de477650df3c4a120b229c96b)

[View file @5164c38](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/5164c3886656ccfe8cde15e2d4b514559ddb59a0/scaffoldAnalyticsJCIM.tex)

|  |  |  |
| --- | --- | --- |
| ... | ... | @@ -4,7 +4,7 @@ |
|  |  | *%% The document class accepts keyval options, which should include* |
|  |  | *%% the target journal and optionally the manuscript type.* |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | **\documentclass**[journal=jacsat,manuscript=article]{achemso} |
|  |  | **\documentclass**[journal=jacsat,biochem,manuscript=article]{achemso} |
|  |  |  |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | *%% Place any additional packages needed here. Only include packages* |
| ... | ... | @@ -14,6 +14,7 @@ |
|  |  | *%% servers.* |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | *%%\usepackage[version=3]{mhchem} % Formula subscripts using \ce{}* |
|  |  | **\usepackage**{mciteplus} |
|  |  | **\usepackage**[T1]{fontenc} *% Use modern font encodings* |
|  |  | **\usepackage**{hyperref} |
|  |  |  |
| ... | ... | @@ -238,7 +239,10 @@ removing molecules that contain them until a threshold is met, |
|  |  | yielding a set of disjoint frameworks. Other methods have used |
|  |  | multiple common substructure (MCS), first proposed for finding protein |
|  |  | structural similarity**\cite**{Koch1997MCSprot}, for example |
|  |  | **\citet**{Quintus2009MCS} and the ChemAxon product LibraryMCS. |
|  |  | **\citet**{Quintus2009MCS} and the ChemAxon product LibraryMCS. Recent work |
|  |  | has leveraged the Graph Edit Distance to define the MCS**\cite**{SmallWorld}, |
|  |  | and calculated it on reduced graphs(**\cite**{GH2019GED},**\cite**{Harper2004DDclus}) |
|  |  | as a more general way to define similarity and retrieve analogs. |
|  |  |  |
|  |  | *% To overcome the effect of small variations in heteroatoms (eg. O to S)* |
|  |  | *% mapping otherwise similar molecules to different scaffolds,* |
| ... | ... | @@ -267,8 +271,10 @@ reduces repeatability. |
|  |  | \end{figure} |
|  |  |  |
|  |  | Assigning each molecule to a single cluster partitions the dataset into |
|  |  | disparate, non-overlapping groups (also termed a hard clustering; |
|  |  | agglomerative fingerprint-based clustering methods such as sphere exclusion, single and complete linkage clustering**\cite**{Downs2003} work this way. On the other hand, a molecule could be assigned simultaneously to multiple clusters, depending on the structural features present. Traditional |
|  |  | disparate, non-overlapping groups, also termed a hard clustering; |
|  |  | agglomerative fingerprint-based clustering methods such as sphere exclusion, single and complete linkage clustering**\cite**{Downs2003} work this way. These methods tend to place structurally analogous molecules in different clusters where they would not be discovered as analogs of each other - see an example in Supplementary Material Section S8. |
|  |  |  |
|  |  | On the other hand, a molecule could be assigned simultaneously to multiple clusters, depending on the structural features present. Traditional |
|  |  | hard clustering methods will not allow this and will assign such molecules |
|  |  | to their own cluster, thus identifying them |
|  |  | as singletons. Such a result can be observed in real chemical datasets |
| ... | ... | @@ -276,6 +282,7 @@ as well -- for example, the complete-linkage clustering of the TCAMS |
|  |  | dataset**\cite**{Gamo2010,Calderon2011} has nearly 25**\%** of the 2000 |
|  |  | clusters identified with just one compound, as shown in **\fref**{platypus}. |
|  |  |  |
|  |  |  |
|  |  | In contrast to hard clustering, the methods described |
|  |  | here rely on fuzzy clustering, where a molecule may be assigned to |
|  |  | multiple clusters. Fuzzy clusters have rarely been used in |
| ... | ... | @@ -310,13 +317,12 @@ assays: |
|  |  | **\item** A {**\it** full HTS} run in 2012 with {**\bf** 4564} {pIC50s} measured after screening 2 million compounds at a single concentration (10 uM). |
|  |  | **\item** A {**\it** top-up HTS} run in 2014 with {**\bf** 3613} {pIC50s} from screening 350,000 compounds. |
|  |  | \end{itemize} |
|  |  | A fourth dataset comprises {**\bf** 824} virtual compounds selected from |
|  |  | {**\it** ELT}, a DNA Encoded Library screen**\cite**{ELT} of 130 |
|  |  | combinatorial libraries with billions of potential molecules. |
|  |  | A fourth dataset comprises {**\bf** 824} so-called virtual compounds whose activity was inferred from {**\it** ELT}, a DNA Encoded Library screen**\cite**{ELT} of 130 |
|  |  | combinatorial libraries comprising billions of potential molecules. Triaged ELT compounds are synthesized off-DNA to check if they are indeed active. |
|  |  |  |
|  |  | The GSK Kinase dataset also contains physico-chemical properties of interest for |
|  |  | developability, notably the Property Forecast Index (PFI**\cite**{Young2011}). |
|  |  | We chosen this dataset to illustrate the power of Scaffold Analytics in joining |
|  |  | We chose this dataset to illustrate the power of Scaffold Analytics in joining |
|  |  | and merging datasets from multiple screens, combining their SAR to |
|  |  | design hybrid molecules, and making inferences about unknown activity |
|  |  | in one screen based on known activity in another screen. |
| ... | ... | @@ -346,7 +352,7 @@ created a simple workflow in Pipeline Pilot\cite{PPilot}, but this can also be d |
|  |  | via standard cheminformatics toolkits such as JChem**\cite**{JChem} or RDKit**\cite**{RDKit}. Prior to |
|  |  | SDF conversion, activity or property columns that are not to be |
|  |  | aggregated at the scaffold level should be deleted from the CSV file, |
|  |  | to accelerate analysis and aggregation. Further quirks specific to the NCATS R-group tool |
|  |  | to accelerate analysis and aggregation. Further quirks specific to one of the methods, the NCATS R-group tool, |
|  |  | will be described in the Supplementary Material in section S3. |
|  |  |  |
|  |  | **\subsection**{Partitioning Method: Complete Linkage Clustering} |
| ... | ... | @@ -360,7 +366,7 @@ as \citet{Jain2010,Downs2003}. |
|  |  |  |
|  |  | **\subsection**{Fragmentation Method: NCATS R-Group Tool} |
|  |  | **\label**{sec:rgtool} |
|  |  | The NCATS R-group analysis tool (**\url**{https://tripod.nih.gov/?p=46}) |
|  |  | The NCATS R-group analysis tool (**\url**{https://tripod.nih.gov/?p=46}, **\cite**{RGTool}) |
|  |  | was developed to automatically and exhaustively generate R-group |
|  |  | tables from a dataset using all scaffolds, defined as chemical |
|  |  | substructures shared by two or more molecules. The scaffolds are |
| ... | ... | @@ -373,7 +379,7 @@ symmetry and additional constraints (\eg reactivity, synthetic |
|  |  | accessibility). Briefly, the fragments are generated based on the |
|  |  | following rules: |
|  |  | \begin{enumerate} |
|  |  | **\item** Generate Bemis-Murcko framework by iteratively pruning all foliage |
|  |  | **\item** Generate Bemis-Murcko framework by iteratively pruning all pendant atoms |
|  |  | except for carbonyl (or any terminal double bond). |
|  |  | **\item** Exhaustively enumerate all possible combinations of ring |
|  |  | system. This is achieved by iteratively breaking non-aromatic |
| ... | ... | @@ -394,12 +400,12 @@ the TCAMS dataset. |
|  |  | *%(d)\includegraphics[width=3in]{fig/tcam1\_SNG3.png}* |
|  |  | *%\caption{Scaffold Decompositions for a molecule (a) from the TCAMS dataset with PubChem Compound ID 536182. (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 frameworks -- first 21 Bemis-Murcko Like and last 3 in the bottom row RECAP. (d) Scaffold Network generated by SNG, starting from top-level scaffold with four rings down to to all subscaffolds with two rings.}* |
|  |  |  |
|  |  | (a)**\includegraphics**[width=3in]{fig/tcam1\_mol.png}**\\** |
|  |  | (a)**\includegraphics**[width=3in]{fig/tcam1\_mol\_v2.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | (b)**\includegraphics**[width=5in]{fig/tcam1\_RGscaf.png}**\\** |
|  |  | (b)**\includegraphics**[width=5in]{fig/tcam1\_RGscaf\_v2.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | (c)**\includegraphics**[width=5in]{fig/tcam1\_GSKframes.png} |
|  |  | **\caption**{Scaffold Decompositions for (a) Molecule 1 from the TCAMS dataset with Compound ID 536182 (PubChem CID 44522854). (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 frameworks -- first 21 Bemis-Murcko Like and last 3 in the bottom row RECAP.} |
|  |  | (c)**\includegraphics**[width=5in]{fig/tcam1\_FW\_v2.png} |
|  |  | **\caption**{Scaffold Decompositions for (a) Molecule 1 from the TCAMS dataset with Compound ID 536182 (PubChem CID 44522854). (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 Bemis-Murcko Like frameworks.} |
|  |  | **\label**{fig:scafmethod} |
|  |  | \end{figure} |
|  |  |  |
| ... | ... | @@ -521,7 +527,7 @@ datasets. What gets added follows the schema ``Molecule --> Scaffold (including |
|  |  | dataset, we connect it to every Scaffold/Framework/Cluster it |
|  |  | contains, and then to every other molecule containing any of these |
|  |  | Scaffolds/Frameworks/Clusters. Slight differences for each individual |
|  |  | method are detailed in the Supplementary Material, Section S6. |
|  |  | method are detailed in the Supplementary Material, Section S5. |
|  |  |  |
|  |  | **\subsection**{Visualization of Molecules, Scaffolds and Related Molecules} |
|  |  |  |
| ... | ... | @@ -536,18 +542,19 @@ following elements: |
|  |  | selectivity, ligand efficiency and molecular properties may be |
|  |  | highlighted on the X, Y, shape, size and color axes on a scatter plot. |
|  |  | This view is depicted in compressed form in the top half of **\fref**{spotviz}(a). |
|  |  | **\item** {**\bf** Related Molecules Tab}: The purpose of this tab is to |
|  |  | implement the Scaffold Walking navigation described briefly earlier. |
|  |  | **\item** {**\bf** Related Molecules Tab}: The purpose of this tab is to explore the |
|  |  | set of Related Molecules to molecule(s) of interest, and thus implement the Scaffold Walking navigation briefly later in **\sref**{scafwalk}. |
|  |  | The setup is described for the NCATS R-group Tool decomposition, |
|  |  | though this tab applies to and can be set up analogously for any |
|  |  | other decomposition. The tab consists of two visualizations, |
