Exhaustive SAR analysis tools for HTS hits based on intuition of medicinal chemists

Hironobu Koga

Screening Science and Technology Div., Astellas Research Technology Co., Ltd., 21, Miyukigaoka, Tsukuba-shi, Ibaraki, Japan  
*E-mail: hironobu.koga@astellas.com

(Received February 27, 2019; accepted March 29, 2019; published online April 15, 2019)

Abstract

High-throughput screening (HTS) is a common practice in drug discovery. Although the chemoinformatics community has proposed various approaches for HTS data analysis, medicinal chemists continue to long for intuitive tools for the structure-activity relationship (SAR) analysis of HTS hits. Here, the author propose SAR analysis tools that were designed to help medicinal chemists grasp the chemical space of interest with conventional SAR tables. These tools comprise an on-the-fly analysis environment and a series of computational protocols for data processing prior to the interactive analysis. The protocols are designed for the following processes: i) structural classification based on simple rules to mimic visual inspection by medicinal chemists; ii) exhaustive generation of promising SAR tables using Pharmacofragment (PHF), a novel substructure concept; and iii) comprehensive analogue search to identify compounds that correspond to blank cells in SAR tables from compounds at hand. A case study using data from a screen for ribosomal protein S6 phosphorylation inhibitors (PubChem AID:493208) suggests that these tools are useful for generating conventional SAR tables for practical application to large-scale data such as HTS.

Key Words: Structure-activity relationship, Similarity search

Area of Interest: Information and computing approach for drug design and ADMET study

1. Introduction

“When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.” [1]

High-throughput screening (HTS) is a common practice in drug discovery. It produces large amounts of assay data for compounds in an HTS library. This data set can serve as a critical source of information for triaging hit compounds. In HTS hit triage, it is important to examine not only each
hit compound but also sets of structurally-related compounds, or chemotypes. Chemotypes are usually analyzed for structure-activity relationships (SARs) to determine whether the hits of the respective chemotypes are promising for further exploration of the chemical space of interest.

One of the conventional approaches in SAR analysis is visual inspection of SAR tables, which summarize activity changes due to substituent conversion for a group of compounds with a common structural core. Many examples can be found in research articles in the medicinal chemistry field aimed at optimizing a lead compound. Although SAR tables are intuitively comprehensible, it is not common for medicinal chemists to apply them exhaustively to HTS data due to the laboriousness of their preparation, which usually requires intensive visual inspection of the chemical structures and manual dissection of the core and substituents. In addition, consequent SAR tables for HTS hits can be sparse because the HTS library is generally designed to select from an astronomical number of drug-like compounds, estimated at $10^{60}$ [2,3], and reduce it to a more practical number such as $10^4$ to $10^6$.

The chemoinformatics community has tackled SAR analysis for HTS data and proposed various techniques for structural classification prior to SAR analysis [4]. Most of these are based on similarity measures of the properties of the whole structure and are effective for classifying compounds. However, the results are not always convincing to medicinal chemists because there is little or no consideration for distinct scaffolds or substituents, on which medicinal chemists generally focus when they aim to uncover promising common substructures.

As a practical compromise, some medicinal chemists visually inspect potent compounds, manually extract common substructures, and subject them to existing computational R-deconvolution tools such as the R-group decomposition function in TIBCO®Spotfire® 7.5.0. In their visual inspection, they usually sort the compounds in descending order of activity and attempt to extract the substructures in a group of active compounds. Once the substructures are identified, they examine potential SARs for each of the compounds in the respective groups. This type of analysis occasionally proceeds in a trial-and-error fashion rather than in an intentionally planned and straightforward manner. Further, even in cases where consequent SAR tables cover only sparse chemical spaces, medicinal chemists may be able to intuitively determine SARs, allowing them to push forward in drug discovery projects.

In this article, the author reports the efforts to help medicinal chemists to more efficiently determine SARs. The author adopted a strategy for structuring the data prior to intuitive analysis by computationally mimicking the methods of medicinal chemists described above. The author first developed a computational protocol for compound classification by focusing on their Bemis-Murcko framework [5] and on the preferential properties determined using the assay. The author subsequently proposed a novel substructure concept called Pharmacofragment (PHF) and implemented it into the protocol used to produce tables containing all promising SARs beforehand. Unlike compound classification, this protocol allowed compounds