|  |  | illustrated for the TCAMS dataset in **\fref**{spotviz}(a): **\subitem** The |
|  |  | first one is a miniature version of the Main window, allowing the |
|  |  | illustrated for the TCAMS dataset in **\fref**{spotviz}(a): |
|  |  | **\subitem** Miniature version of the {**\bf** Main window}, allowing the |
|  |  | user to select (in Spotfire, mark) molecules of interest without |
|  |  | flipping over to the Main tab. Doing so drives one of the following |
|  |  | two Details Visualizations on the {RG}decomp table, showing only |
|  |  | molecules from the scaffolds contained in the marked molecule, **\ie** |
|  |  | Related Molecules: **\subitem** {**\bf** Scaffold Trellis}: This scatter |
|  |  | Related Molecules. |
|  |  | **\subitem** {**\bf** Scaffold Trellis}: This scatter |
|  |  | plot is trellised by Scaffold ID and ideally displays the same |
|  |  | properties on the axes as the Main visualization above it. An |
|  |  | example is shown in **\fref**{ELT}. The trellis allows us to break up |
| ... | ... | @@ -559,10 +566,11 @@ following elements: |
|  |  | groups of points that are laid out similarly across multiple trellis |
|  |  | panels. Though some of our users still prefer this approach, we now |
|  |  | describe a newer solution that better leverages Spotfire's |
|  |  | capabilities. **\subitem** {**\bf** Scaffold Pies}: Instead of using a |
|  |  | capabilities. |
|  |  | **\subitem** {**\bf** Scaffold Pies}: Instead of using a |
|  |  | trellis, the Marker shape is changed to Pies, with Colors (which map |
|  |  | to pie sectors) by Scaffold ID, and sectors sized by the Count of |
|  |  | molecules in each scaffold. The result is illustrated in |
|  |  | molecules containing each scaffold. The result is illustrated in |
|  |  | **\fref**{spotviz}(a). This plot shows only one point per related |
|  |  | molecule but one sector for each scaffold it shares with the parent |
|  |  | molecule. As described in Section **\ref**{sec:results}, this lets the |
| ... | ... | @@ -576,15 +584,20 @@ following elements: |
|  |  | \end{itemize} |
|  |  |  |
|  |  | \begin{figure} |
|  |  | (a)**\includegraphics**[width=5in]{fig/spotviz\_scafpie\_tooltip.png}**\\** |
|  |  | (b)**\includegraphics**[width=5in]{fig/spotviz\_scaffolds\_aggr.png}**\\** |
|  |  | (c)**\includegraphics**[width=5in]{fig/spotviz\_RGtable.png} |
|  |  | **\caption**{Elements of the minimal Spotfire interface we developed for Scaffold-Based Analytics: (a) Related Molecules page, with molecule selection in the top panel and related molecules viewed as one pie sector per scaffold shared; (b) Aggregate plot of scaffold statistics (with often a detailed drill-down into the selected scaffold); (c) R-group table on selected scaffolds, sorted by scaffold ID and decreasing pIC50 or other desired properties.} |
|  |  | *%(a)\includegraphics[width=5in]{fig/spotviz\_scafpie\_tooltip.png}\\* |
|  |  | *%(b)\includegraphics[width=5in]{fig/spotviz\_scaffolds\_aggr.png}\\* |
|  |  | *%(c)\includegraphics[width=5in]{fig/spotviz\_RGtable.png}* |
|  |  | (a)**\includegraphics**[width=5in]{fig/Fig4a\_top\_struc.png}**\\** |
|  |  | (b)**\includegraphics**[width=5in]{fig/Fig4b\_scafpie\_struc.png}**\\** |
|  |  | (c)**\includegraphics**[width=5in]{fig/Fig4c\_scafaggr\_struc.png}**\\** |
|  |  | (d)**\includegraphics**[width=5in]{fig/Fig4d\_RGtable\_struc.png} |
|  |  |  |
|  |  | **\caption**{Elements of the minimal Spotfire interface we developed for Scaffold-Based Analytics: Related Molecules page, with (a) molecule selection and (b) related molecules viewed as one pie sector per scaffold shared; (c) Aggregate plot of scaffold statistics; (d) R-group table on selected scaffolds, sorted by scaffold ID and decreasing pIC50 or other desired properties.} |
|  |  | **\label**{fig:spotviz} |
|  |  | \end{figure} |
|  |  |  |
|  |  | Alternative visualizations are possible and we describe some of them |
|  |  | in the Supplementary Material, Section S7. In addition, we discuss several |
|  |  | in the Supplementary Material, Section S6. In addition, we discuss several |
|  |  | Spotfire tricks that are instrumental in making our visualization |
|  |  | useful to the chemist or biologist |
|  |  | user. *%\textbf{\textcolor{red}{TODO}}* |
| ... | ... | @@ -593,10 +606,10 @@ user. %\textbf{\textcolor{red}{TODO}} |
|  |  | **\label**{sec:statmethod} |
|  |  | In this section we define the statistical methods used to compare the |
|  |  | scaffold generation methods in the context of annotating molecules in |
|  |  | a screening dataset with multiple, possibly overlapping scaffolds (as |
|  |  | opposed to partitioning the set of molecules into non-overlapping |
|  |  | categories, a.k.a. clustering). The code to implement these methods is released |
|  |  | as part of the Supplementary Material in Section S8. |
|  |  | a screening dataset with multiple, possibly overlapping scaffolds -- a |
|  |  | generalization of partitioning the set of molecules into non-overlapping |
|  |  | categories, a.k.a. clustering. The code to implement these methods is released |
|  |  | as part of the Supplementary Material in Section S10. |
|  |  |  |
|  |  | The similarities and differences between non-overlapping clusterings |
|  |  | have been analyzed in the literature, and one example is the method of |
| ... | ... | @@ -644,7 +657,7 @@ PI\_A(C) = \| C\_A \| / \| C\_A \cup C\_B \| |
|  |  | In constrast, the Proportion of Information Unique to $A, PIU\_A$ uses the set difference between $C\_A$ and $C\_B$ to get at the question: if I have $B$, do I still need $A$? |
|  |  |  |
|  |  | \begin{equation} |
|  |  | PIU\_A(C) = **\|** C\_A **\setminus** C\_B **\|** / **\|** C\_A **\cup** C\_B **\|** = 1 - PI\_c(B) |
|  |  | PIU\_A(C) = **\|** C\_A **\setminus** C\_B **\|** / **\|** C\_A **\cup** C\_B **\|** = 1 - PI\_B(C) |
|  |  | \end{equation} |
|  |  |  |
|  |  | When comparing one fragmentation method against another, we often see |
| ... | ... | @@ -661,23 +674,26 @@ FragEff\_A(C) = \| C\_A \| / \| frag\_A(C) \| |
|  |  | \end{equation} |
|  |  |  |
|  |  |  |
|  |  | These statistics can be averaged over all compounds in a dataset to |
|  |  | yield the Average Common Proportion (ACP), Average Proportion of |
|  |  | Information (API), Average Proportion of Information Unique to A |
|  |  | (APIU) or Average Fragment Efficiency (AFE). Other statistics can |
|  |  | also be applied to the distribution of $CP$, $PI$ or $PIU$ to |
|  |  | characterize the dataset and overlapping scaffolds used to |
|  |  | characterize it. |
|  |  |  |
|  |  | Our methods extend the similarity score of **\citet**{Torres2009} to |
|  |  | overlapping clusters by using a set of clusters derived from them, |
|  |  | defined as follows: |
|  |  | \begin{itemize} |
|  |  | **\item** Replace the original clusters by new clusters, one for each item |
|  |  | in any cluster (in this case, any compound in the dataset). |
|  |  | **\item** The cluster for Compound $C$ contains all compounds that appear |
|  |  | with Compound $C$ in any cluster, **\ie** $C\_A$. |
|  |  | \end{itemize} |
|  |  | The distribution of these statistical measures for our dataset is |
|  |  | used to compare the overlapping clustering methods (NCATS RG-Decomposition and Molecular Frameworks) and a partitioning clustering method (Complete linkage) in **\sref**{statcomp}. |
|  |  | *%complete-* |
|  |  | *%can be averaged over all compounds in a dataset to* |
|  |  | *%yield the Average Common Proportion (ACP), Average Proportion of* |
|  |  | *%Information (API), Average Proportion of Information Unique to A* |
|  |  | *%(APIU) or Average Fragment Efficiency (AFE). Other statistics can* |
|  |  | *%also be applied to the distribution of $CP$, $PI$ or $PIU$ to* |
|  |  | *%characterize the dataset and partitioning or overlapping scaffolds used to* |
|  |  | *%characterize it.* |
|  |  |  |
|  |  | *%Our methods extend the similarity score of \citet{Torres2009} to* |
|  |  | *%overlapping clusters by using a set of clusters derived from them,* |
|  |  | *%defined as follows:* |
|  |  | *%\begin{itemize}* |
|  |  | *%\item Replace the original clusters by new clusters, one for each item* |
|  |  | *% in any cluster (in this case, any compound in the dataset).* |
|  |  | *%\item The cluster for Compound $C$ contains all compounds that appear* |
|  |  | *% with Compound $C$ in any cluster, \ie $C\_A$.* |
|  |  | *%\end{itemize}* |
|  |  |  |
|  |  |  |
|  |  | **\section**{Results} |
| ... | ... | @@ -696,8 +712,8 @@ molecule's activity, and use this information to design or test |
|  |  | further compounds. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | (a)**\includegraphics**[width=2in]{fig/tcam2\_mol\_541564.png} |
|  |  | (b)**\includegraphics**[width=5in]{fig/RGT\_aggr\_prop2.png} |
|  |  | (a)**\includegraphics**[width=2in]{fig/tcam2\_mol\_541564\_v2.png} |
|  |  | (b)**\includegraphics**[width=5in]{fig/RGT\_aggr\_prop3.png} |
|  |  | **\caption**{(a) Molecule 2 (TCAMS Compound ID: 541564; PubChem CID: 44531725). |
|  |  | (b) All **\~**5000 scaffolds from the TCAMS dataset ranked by |
|  |  | average pIC50 in the P.~falciparum 3D7 strain and Inhibition |
| ... | ... | @@ -720,8 +736,7 @@ molecules sharing this scaffold are shown in \fref{KinaseX}(b). Clear differenti |
|  |  | **\label**{fig:KinaseX} |
|  |  | \end{figure} |
|  |  |  |
|  |  |  |
|  |  | To conclude, we have observed how scaffold-based analytics allows us to make decisions based on the aggregate (average, SD, median, min, max, ...) properties of molecules in scaffolds. An entire scaffold may be rejected or prioritized at a time instead of keeping track of individual molecules. It is important to note that rejecting a scaffold does not reject all molecules containing that scaffold - otherwise rejecting the benzene ring would remove a large percentage of valid leads. |
|  |  | In summary, scaffold-based analytics allows us to make decisions based on the aggregate (average, SD, median, min, max, ...) properties of molecules in scaffolds. An entire scaffold may be rejected or prioritized at a time instead of keeping track of individual molecules. It is important to note that rejecting a scaffold does not reject all molecules containing that scaffold - otherwise rejecting the benzene ring would remove a large percentage of valid leads. |
|  |  | *%Also at any time we are able to drill down into the data for individual compounds in interesting scaffolds, whether in an R-group table or a custom plot, and incorporate that into our prioritization.* |
|  |  |  |
|  |  | **\subsection**{Use Case: Dataset Fusion and Hybridization} |
| ... | ... | @@ -730,15 +745,11 @@ In the previous section and \fref{KinaseX}(b), we noted that the 8 molecules sha |
|  |  | In this way, Scaffold-Based analytics may be used to combine multiple datasets using the scaffold as the common unit of comparison and merging, with dataset labels identifying molecules in the resulting merged dataset. If the datasets have comparable activities, as the Kinase X HTS datasets did, aggregate statistics are easy to compute. If they do not, **\eg** if one dataset has $pIC\_{50}$ and others primary screening activity or a different measure of desirability, normalized activities may be computed to enable aggregation at the scaffold level. |
|  |  |  |
|  |  | Sometimes one of the datasets does not contain activities at all, **\eg** |
|  |  | in the case of virtual molecules from a DNA Encoded Library Technology**\cite**{ELT} |
|  |  | (ELT) screen, or for molecules from a vendor catalog that are yet to |
|  |  | be ordered and assayed. In **\fref**{ELT} we show an example where two |
|  |  | computed properties, Molecular Weight and PFI were used to identify |
|  |  | promising virtual molecules from the ELT dataset to make, and untested |
|  |  | molecules from the FBDD**\cite**{FBDD} (Fragment-Based Drug Design) dataset to test. |
|  |  |  |
|  |  | in the case of the ``virtual'' hits from DNA Encoded Library Technology**\cite**{ELT} |
|  |  | (ELT) screening that are yet to be synthesized off-DNA, or for molecules from a vendor catalog that are yet to be ordered and assayed. In **\fref**{ELT} we show an example where two computed properties, Molecular Weight and PFI were used to identify promising untested molecules to make (ELT dataset) or test (FBDD dataset**\cite**{FBDD}). |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/KinaseX\_ELT.png} |
|  |  | **\includegraphics**[width=6in]{fig/KinaseX\_ELT\_v2.png} |
|  |  | **\caption**{ |
|  |  | Two scaffolds with ID 1920 (pyridophenyl) and 3116 (benzimidazole) shown in a trellised plot of molecules related to a weak active in the Kinase X 2014 HTS screen. The axes of the plots are computed properties, MW and PFI, to allow molecules without measured activity or virtual molecules not yet synthesized to be included in the plot. The compass arrow device points in the direction of improved properties -- lower MW and PFI. Interesting related molecules are labeled on the plot, including untested fragment-like molecules from the FBDD dataset, virtual molecules from the ELT dataset, and another ligand-efficient hit from the HTS2014 dataset, all with low MW/PFI. The unmeasured and unmade molecules are good candidates for future synthesis and testing. |
|  |  | } |
| ... | ... | @@ -748,94 +759,31 @@ Two scaffolds with ID 1920 (pyridophenyl) and 3116 (benzimidazole) shown in a tr |
|  |  |  |
|  |  | **\subsection**{Use Case: Scaffold Walking Navigation} |
|  |  | **\label**{sec:scafwalk} |
|  |  | Scaffold Walking is our term for navigating from molecule(s) through scaffolds (implicitly) to Related Molecules. This contrasts with Scaffold Hopping, which is usually defined as a complete replacement of a scaffold by another 2D-dissimilar but 3D-similar or bioisosteric scaffold. Scaffold Walking is meant to be a gradual change to the molecule, at each step retaining at least one element of its maximal Murcko scaffold (**\ie** at least one among the multiple scaffolds it shares with other molecules in the dataset). In the process the SAR gets deconvoluted in terms of these scaffolds, allowing us to determine visually both the most essential scaffolds in a molecule and the best Related Molecules containing them. |
|  |  | Scaffold Walking is our term for navigating from molecule(s) through scaffolds (implicitly) to Related Molecules. This contrasts with Scaffold Hopping, which is usually defined as a complete replacement of a scaffold by another 2D-dissimilar but 3D-similar or bioisosteric scaffold. Scaffold Walking is meant to be a gradual change to the molecule, at each step retaining at least one element of its maximal Murcko scaffold (**\ie** at least one among the multiple scaffolds it shares with other molecules in the dataset). The concept is similar to using Graph Edit Distance**\cite**{GH2019GED}. In the process the SAR gets deconvoluted in terms of these scaffolds, allowing us to determine visually both the most essential scaffolds in a molecule and the best Related Molecules containing them. |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/mol1\_RGtool\_scafpie.png} |
|  |  | **\caption**{Related Molecules scaffold pies visualization for Molecule 1 (TCAMS Compound ID: 536182, PubChem CID 44522854). Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the figure and also discussed in the text.} |
|  |  | **\includegraphics**[width=6in]{fig/mol1\_RGtool\_scafpie\_top\_v4.png} |
|  |  | **\includegraphics**[width=6in]{fig/mol1\_RGtool\_scafpie\_struc\_v4.png} |
|  |  | **\caption**{Related Molecules scaffold pies visualization for Molecule 1 (TCAMS Compound ID: 536182, PubChem CID 44522854). Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the bottom part of the figure (showing structures) and also discussed in the text.} |
|  |  | **\label**{fig:scafwalk1} |
|  |  | \end{figure} |
|  |  |  |
|  |  | As an example, consider Molecule 1 (TCAMS Compound ID: 536182, PubChem CID 44522854) in **\fref**{scafwalk1}, as a hit that we want to explore SAR of and optimize. This molecule contains 5 scaffolds as determined by the NCATS R-group tool. Using the Scaffold Pie visualization, we observe that the naphthyl scaffold (**\#**978, cyan) is by itself only moderately active. The dihydrotriazine (**\#**4719, blue) scaffold is observed to always occur where the dihydrotriazine-phenethyl-ether (**\#**2467, pink) one does, implying the substructure relationship between them visually even if one did not know it beforehand. Scaffold **\#**4719 also exists and is active without **\#**2467, implying that the phenethyl ether can be substituted and the dihydrotriazine may be sufficient for activity by itself. |
|  |  | As an example, consider Molecule 1 (TCAMS Compound ID: 536182, PubChem CID 44522854) in **\fref**{scafwalk1}, as a hit that we want to explore SAR of and optimize. This molecule contains 5 scaffolds as determined by the NCATS R-group tool. Using the Scaffold Pie visualization, we observe that the naphthyl scaffold (**\#**978, yellow) is by itself only moderately active. The dihydrotriazine (**\#**4719, blue) scaffold is observed to always occur where the dihydrotriazine-phenethyl-ether (**\#**2467, pink) one does, implying the substructure relationship between them visually even if one did not know it beforehand. Scaffold **\#**4719 also exists and is active without **\#**2467, implying that the phenethyl ether can be substituted and the dihydrotriazine may be sufficient for activity by itself. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/mol2\_RGtool\_scafpie2.png} |
|  |  | **\caption**{Related Molecules scaffold pies visualization for Molecule 3 (TCAMS Compound ID: 533945; PubChem CID: 44531163). Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the figure and also discussed in the text.} |
|  |  | **\label**{fig:scafwalk2} |
|  |  | \end{figure} |
|  |  | Further examples of the value of Scaffold Walking are in the Supplementary Material, Section S7. |
|  |  |  |
|  |  | Molecule 3 in **\fref**{scafwalk2} is another case where we might want to optimize the physical and chemical properties of the molecule without sacrificing activity or increasing promiscuity. Since solubility has been shown to decrease with number of aromatic rings independent of lipophilicity**\cite**{Hill2010}, walks that remove one or more of the three fused rings might be beneficial. By exploring the Related Molecules, we observe the SAR for three scaffolds: quinazolines (**\#**305, brown), indazoloquinazolines (**\#**574, blue) and indazoles (**\#**3822, pink). We observe from the plot a few more molecules containing all three scaffolds (tricolored pies, **\ie** exact analogs of the parent molecule); all of these are less active than the parent. The indazoles when they occur alone (pink circles) are far less active than the parent, suggesting they do not contribute significantly to activity and may be substituted. Lastly the quinazolines (brown) include several analogs that are more active and also less promiscuous than the parent. Drilling down into these structures, we observe several that contain only two aromatic rings (**\ie** no more are either fused or attached); these provide novel, active and ligand-efficient templates on which to build new analogs with enhanced solubility or other properties. Suggestions for which analogs to make can often be obtained by examining the SAR -- for example, disparate aliphatic and aromatic analogs at two adjacent positions on the quinazoline phenyl are active, which suggests hybridizing them or designing further analogs substituted at these positions. |
|  |  | **\subsection**{Statistical Comparison of Scaffold-Generation Methods and Clustering}**\label**{sec:statcomp} |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=5.5in]{fig/mol3\_RGtool\_scafpie\_iter.png} |
|  |  | **\caption**{Iterative Related Molecules scaffold pies visualization for Molecule 4 (TCAMS Compound ID: 541531, PubChem CID: 44531903), a scaffold hop from Molecule 3 shown in **\fref**{scafwalk2}. Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the figure and also discussed in the text.} |
|  |  | **\label**{fig:scafwalk3} |
|  |  | \end{figure} |
|  |  | Recall the concepts of structure group and common proportion defined in **\sref**{statmethod}, and illustrated with an example in Supplementary Material, Section S9. Let us now apply these to analyze and compare the scaffold generation methods Frameworks (denoted $A$ below) and NCATS R-group tool ($B$), and compare $B$ with clustering method Complete-Linkage clustering ($D$). |
|  |  |  |
|  |  | Another intriguing result is seen by observing a new tricyclic series that shares the quinazoline but adds an indole instead of indazole as the third fused ring. This alternative tricyclic template, while it may not confer solubility advantages, opens up a new area of chemical space. By iteratively seeking the Related Molecules for this new hit as shown in **\fref**{scafwalk3}, we observed that most of the active quinazoline analogs have this new tricyclic scaffold (**\#**1824, tan wedges in tricolored pies) and relatively few contain only quinazolines (**\#**305, brown circles). Also, indoles by themselves (**\#**1188, blue circles) do not much better than indazoles, so this is a new synergistic effect discovered by scaffold walking from the original hit. |
|  |  |  |
|  |  |  |
|  |  | **\subsection**{Qualitative Comparison of Scaffold-Generation Methods and Clustering} |
|  |  | {**\bf** Complete-Linkage Clustering}: As shown in **\fref**{clusterlanes}, the defining feature of a partitioning clustering is that every molecule maps to one and only one cluster. Thus if a chemotype is broken up among two or more clusters, using the cluster ID to map Related Molecules can retrieve only neighbors from the same cluster, ignoring the other cluster. This is not ideal for purposes of the visualization and navigation method presented here, as arbitrary neighbors would be excluded depending on how the clustering is defined. Thus we do not advocate the use of clustering, unless it is a fuzzy clustering where all meaningful class memberships a molecule might have are considered. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/clusterlanes.png} |
|  |  | **\caption**{Illustrating one problem with clustering: bifurcation of related molecules. When two molecules of the same chemotype differing by a halogen are split across Complete Linkage Clusters, searches of cluster neighbors for one molecule do not find its analogs in the other cluster, **\ie** the two related clusters are not linked.} |
|  |  | **\label**{fig:clusterlanes} |
|  |  | \end{figure} |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  | {**\bf** NCATS R-group tool}: As opposed to the clustering method, |
|  |  | if any two molecules share a common substructure that meets the standards required of a scaffold by the NCATS method (**\eg** being bordered by rings on each end), then those molecules will be found to contain that shared substructure as a scaffold and their activities will be used to compute aggregate properties for it. |
|  |  |  |
|  |  | {**\bf** Other Scaffold Generation Methods}: As shown earlier, even though another scaffold generation method (represented here by the **\citet**{Harper2004DDclus} implementation of Frameworks) differed in its implementation details and produced different numbers of scaffolds for the same molecule, it was roughly equivalent in a qualitative sense with regard to the insights obtained. Due to substantial overlap between sets of scaffolds, ring systems responsible for activity of a molecule were generally revealed by either method. For example, the insights mentioned in **\sref**{scafwalk} were more or less consistent across the methods. However, there were cases where the Frameworks revealed negative information about a fragment being not important for activity that is also useful for a drug discovery scientist. For example, in **\fref**{frameswalk} a substructure is highlighted that is on the aggregate inactive and could be removed or substituted. This insight is not available from SSSR-based scaffolding methods such as the NCATS R-group tool since they don't define or find that fragment as a scaffold. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=5in]{fig/mol1\_frames\_scafpie.png} |
|  |  | **\caption**{Using Frameworks with the Scaffold Pies visualization. One framework is highlighted that has no equivalent in the NCATS scaffolds, but is shown to reduce activity as related molecules containing it are less active than the parent molecule. The star symbol shows the location of the parent molecule in this Related Molecules plot, and the compass device at the origin shows the direction of favorable properties (+X and +Y axes).} |
|  |  | **\label**{fig:frameswalk} |
|  |  | \end{figure} |
|  |  |  |
|  |  | *%\begin{figure}* |
|  |  | *%\includegraphics[width=5in]{fig/mol2\_SNG\_relmol\_trellis.png}* |
|  |  | *%\caption{Using SNG to find Related Molecules a few hops away in the Scaffold Network that have a desirable property, in this case activity against the resistant DD2 strain of P.~falc. The star symbol shows the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (+X and +Y axes).}* |
|  |  | *%\label{fig:SNGwalk}* |
|  |  | *%\end{figure}* |
|  |  | *%* |
|  |  | *%Using the Scaffold Network to visualize Related Molecules, one is able to deconvolute scaffolds with different numbers of rings (which serves as the level in the hierarchy) by trellising on it, and thus gather together substructures of comparable size. As shown in \fref{SNGwalk}, with the same tricyclic molecule as in \fref{scafwalk2}, we use Scaffold Networks to solve a selectivity issue. Members of the original scaffold, shown in light brown, tend to have some activity against the susceptible 3D7 strain of $P.~falciparum$ but be inactive against the multidrug-resistant DD7 strain. Using the scaffold network to go one level down (to molecules with two rings) and then two levels up (up to 5 rings) we obtain many new molecules that are a few hops away from the top-level tricyclic scaffold while still sharing some aspect of it. Looking at where these new Related Molecules are placed, we mark one scaffold with 3 rings that has a quinazoline linked to a phenyl, and one with 4 rings that contains an indoloquinazoline (also discovered by scaffold walking with the NCATS R-group tool as described above) linked to a phenyl. These two scaffolds are highlighted since their members consistently have high activity against the resistant DD2 strain, and thus this would be a desirable scaffold hop (or walk) towards chemical space that is more useful in overcoming resistance from the malarial parasite.* |
|  |  |  |
|  |  | *%To summarize, all three multiple-scaffold decomposition methods considered in this study, \ie NCATS R-group Tool, Frameworks and Scaffold Network Generator give comparable insights when exploring the TCAMS dataset, with some differences stemming from individual substructures that are considered shared scaffolds or not by the individual methods. We now explore these overlaps, similarities and differences in the aggregate using the statistical methods described earlier in \sref{statmethod}.* |
|  |  |  |
|  |  | To summarize, both multiple-scaffold decomposition methods considered in this study, **\ie** NCATS R-group Tool and Frameworks give comparable insights when exploring the TCAMS dataset, with some differences stemming from individual substructures that are considered shared scaffolds or not by the individual methods. We now explore these overlaps, similarities and differences in the aggregate using the statistical methods described earlier in **\sref**{statmethod}. |
|  |  |  |
|  |  |  |
|  |  | **\subsection**{Statistical Comparison of Scaffold-Generation Methods}**\label**{sec:statcomp} |
|  |  |  |
|  |  | Recall the concepts of structure group and common proportion defined in **\sref**{statmethod}. Let us now apply these to analyze and compare the scaffold generation methods Frameworks ($A$) and NCATS R-group tool ($B$). As an illustration, **\fref**{strucgroup} (a)-(c) shows the structure group of the compound $C$ with Compound ID 541564 (PubChem CID: 44531725). |
|  |  | \begin{figure} |
|  |  | **\centering** |
|  |  | \begin{minipage}[b][0.2**\textheight**][s]{0.3**\textwidth**} |
|  |  | **\centering** |
|  |  | (a)**\includegraphics**[width=1.5in]{fig/howmany\_scaf.png} |
|  |  | (b)**\includegraphics**[width=1.5in]{fig/tcam2\_mol\_541564.png} |
|  |  | \end{minipage} |
|  |  | (c)**\includegraphics**[width=3in]{fig/structure\_group\_C.png} |
|  |  | **\includegraphics**[width=0.5in]{fig/tcam2\_541564\_6scaf\_col.png} |
|  |  | **\caption**{ |
|  |  | (a) Distribution of number of fragments, out of nearly 6000 total, found in each molecule in the TCAMS dataset. (b) Compound $C$, chosen here as Molecule 2 (TCAMS Compound ID 541564, PubChem CID: 44531725) from previous figures. |
|  |  | (c) Compound $C$ has 6 fragments derived using the NCATS R-group tool, shown in legend at right. The scaffold pie plot shows the structure group of $C$, restricted here to only compounds that share two or more fragments with $C$. This ensures that all compounds sharing just a single heteroaromatic ring are not in the same group, as chemists expect, and also reduces the group size -- 80 related molecules for this particular compound $C$.} |
|  |  | **\label**{fig:strucgroup} |
|  |  | \end{figure} |
|  |  |  |
|  |  |  |
|  |  | Comparing the Frameworks and NCATS R-Group Tool for the TCAMS dataset using the above mentioned statistical metrics, we show the aggregate statistics over the entire dataset in **\fref**{statcomparetable}. |
|  |  | Comparing these methods for the TCAMS dataset, we show the aggregate statistics over the entire dataset in **\fref**{statcomparetable}. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | \begin{minipage}[c]{**\linewidth**} |
|  |  | **\vspace**{0pt} |
|  |  | **\centering** |
|  |  | **\scalebox**{.8}{ |
|  |  | (a)\begin{tabular}{|c|ccc|c|} |
|  |  | **\small** |
|  |  | **\vspace**{0pt} |
|  |  | **\centering** |
|  |  | **\scalebox**{0.7}{ |
|  |  | (a)\begin{tabular}{|c|ccc|c|} |
|  |  | **\hline** |
|  |  | {**\bf** Column} & {**\bf** P10} & {**\bf** Median} & {**\bf** P90} & {**\bf** Avg} **\\** |
|  |  | **\hline** |
| ... | ... | @@ -853,29 +801,58 @@ Comparing the Frameworks and NCATS R-Group Tool for the TCAMS dataset using the |
|  |  | **\hline** |
|  |  | \end{tabular} |
|  |  | } |
|  |  | (b)**\includegraphics**[width=2in]{fig/CP\_TCAMS\_GSKFW\_RGT.png} |
|  |  | **\scalebox**{0.7}{ |
|  |  | (b)\begin{tabular}{|c|ccc|c|} |
|  |  | **\hline** |
|  |  | {**\bf** Column} & {**\bf** P10} & {**\bf** Median} & {**\bf** P90} & {**\bf** Avg} **\\** |
|  |  | **\hline** |
|  |  | {**\bf** $C\_D$} & 3.00 & 17.00 & 88.00 & 33.11 **\\** |
|  |  | {**\bf** $C\_B$} & 83.50 & 888.00 & 2151.00 & 1047.93 **\\** |
|  |  | {**\bf** $CP$} & 0.00 & 0.02 & 0.17 & 0.07 **\\** |
|  |  | {**\bf** $PI\_D$} & 0.00 & 0.02 & 0.2 & 0.09 **\\** |
|  |  | {**\bf** $PI\_B$} & 0.98 & 1.00 & 1.00 & 0.98 **\\** |
|  |  | {**\bf** $PIU\_D$} & 0.00 & 0.00 & 0.02 & 0.02 **\\** |
|  |  | {**\bf** $PIU\_B$} & 0.79 & 0.98 & 1.00 & 0.91 **\\** |
|  |  | {**\bf** $Frag\_D$} & 1.00 & 1.00 & 1.00 & 1.00 **\\** |
|  |  | {**\bf** $Frag\_B$} & 2.00 & 5.00 & 11.00 & 5.93 **\\** |
|  |  | {**\bf** $FragEff\_D$} & 3.00 & 17.00 & 88.00 & 33.11 **\\** |
|  |  | {**\bf** $FragEff\_B$} & 21.25 & 150.80 & 454.00 & 211.80 **\\** |
|  |  | **\hline** |
|  |  | \end{tabular} |
|  |  | }**\\** |
|  |  | (c)**\includegraphics**[height=1.3in]{fig/CP\_TCAMS\_GSKFW\_RGT\_v2.png} |
|  |  | (d)**\includegraphics**[height=1.25in]{fig/CP\_TCAMS\_CLink\_RGT\_v3.png}**\\** |
|  |  |  |
|  |  | \end{minipage} |
|  |  | **\caption**{Statistics computed for 10th/90th percentile, median and average value for structure group size ($Ca$/$Cb$), common proportion, proportional information unique to A and B, number of fragments in and fragment efficiency of method A (frameworks) and B (NCATS R-group tool).} |
|  |  | **\caption**{(a) Statistics computed for 10th/90th percentile, median and average value for structure group size ($Ca$/$Cb$), common proportion, proportional information unique to $A$ and $B$, number of fragments in and fragment efficiency of method $A$ (frameworks) and $B$ (NCATS R-group tool). (b) Same comparison for $B$ (NCATS R-group tool) and $D$ (Complete-Linkage Clustering). (c)-(d) Histograms of Common Proportion for the datasets tabulated above in (a)-(b).} |
|  |  | **\label**{fig:statcomparetable} |
|  |  | \end{figure} |
|  |  |  |
|  |  |  |
|  |  | We compare the Common Proportion, Proportion of Information Unique (PIU) and Fragment Efficiency (FragEff) statistics for the Frameworks (method ``A'') and the NCATS R-group tool (method ``B'') for all the 13.5k compounds in TCAMS in **\fref**{statcompare}. This figure has PIU of methods $A$ and $B$ on the axes, and is sized by the ratio of their Fragment Efficiencies for all 13.5k molecules. |
|  |  | We compare the Common Proportion, Proportion of Information Unique (PIU) and Fragment Efficiency (FragEff) statistics for the NCATS R-group tool (method ``B'') with both the Frameworks (method ``A'') and the Complete Linkage Clusters (method ``D'') for all the 13.5k compounds in TCAMS in **\fref**{statcompare}. This figure has the PIU of these methods on the axes, and is sized by the ratio of their Fragment Efficiencies for all 13.5k molecules. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/statcompare\_frames\_RGtool\_transparent\_density.png} |
|  |  | **\caption**{Comparison of the Common Proportion, PIU and FragEff statistics (described in the text) between Frameworks (method ``{**\bf** A}'') and NCATS R-group Tool (``{**\bf** B}'') for all molecules in TCAMS. PIU of A and B are on the X and Y axes, and the points are colored by the log ratio of FragEff(B) to FragEff(A) and sized by Common Proportion.} |
|  |  | (a)**\includegraphics**[width=3.25in]{fig/statcompare\_frames\_RGtool\_transparent\_density\_v3.png} |
|  |  | (b)**\includegraphics**[width=2.5in]{fig/statcompare\_CLink\_RGtool\_transparent\_density\_v2.png} |
|  |  | **\caption**{Comparison of the Common Proportion, PIU and FragEff statistics (described in the text) between the NCATS R-group Tool (``{**\bf** B}''), |
|  |  | PIU on Y axis, |
|  |  | and (a) Frameworks (method ``{**\bf** A}'') |
|  |  | or (b) Complete Linkage Clustering (method ``{**\bf** D}''). |
|  |  | PIU for the methods being compared are on X and Y axes, and points are |
|  |  | colored by the log ratio of Fragment Efficiencies between the two methods being compared and sized by their Common Proportion.} |
|  |  | **\label**{fig:statcompare} |
|  |  | \end{figure} |
|  |  |  |
|  |  | We can make a few observations from this data and plot: |
|  |  | \begin{enumerate} |
|  |  | **\item** The two methods allow one to access different sets of molecules starting from any molecule in TCAMS - the average overlap in their coverage ($CP$) is 40**\%**, seen in Row 3 of **\fref**{statcomparetable}(a). The distribution of $CP$ seen in **\fref**{statcomparetable}(b) is bimodal with peaks at 0 and 55**\%**. The few molecules with almost complete overlap in coverage have high Common Proportion, low $PIU$ for both methods $A$ and $B$, and hence are represented by the large squares towards the bottom left of **\fref**{statcompare}. |
|  |  | **\item** On the average, frameworks connected to a compound add more unique information than NCATS scaffolds connected to the same compound - this is seen in the higher density near the X-axis in **\fref**{statcompare}, and higher numbers for $PIU\_A$ than $PIU\_B$ in **\fref**{statcomparetable}. |
|  |  |  |
|  |  | We can make a few observations from this data and these plots: |
|  |  | \begin{enumerate} |
|  |  | **\item** {**\bf** Clustering (method $D$) is able to access significantly fewer molecules} as structurally related analogs for most molecules in TCAMS, compared to the overlapping scaffold-based methods $A$ and $B$ (which were comparable). Most molecules have a low Common Proportion between clustering and the scaffold-based method, shown in **\fref**{statcompare}(d). Almost all the information (PIU) is contributed by the scaffold method ($B$) for most molecules, not surprising when we recall the high singleton rate from **\fref**{platypus} where the clustering contributes no information. Even allowing for fewer (one) fragments per molecule, the partitioning cluster method still had a lower Fragment Efficiency than any scaffold-based method. |
|  |  | **\item** The two scaffold-based methods ($A$: Frameworks and $B$: NCATS R-Group Tool) allow one to access different sets of molecules starting from any molecule in TCAMS - the average overlap in their coverage ($CP$) is 40**\%**, seen in Row 3 of **\fref**{statcomparetable}(a). The distribution of $CP$ seen in **\fref**{statcomparetable}(c) is bimodal with peaks at 0 and 55**\%**. The few molecules with almost complete overlap in coverage have high Common Proportion, low $PIU$ for both methods $A$ and $B$, and hence are represented by the large squares towards the bottom left of **\fref**{statcompare}(a). |
|  |  | **\item** On the average, frameworks connected to a compound add more unique information than NCATS scaffolds connected to the same compound - this is seen in the higher density near the X-axis in **\fref**{statcompare}(a), and higher numbers for $PIU\_A$ than $PIU\_B$ in **\fref**{statcomparetable}. |
|  |  | **\item** On the average, one can link to about twice as many molecules with the frameworks, as seen by comparing $C\_A$ to $C\_B$ in **\fref**{statcomparetable}(a); however, this is because on the average there are 6 times more frameworks ($Frag\_A$) than NCATS scaffolds ($Frag\_B$). *%, \ie NCATS scaffolds are three times more fragment-efficient for this dataset. %, so the fragment efficiency is actually 3 times greater for NCATS scaffolds. %One could also argue that many of the framework-only links (\eg variously decorated benzene rings) are not useful.* |
|  |  | **\item** The outliers are interesting. At one end, compounds in a rare tautomer are unified with the dominant one by the NCATS tool (high fragment efficiency), but left as singletons by the frameworks (low fragment efficiency). And compounds whose only link with other molecules would be a benzene ring or similar low complexity scaffold remain singletons with the R-group tool (lower fragment efficiency). |
|  |  | **\item** The outliers in **\fref**{statcompare}(a) are interesting. At one end, compounds in a rare tautomer are unified with the dominant one by the NCATS tool (high fragment efficiency), but left as singletons by the frameworks (low fragment efficiency). And compounds whose only link with other molecules would be a benzene ring or similar low complexity scaffold remain singletons with the R-group tool (lower fragment efficiency). |
|  |  | **\item** Singleton clusters are outliers in **\fref**{statcompare}(b), connected to other molecules via scaffolds but not clusters. And a minority of molecules are connected to others via clusters only, mostly those with low complexity linear scaffolds whose benzene rings were excluded from scaffold but not cluster analysis. |
|  |  | \end{enumerate} |
|  |  |  |
|  |  | **\subsection**{Statistical Basis of Structure-Activity Relationships (SAR)}**\label**{sec:statSAR} |
| ... | ... | @@ -898,10 +875,10 @@ We computed this NormCP measure for all compounds, which fit a normal distributi |
|  |  | \begin{figure} |
|  |  | (a)**\includegraphics**[width=4.5in]{fig/NormCP\_all.png}**\\** |
|  |  | (b)**\includegraphics**[width=4.5in]{fig/NormCP\_top200.png}**\\** |
|  |  | (c)**\includegraphics**[width=1.5in]{fig/NormCP=42\_5\_pIC50=8\_49\_CID524739.png} |
|  |  | (d)**\includegraphics**[width=1.5in]{fig/NormCP=0\_17\_pIC50=8\_22\_CID541941.png} |
|  |  | (e)**\includegraphics**[width=1.7in]{fig/NormCP=0\_16\_pIC50=7\_73\_CID531249.png} |
|  |  | **\caption**{(a) Normalized common proportion scores between structure (data-driven frameworks) and activity ($pIC50\_{3D7}$) for all 13.5k molecules in TCAMS. The histogram has 1000 bins and a log scale on the X-axis. The top 200 most active compounds ($pIC50\_{3D7}$ > 7.6) are marked in red, and occupy many of the outlier bars with high NormCP > 10. (b) Making the distribution of the top 200 active compounds clearer by giving them a separate scale, at the RHS. (c) One of the 23 compounds with NormCP = 42.5, Compound ID 524739. (d)--(e) two active compounds with low NormCP in the 0.16-0.17 range: (d) Compound ID 542941 and (e) Compound ID 531249. For both compounds, they alone are highly active among a structure group of 96 compounds, raising the probability that these compounds are flagpoles and not worth following up as hits.} |
|  |  | (c)**\includegraphics**[width=1.8in]{fig/NormCP=42\_5\_pIC50=8\_49\_CID524739\_v2.png} |
|  |  | (d)**\includegraphics**[width=1.3in]{fig/NormCP=0\_17\_pIC50=8\_22\_CID541941\_v2.png} |
|  |  | (e)**\includegraphics**[width=1.9in]{fig/NormCP=0\_16\_pIC50=7\_73\_CID531249\_v2.png} |
|  |  | **\caption**{(a) Normalized common proportion scores between structure (data-driven frameworks) and activity ($pIC50\_{3D7}$) for all 13.5k molecules in TCAMS. The histogram has 1000 bins and a log scale on the X-axis. The top 200 most active compounds ($pIC50\_{3D7}$ > 7.6) are marked in red, and occupy many of the outlier bars with high NormCP > 10. (b) Making the distribution of the top 200 active compounds clearer by giving them a separate scale, at the RHS. (c) One of the 23 compounds with NormCP = 42.5, Compound ID 524739. (d)--(e) two active compounds with low NormCP in the 0.16-0.17 range: (d) Compound ID 541941 and (e) Compound ID 531249. For both compounds, they alone are highly active among a structure group of 96 compounds, raising the probability that these compounds are flagpoles and not worth following up as hits.} |
|  |  | **\label**{fig:NormCP} |
|  |  | \end{figure} |
|  |  |  |
| ... | ... | @@ -929,7 +906,7 @@ for the comparison of different scaffold-based analysis schemes. |
|  |  | Also, the overlapping scaffold approach described here |
|  |  | avoids the phenomenon of similar molecules being arbitrarily assigned to |
|  |  | exclusive clusters, which affects partitioning-based methods such as |
|  |  | complete-linkage clustering as shown in **\fref**{clusterlanes}. |
|  |  | complete-linkage clustering.*% as shown in \fref{clusterlanes}.* |
|  |  | Instead, using overlapping scaffolds ensures that |
|  |  | molecules that differ only in decorations off a shared scaffold |
|  |  | will be considered within the same group. |
| ... | ... | @@ -980,7 +957,9 @@ combines data and intuitive visualizations to help scientists combat such biases |
|  |  | \begin{acknowledgement} |
|  |  | The GSK authors thank Subhas Chakravorty, Neysa Nevins, Ami Lakdawala Shah, |
|  |  | Eric Manas, Todd Graybill, Stan Martens, Mike Ouellette, Tony Jurewicz, Rob |
|  |  | Young, Ken Lind and Jeff Messer for valuable feedback and suggestions while developing the method and visualizations. We dedicate this work to the memory of Christopher Louer, our colleague and cheminformatics wizard at GSK who always encouraged us to innovate in order to help chemists. |
|  |  | Young, Ken Lind and Jeff Messer for valuable feedback and suggestions while developing the method and visualizations. We dedicate this work to the memory of Christopher Louer, our colleague and cheminformatics wizard at GSK who always encouraged us to innovate in order to help chemists. |
|  |  |  |
|  |  | The authors declare the following sources of funding: author DB was funded by GSK while undertaking this work and by Janssen Pharmaceuticals in the later stages of writing it up. Authors CK, PB, JB, ZH and GS were funded by GSK. Authors TP, DTN and AJ were funded by NIH. Author RG was funded by NIH while undertaking this work and by Vertex Pharmaceuticals in the later stages of writing it up. |
|  |  | \end{acknowledgement} |
|  |  |  |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
| ... | ... | @@ -996,6 +975,13 @@ Supplementary material is available online for this article. |
|  |  | *%% Notice that the class file automatically sets \bibliographystyle* |
|  |  | *%% and also names the section correctly.* |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | *%%\bibliographystyle{achemso}* |
|  |  | **\bibliography**{bibliography} |
|  |  |  |
|  |  | **\newpage** |
|  |  | *%% w=3.25in, h=1.75in, scaled 1.5x: 3.25+1.625, 1.75+0.875: w=4.875, h=2.625* |
|  |  | \begin{figure} |
|  |  | *% \includegraphics[width=4.875in,height=2.625in]{fig/TOC\_image\_v2.png}* |
|  |  | **\includegraphics**[width=3.25in,height=1.75in]{fig/TOC\_image\_v4.png} |
|  |  | **\caption**{For Table of Contents Only} |
|  |  | \end{figure} |
|  |  | \end{document} |

[**SI/SI.tex**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/28575ec708dd8bb0f33062c15d36b8d3781e911b...5164c3886656ccfe8cde15e2d4b514559ddb59a0#3cc0701a2668b990684d9d4f9f430f46f63b24aa)

[View file @5164c38](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/5164c3886656ccfe8cde15e2d4b514559ddb59a0/SI/SI.tex)

|  |  |  |
| --- | --- | --- |
| ... | ... | @@ -3,6 +3,13 @@ |
|  |  | **\usepackage**{ctable} |
|  |  |  |
|  |  | **\usepackage**{etoolbox} |
|  |  | **\usepackage**{listings} |
|  |  | **\lstset**{ *%* |
|  |  | language=R, *% choose the language of the code* |
|  |  | numbers=none, *% where to put the line-numbers* |
|  |  | numberstyle=**\footnotesize**, *% the size of the fonts that are used for the line-numbers* |
|  |  | basicstyle=**\tiny** *% the size of the fonts that are used for the line-numbers* |
|  |  | } |
|  |  | **\makeatletter** |
|  |  | **\patchcmd**{**\@**verbatim} |
|  |  | {**\verbatim**@font} |
| ... | ... | @@ -13,6 +20,10 @@ |
|  |  | **\newcommand\*\fref**[1]{Figure~**\ref**{fig:#1}} |
|  |  | **\newcommand\*\tref**[1]{Table~**\ref**{table:#1}} |
|  |  | **\newcommand\*\sref**[1]{Section~**\ref**{sec:#1}} |
|  |  | **\newcommand\*\eg**{e.g.,~} |
|  |  | *% \newcommand\*\ie{i.e.,~}* |
|  |  | **\newcommand\*\vs**{vs.~} |
|  |  | **\newcommand\*\viz**{viz.~} |
|  |  |  |
|  |  | *% failed attempt to use \citet in SI - used successfully in main paper* |
|  |  | *% will hard-code the 2 refs where needed* |
| ... | ... | @@ -186,7 +197,7 @@ Bemis-Murcko-like and Recap fragments can be built from datasets within the GSK |
|  |  | | awk '{print $2, $1}' | uniq -D -f 1 |
|  |  | > chemblntd\_gsk2\_spc\_frames\_shared.txt |
|  |  | \end{verbatim} |
|  |  | **\item** As an optional step we did not implement, a scoring function may be computed from the aggregate activity of a fragment and used to triage frameworks as was done for scaffolds from the NCATS R-group tool. |
|  |  | **\item** Optionally (not implemented) frameworks may be triaged by aggregate activity as was done for scaffolds from the NCATS R-group tool. |
|  |  |  |
|  |  | \end{itemize} |
|  |  |  |
| ... | ... | @@ -281,8 +292,8 @@ within Spotfire: |
|  |  | *%% outside Spotfire (\eg in a script that integrates the data); and using* |
|  |  | *%% various add-ins for Spotfire such as the internally developed GSK SAR* |
|  |  | *%% Toolkit and commercial tools from Discngine SA.* |
|  |  |  |
|  |  | **\newpage** |
|  |  |  |
|  |  | **\newpage** |
|  |  |  |
|  |  | **\section**{Spotfire alternate visualizations and usability tips} |
|  |  | **\label**{sec:spotviz-trick} |
| ... | ... | @@ -301,137 +312,248 @@ A few Spotfire techniques were employed to make the users' experience with our S |
|  |  | \end{itemize} |
|  |  |  |
|  |  | **\newpage** |
|  |  | **\section**{Use Case: Scaffold Walking Further Examples}**\label**{sec:scafwalk\_eg} |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\centering** |
|  |  | **\includegraphics**[width=6in]{../fig/mol2\_RGtool\_scafpie\_plot\_v4.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | **\includegraphics**[width=5.5in]{../fig/mol2\_RGtool\_scafpie\_struc\_v2.png}**\\** |
|  |  | **\caption**{Related Molecules scaffold pies visualization for Molecule 3 (TCAMS Compound ID: 533945; PubChem CID: 44531163). Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the figure and also discussed in the text.} |
|  |  | **\label**{fig:scafwalk2} |
|  |  | \end{figure} |
|  |  |  |
|  |  | Molecule 3 in **\fref**{scafwalk2} is another case where we might want to optimize the physical and chemical properties of the molecule without sacrificing activity or increasing promiscuity. Since solubility has been shown to decrease with number of aromatic rings independent of lipophilicity**\cite**{Hill2010}, walks that remove one or more of the three fused rings might be beneficial. By exploring the Related Molecules, we observe the SAR for three scaffolds: quinazolines (**\#**305, brown), indazoloquinazolines (**\#**574, blue) and indazoles (**\#**3822, pink). We observe from the plot a few more molecules containing all three scaffolds (tricolored pies, **\ie** exact analogs of the parent molecule); all of these are less active than the parent. The indazoles when they occur alone (pink circles) are far less active than the parent, suggesting they do not contribute significantly to activity and may be substituted. Lastly the quinazolines (brown) include several analogs that are more active and also less promiscuous than the parent. Drilling down into these structures, we observe several that contain only two aromatic rings (**\ie** no more are either fused or attached); these provide novel, active and ligand-efficient templates on which to build new analogs with enhanced solubility or other properties. Suggestions for which analogs to make can often be obtained by examining the SAR -- for example, disparate aliphatic and aromatic analogs at two adjacent positions on the quinazoline phenyl are active, which suggests hybridizing them or designing further analogs substituted at these positions. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\centering** |
|  |  | **\includegraphics**[width=5.5in]{../fig/mol3\_RGtool\_scafpie\_iter\_plot\_v2.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | **\includegraphics**[width=5in]{../fig/mol3\_RGtool\_scafpie\_iter\_struc\_v2.png} |
|  |  | **\caption**{Iterative Related Molecules scaffold pies visualization for Molecule 4 (TCAMS Compound ID: 541531, PubChem CID: 44531903), a scaffold hop from Molecule 3 shown in **\fref**{scafwalk2}. Each pie here is one related molecule, and each pie sector and color is a scaffold that it shares with the parent molecule. The star symbol is added to show the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (in this case towards the +X and -Y axes). Insights derived from the plot are highlighted in the figure and also discussed in the text.} |
|  |  | **\label**{fig:scafwalk3} |
|  |  | \end{figure} |
|  |  |  |
|  |  | Another intriguing result is seen by observing a new tricyclic series that shares the quinazoline but adds an indole instead of indazole as the third fused ring. This alternative tricyclic template, while it may not confer solubility advantages, opens up a new area of chemical space. By iteratively seeking the Related Molecules for this new hit as shown in **\fref**{scafwalk3}, we observed that most of the active quinazoline analogs have this new tricyclic scaffold (**\#**1824, tan wedges in tricolored pies) and relatively few contain only quinazolines (**\#**305, brown circles). Also, indoles by themselves (**\#**1188, blue circles) do not much better than indazoles, so this is a new synergistic effect discovered by scaffold walking from the original hit. |
|  |  |  |
|  |  |  |
|  |  | **\newpage** |
|  |  |  |
|  |  | **\section**{Qualitative Comparison of Scaffold-Generation Methods and Clustering}**\label**{sec:qualcomp} |
|  |  | {**\bf** Complete-Linkage Clustering}: As shown in **\fref**{clusterlanes}, the defining feature of a partitioning clustering is that every molecule maps to one and only one cluster. Thus if a chemotype is broken up among two or more clusters, using the cluster ID to map Related Molecules can retrieve only neighbors from the same cluster, ignoring the other cluster. This is not ideal for purposes of the visualization and navigation method presented here, as arbitrary neighbors would be excluded depending on how the clustering is defined. Thus we do not advocate the use of clustering, unless it is a fuzzy clustering where all meaningful class memberships a molecule might have are considered. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{../fig/clusterlanes\_v3.png} |
|  |  | **\caption**{Illustrating one problem with clustering: bifurcation of related molecules. When two molecules of the same chemotype differing by a halogen (TCAMS IDs: 79271 and 79711) are split across Complete Linkage Clusters, searches of cluster neighbors for one molecule (**\eg** IDs 540816 and 528977 shown within Cluster 3) do not find its analogs in the other cluster (**\eg** 540640 and 528006 in Cluster 10), **\ie** the two related clusters are not linked.} |
|  |  | **\label**{fig:clusterlanes} |
|  |  | \end{figure} |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  | {**\bf** NCATS R-group tool}: As opposed to the clustering method, |
|  |  | if any two molecules share a common substructure that meets the standards required of a scaffold by the NCATS method (**\eg** being bordered by rings on each end), then those molecules will be found to contain that shared substructure as a scaffold and their activities will be used to compute aggregate properties for it. |
|  |  |  |
|  |  | {**\bf** Other Scaffold Generation Methods}: Even though another scaffold generation method (represented here by molecular frameworks as implemented in **\cite**{Harper2004DDclus}) differed in its implementation details and produced different numbers of scaffolds for the same molecule, it was roughly equivalent in a qualitative sense with regard to the insights obtained during Scaffold Walking. Due to substantial overlap between sets of scaffolds, ring systems responsible for activity of a molecule were generally revealed by either method. However, there were cases where the Frameworks revealed negative information about a fragment being not important for activity that is also useful for a drug discovery scientist. For example, in **\fref**{frameswalk} a substructure is highlighted that is on the aggregate inactive and could be removed or substituted. This insight is not available from SSSR-based scaffolding methods such as the NCATS R-group tool since they don't define or find that fragment as a scaffold. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=5in]{../fig/mol1\_frames\_scafpie\_v2.png} |
|  |  | **\caption**{Using Frameworks with the Scaffold Pies visualization. One framework is highlighted that has no equivalent in the NCATS scaffolds, but is shown to reduce activity as related molecules containing it are less active than the parent molecule. The star symbol shows the location of the parent molecule in this Related Molecules plot, and the compass device at the origin shows the direction of favorable properties (+X and +Y axes).} |
|  |  | **\label**{fig:frameswalk} |
|  |  | \end{figure} |
|  |  |  |
|  |  |  |
|  |  | To summarize, both multiple-scaffold decomposition methods considered in this study, **\ie** NCATS R-group Tool and Frameworks give comparable insights when exploring the TCAMS dataset, with some differences stemming from individual substructures that are considered shared scaffolds or not by the individual methods. |
|  |  | *%We now explore these overlaps, similarities and differences in the aggregate using the statistical methods described earlier in \sref{statmethod}.* |
|  |  |  |
|  |  | **\newpage** |
|  |  | **\section**{Statistical analysis: structure group concept illustrated} |
|  |  | **\label**{sec:strucgroup} |
|  |  | As an illustration, **\fref**{strucgroup} (a)-(c) shows the structure group of the compound $C$ with Compound ID 541564 (PubChem CID: 44531725). |
|  |  | \begin{figure} |
|  |  | **\centering** |
|  |  | *% \begin{minipage}[b][0.2\textheight][s]{0.7\textwidth}* |
|  |  | *% \centering* |
|  |  | (a)**\includegraphics**[width=1.5in]{../fig/howmany\_scaf.png} |
|  |  | (b)**\includegraphics**[width=2in]{../fig/tcam2\_mol\_541564\_v2.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | *% \end{minipage}* |
|  |  | (c)**\includegraphics**[width=4in]{../fig/structure\_group\_C\_v2.png}**\\** |
|  |  | **\includegraphics**[width=4in]{../fig/tcam2\_541564\_6scaf\_row\_v2.png} |
|  |  | **\caption**{ |
|  |  | (a) Distribution of number of scaffolds in TCAMS molecules. (b) Compound $C$, chosen as Molecule 2 (TCAMS Compound ID 541564, PubChem CID: 44531725) from previous figures in the manuscript. |
|  |  | (c) Compound $C$ has 6 fragments derived using the NCATS R-group tool, shown in bottom legend. The scaffold pie plot shows the structure group of $C$, restricted for illustration to only the 80 compounds that share two or more fragments with $C$.} *%This ensures that all compounds sharing just a single heteroaromatic ring are not in the same group, as chemists expect, and also reduces the group size -- 80 related molecules for this particular compound $C$.}* |
|  |  | **\label**{fig:strucgroup} |
|  |  | \end{figure} |
|  |  |  |
|  |  | **\newpage** |
|  |  |  |
|  |  | **\section**{R code for statistical comparison of scaffold generation methods} |
|  |  | **\label**{sec:statcode-scafcomp} |
|  |  |  |
|  |  | The code below expects two files, DDframes.txt (from framework clustering, Method A in the results) and RGD.txt (NCATS R-group tool decomposition, Method B). They both have a column named Compound**\\_**ID. DDframes.txt has a column StrucUniqueID which is computed for example using the DenseRank function in Spotfire, assigning a unique numerical ID to each fragment having different Canonical SMILES. RGD.txt has a column SCAFFOLD**\\_**ID that is already numerically distinct for each separate scaffold. The code can be generalized for any pair of methods as long as the map from numerical compound IDs to scaffold IDs exists in the input data. |
|  |  | The code below expects three files, DDframes.txt (from framework clustering, Method $A$ in the results), RGD.txt (NCATS R-group tool decomposition, Method $B$), and CLinkClusters.txt (Complete Linkage Clustering, Method $D$). All three files have a column named Compound**\\_**ID. DDframes.txt has a column StrucUniqueID which is computed for example using the DenseRank function in Spotfire, assigning a unique numerical ID to each fragment having different Canonical SMILES. RGD.txt has a column SCAFFOLD**\\_**ID that is already numerically distinct for each separate scaffold. And CLinkClusters.txt contains the Cluster number for each compound in column CLink. The computation, summarization and graphing of our statistics is parameterized into functions to allow Methods $A$ and $D$ to be compared in turns to Method $B$. The code can be generalized to compare any pair of methods as long as the map from numerical compound IDs to scaffold IDs exists in the input data. |
|  |  |  |
|  |  | \begin{verbatim} |
|  |  | *% \begin{verbatim}* |
|  |  | {**\tiny** |
|  |  | \begin{lstlisting} |
|  |  | #Note: To do the calculations comparing 2 fragmentation methods |
|  |  | #took about 10 minutes of computer time. |
|  |  |  |
|  |  | #Set working directory to location of the data files |
|  |  | setwd("C:**\\**Work**\\**Consulting**\\**MDR**\\**Other MDR Issues**\\**CSC**\\**FragmentOntologies") |
|  |  | #setwd("C:**\\**Work**\\**Consulting**\\**MDR**\\**Other MDR Issues**\\**CSC**\\**FragmentOntologies") # Joe's folder |
|  |  | setwd("C:/Work/git/NIH/scaffoldanalytics/stats") # Deepak's folder |
|  |  | ##################### |
|  |  | rm(list = ls()) #Erase anything in R's working memory |
|  |  |  |
|  |  | DDData <- read.table("DDframes.txt", header = T, sep = "**\t**") |
|  |  | RGData <- read.table("RGD.txt", header = T, sep = "**\t**") |
|  |  | CLData <- read.table("CLinkClusters.txt", header = T, sep = "**\t**") |
|  |  |  |
|  |  | #Create a list of compounds shared between two datasets. |
|  |  | #Scores will not make sense if we include non-common compounds |
|  |  |  |
|  |  | DDCompounds <- unique(DDData$COMPOUND\_ID**)** |
|  |  | RGCompounds <**-** unique**(**RGData$COMPOUND\_ID) |
|  |  | CLCompounds <- unique(CLData$COMPOUND\_ID**)** |
|  |  |  |
|  |  | # Compounds <**-** intersect**(**DDCompounds, RGCompounds**)** |
|  |  | # if comparing 3 methods pairwise, make sure we use the same set of compounds |
|  |  | Compounds <**-** intersect**(**intersect**(**RGCompounds, DDCompounds**)**, CLCompounds**)** |
|  |  |  |
|  |  | Compounds <- intersect(DDCompounds, RGCompounds) |
|  |  | # Compounds <**-** as.numeric**(**as.character**(**intersect**(**DDCompounds, RGCompounds**)))** |
|  |  | N <**-** length**(**Compounds**)** |
|  |  |  |
|  |  | #Calculate PI's and common proportions for each compound |
|  |  |  |
|  |  | #Proportions by compound |
|  |  |  |
|  |  | PropByCompound <- data.frame(CompoundID = rep(NA, N), |
|  |  | FragA = rep(NA, N), |
|  |  | FragB = rep(NA, N), |
|  |  | Ca = rep(NA, N), |
|  |  | Cb = rep(NA, N), |
|  |  | IntAB = rep(NA, N), |
|  |  | UnionAB = rep(NA, N), |
|  |  | CommonProp = rep(NA, N), |
|  |  | PIa = rep(NA, N), |
|  |  | PIb = rep(NA, N), |
|  |  | PIaU = rep(NA, N), |
|  |  | PIbU = rep(NA, N), |
|  |  | FragEffA = rep(NA, N), |
|  |  | FragEffB = rep(NA, N)) |
|  |  |  |
|  |  | for (index in 1:N){ |
|  |  | statcompare <**-** function**(**ScafSetA, ScafSetB, |
|  |  | ScafColA**=**"StrucUniqueID", ScafColB**=**"SCAFFOLD\_ID", |
|  |  | CompoundSet**)** { |
|  |  | N <**-** length**(**CompoundSet**)** |
|  |  |  |
|  |  | PropByCompound$CompoundID**[**index**]** <**-** Compounds**[**index**]** |
|  |  | PropByCompound <**-** data.frame**(**CompoundID **=** rep**(**NA, N**)**, |
|  |  | FragA **=** rep**(**NA, N**)**, FragB **=** rep**(**NA, N**)**, |
|  |  | Ca **=** rep**(**NA, N**)**, Cb **=** rep**(**NA, N**)**, |
|  |  | IntAB **=** rep**(**NA, N**)**, UnionAB **=** rep**(**NA, N**)**, |
|  |  | CommonProp **=** rep**(**NA, N**)**, |
|  |  | PIa **=** rep**(**NA, N**)**, PIb **=** rep**(**NA, N**)**, |
|  |  | PIaU **=** rep**(**NA, N**)**, PIbU **=** rep**(**NA, N**)**, |
|  |  | FragEffA **=** rep**(**NA, N**)**, FragEffB **=** rep**(**NA, N**))** |
|  |  |  |
|  |  | for **(**index in 1:N**)**{ |
|  |  |  |
|  |  | MethodABelongsTo <**-** DDData**[**DDData$COMPOUND\_ID == Compounds[index], |
|  |  | "StrucUniqueID"] |
|  |  | MethodBBelongsTo <- RGData[RGData$COMPOUND\_ID **==** Compounds**[**index**]**, |
|  |  | "SCAFFOLD\_ID"**]** |
|  |  |  |
|  |  | PropByCompound$FragA[index] <- length(unique(MethodABelongsTo)) |
|  |  | PropByCompound$FragB**[**index**]** <**-** length**(**unique**(**MethodBBelongsTo**))** |
|  |  | PropByCompound$CompoundID[index] <- CompoundSet[index] |
|  |  |  |
|  |  | MethodACompoundCluster <**-** unique**(**DDData**[**DDData$StrucUniqueID *%in% MethodABelongsTo,* |
|  |  | "COMPOUND\_ID"]) |
|  |  | MethodBCompoundCluster <- unique(RGData[RGData$SCAFFOLD\_ID *%in% MethodBBelongsTo,* |
|  |  | "COMPOUND\_ID"**])** |
|  |  | MethodABelongsTo <- ScafSetA[ScafSetA$COMPOUND\_ID **==** Compounds**[**index**]**, |
|  |  | ScafColA**]** # StrucUniqueID or CLink |
|  |  |  |
|  |  | MethodBBelongsTo <**-** ScafSetB**[**ScafSetB$COMPOUND\_ID == Compounds[index], |
|  |  | ScafColB] # SCAFFOLD\_ID |
|  |  |  |
|  |  | PropByCompound$FragA**[**index**]** <**-** length**(**unique**(**MethodABelongsTo**))** |
|  |  | PropByCompound$FragB[index] <- length(unique(MethodBBelongsTo)) |
|  |  |  |
|  |  | MethodACompoundCluster <- unique(ScafSetA[ScafSetA[,ScafColA] *%in% MethodABelongsTo,* |
|  |  | "COMPOUND\_ID"]) |
|  |  |  |
|  |  | PropByCompound$Ca[index] <- length(MethodACompoundCluster) |
|  |  | PropByCompound$Cb**[**index**]** <**-** length**(**MethodBCompoundCluster**)** |
|  |  | PropByCompound$IntAB[index] <- length(intersect(MethodACompoundCluster, |
|  |  | MethodBCompoundCluster)) |
|  |  | PropByCompound$UnionAB**[**index**]** <**-** length**(**union**(**MethodACompoundCluster, |
|  |  | MethodBCompoundCluster <- unique(ScafSetB[ScafSetB[,ScafColB] *%in% MethodBBelongsTo,* |
|  |  | "COMPOUND\_ID"]) |
|  |  |  |
|  |  | PropByCompound$Ca**[**index**]** <**-** length**(**MethodACompoundCluster**)** |
|  |  | PropByCompound$Cb[index] <- length(MethodBCompoundCluster) |
|  |  | PropByCompound$IntAB**[**index**]** <**-** length**(**intersect**(**MethodACompoundCluster, |
|  |  | MethodBCompoundCluster**))** |
|  |  | PropByCompound$CommonProp[index] <- PropByCompound$IntAB**[**index**]/** |
|  |  | PropByCompound$UnionAB[index] |
|  |  | PropByCompound$PIa**[**index**]** <**-** PropByCompound$Ca[index]/PropByCompound$UnionAB**[**index**]** |
|  |  | PropByCompound$PIb[index] <- PropByCompound$Cb**[**index**]/**PropByCompound$UnionAB[index] |
|  |  | PropByCompound$PIaU**[**index**]** <**-** 1 **-** PropByCompound$Cb[index]/PropByCompound$UnionAB**[**index**]** |
|  |  | PropByCompound$PIbU[index] <- 1 - PropByCompound$Ca**[**index**]/**PropByCompound$UnionAB[index] |
|  |  | PropByCompound$FragEffA**[**index**]** <**-** PropByCompound$Ca[index]/PropByCompound$FragA**[**index**]** |
|  |  | PropByCompound$FragEffB[index] <- PropByCompound$Cb**[**index**]/**PropByCompound$FragB[index] |
|  |  |  |
|  |  | } |
|  |  | PropByCompound$UnionAB[index] <- length(union(MethodACompoundCluster, |
|  |  | MethodBCompoundCluster)) |
|  |  | PropByCompound$CommonProp**[**index**]** <**-** PropByCompound$IntAB[index]/PropByCompound$UnionAB**[**index**]** |
|  |  | PropByCompound$PIa[index] <- PropByCompound$Ca**[**index**]/**PropByCompound$UnionAB[index] |
|  |  | PropByCompound$PIb**[**index**]** <**-** PropByCompound$Cb[index]/PropByCompound$UnionAB**[**index**]** |
|  |  | PropByCompound$PIaU[index] <- 1 - PropByCompound$Cb**[**index**]/**PropByCompound$UnionAB[index] |
|  |  | PropByCompound$PIbU**[**index**]** <**-** 1 **-** PropByCompound$Ca[index]/PropByCompound$UnionAB**[**index**]** |
|  |  | PropByCompound$FragEffA[index] <- PropByCompound$Ca**[**index**]/**PropByCompound$FragA[index] |
|  |  | PropByCompound$FragEffB**[**index**]** <**-** PropByCompound$Cb[index]/PropByCompound$FragB**[**index**]** |
|  |  | } # for index |
|  |  | return**(**PropByCompound**)** |
|  |  | } # function statcompare |
|  |  |  |
|  |  | # call function **-** A is FW, B is RGT, D is CLink |
|  |  | PropByCompound\_AB <**-** statcompare**(**ScafSetA **=** DDData , ScafSetB **=** RGData, |
|  |  | ScafColA**=**"StrucUniqueID", ScafColB**=**"SCAFFOLD\_ID", |
|  |  | CompoundSet **=** Compounds**)** |
|  |  | PropByCompound\_DB <**-** statcompare**(**ScafSetA **=** CLData , ScafSetB **=** RGData, |
|  |  | ScafColA**=**"CLink", ScafColB**=**"SCAFFOLD\_ID", |
|  |  | CompoundSet **=** Compounds**)** |
|  |  |  |
|  |  | #Create output **--** averages, quantiles, and histograms |
|  |  | summarize\_prop <**-** function**(**PropByCompound**)** { |
|  |  | ACP <**-** mean**(**PropByCompound$CommonProp, na.rm = T) |
|  |  | APIa <- mean(PropByCompound$PIa, na.rm **=** T**)** |
|  |  | APIb <**-** mean**(**PropByCompound$PIb, na.rm = T) |
|  |  | APIaU <- mean(PropByCompound$PIaU, na.rm **=** T**)** |
|  |  | APIbU <**-** mean**(**PropByCompound$PIbU, na.rm = T) |
|  |  | AFragA <- mean(PropByCompound$FragA, na.rm **=** T**)** |
|  |  | AFragB <**-** mean**(**PropByCompound$FragB, na.rm = T) |
|  |  | AFragEffA <- mean(PropByCompound$FragEffA, na.rm **=** T**)** |
|  |  | AFragEffB <**-** mean**(**PropByCompound$FragEffB, na.rm = T) |
|  |  | ACa <- mean(PropByCompound$Ca, na.rm **=** T**)** |
|  |  | ACb <**-** mean**(**PropByCompound$Cb, na.rm = T) |
|  |  |  |
|  |  | CP90 <- quantile(PropByCompound$CommonProp, c**(**0.1,0.5,0.9**))** |
|  |  | PIa90 <**-** quantile**(**PropByCompound$PIa, na.rm = T , c(0.1,0.5,0.9)) |
|  |  | PIb90 <- quantile(PropByCompound$PIb, na.rm **=** T , c**(**0.1,0.5,0.9**))** |
|  |  | PIaU90 <**-** quantile**(**PropByCompound$PIaU, na.rm = T , c(0.1,0.5,0.9)) |
|  |  | PIbU90 <- quantile(PropByCompound$PIbU, na.rm **=** T, c**(**0.1,0.5,0.9**))** |
|  |  | FragA90 <**-** quantile**(**PropByCompound$FragA, na.rm = T, c(0.1,0.5,0.9)) |
|  |  | FragB90 <- quantile(PropByCompound$FragB, na.rm **=** T, c**(**0.1,0.5,0.9**))** |
|  |  | FragEffA90 <**-** quantile**(**PropByCompound$FragEffA, na.rm = T, c(0.1,0.5,0.9)) |
|  |  | FragEffB90 <- quantile(PropByCompound$FragEffB, na.rm **=** T, c**(**0.1,0.5,0.9**))** |
|  |  | Ca90 <**-** quantile**(**PropByCompound$Ca, na.rm = T, c(0.1,0.5,0.9)) |
|  |  | Cb90 <- quantile(PropByCompound$Cb, na.rm **=** T, c**(**0.1,0.5,0.9**))** |
|  |  |  |
|  |  | ret <**-** list**(**ACP**=**ACP, CP90**=**CP90, |
|  |  | APIa**=**APIa, PIa90**=**PIa90, APIb**=**APIb, PIb90**=**PIb90, |
|  |  | APIaU**=**APIaU, PIaU90**=**PIaU90, APIbU**=**APIbU, PIbU90**=**PIbU90, |
|  |  | AFragA**=**AFragA, FragA90**=**FragA90, AFragB**=**AFragB, FragB90**=**FragB90, |
|  |  | AFragEffA**=**AFragEffA, FragEffA90**=**FragEffA90, AFragEffB**=**AFragEffB, FragEffB90**=**FragEffB90, |
|  |  | ACa**=**ACa, Ca90**=**Ca90, ACb**=**ACb, Cb90**=**Cb90**)** |
|  |  |  |
|  |  | return**(**ret**)** |
|  |  | } |
|  |  |  |
|  |  | ACP <- mean(PropByCompound$CommonProp, na.rm **=** T**)** |
|  |  | APIaU <**-** mean**(**PropByCompound$PIaU, na.rm = T) |
|  |  | APIbU <- mean(PropByCompound$PIbU, na.rm **=** T**)** |
|  |  | AFragEffA <**-** mean**(**PropByCompound$FragEffA, na.rm = T) |
|  |  | AFragEffB <- mean(PropByCompound$FragEffB, na.rm **=** T**)** |
|  |  |  |
|  |  | CP95 <**-** quantile**(**PropByCompound$CommonProp, c(0.05,0.25,0.5,0.75, 0.95)) |
|  |  | PIaU95 <- quantile(PropByCompound$PIaU, na.rm **=** T, c**(**0.05,0.25,0.5,0.75, 0.95**))** |
|  |  | PIbU95 <**-** quantile**(**PropByCompound$PIbU, na.rm = T, c(0.05,0.25,0.5,0.75, 0.95)) |
|  |  | FragEffA95 <- quantile(PropByCompound$FragEffA, na.rm **=** T, c**(**0.05,0.25,0.5,0.75, 0.95**))** |
|  |  | FragEffB95 <**-** quantile**(**PropByCompound$FragEffB, na.rm = T, c(0.05,0.25,0.5,0.75, 0.95)) |
|  |  |  |
|  |  | ACP |
|  |  | APIaU |
|  |  | APIbU |
|  |  | AFragEffA |
|  |  | AFragEffB |
|  |  |  |
|  |  | CP95 |
|  |  | PIaU95 |
|  |  | PIbU95 |
|  |  | AFragEffA95 |
|  |  | AFragEffB95 |
|  |  |  |
|  |  | write.table(PropByCompound,"PropByCompound.txt",sep="**\t**",row.name=F,col.name=T) |
|  |  | ######### create summaries ####### |
|  |  | SumProp\_AB <**-** summarize\_prop**(**PropByCompound\_AB**)** |
|  |  | SumProp\_DB <**-** summarize\_prop**(**PropByCompound\_DB**)** |
|  |  |  |
|  |  | # write output to CSV |
|  |  | write.table**(**PropByCompound\_AB,"PropByCompound\_FW\_RGD.txt",sep**=**"\t",row.name**=**F,col.name**=**T**)** |
|  |  | write.table**(**PropByCompound\_DB,"PropByCompound\_Cluster\_RGD.txt",sep**=**"\t",row.name**=**F,col.name**=**T**)** |
|  |  | ######################################################################################### |
|  |  |  |
|  |  | ################# Plot ##################### |
|  |  | attach(PropByCompound) |
|  |  |  |
|  |  | hist(FragA,main="FragA") |
|  |  | hist(FragB,main="FragB") |
|  |  | hist(CommonProp,main="CommonProp") |
|  |  |  |
|  |  | plot(CommonProp~UnionAB) |
|  |  | plot(CommonProp~IntAB) |
|  |  | plot(Ca~Cb) |
|  |  |  |
|  |  | plot(UnionAB, IntAB) |
|  |  | plot(FragA + FragB, CommonProp) |
|  |  |  |
|  |  | detach(PropByCompound) |
|  |  | createPlots <**-** function**(**PropByCompound**)** { |
|  |  | attach**(**PropByCompound**)** |
|  |  |  |
|  |  | hist**(**FragA,main**=**"FragA"**)** |
|  |  | hist**(**FragB,main**=**"FragB"**)** |
|  |  | hist**(**CommonProp,main**=**"CommonProp"**)** |
|  |  |  |
|  |  | plot**(**CommonProp~UnionAB**)** |
|  |  | plot**(**CommonProp~IntAB**)** |
|  |  | plot**(**Ca~Cb**)** |
|  |  |  |
|  |  | plot**(**UnionAB, IntAB**)** |
|  |  | plot**(**FragA **+** FragB, CommonProp**)** |
|  |  |  |
|  |  | detach**(**PropByCompound**)** |
|  |  | } |
|  |  |  |
|  |  | \end{verbatim} |
|  |  | createPlots**(**PropByCompound\_AB**)** |
|  |  | createPlots**(**PropByCompound\_DB**)** |
|  |  | \end{lstlisting} |
|  |  | } |
|  |  | *%\end{verbatim}* |
|  |  |  |
|  |  | \newpage |
|  |  |  |
|  |  |  |
|  |  | \bibliographystyle{unsrt} |
|  |  | \bibliography{bibliography} |
|  |  |  |
| ... | ... |  |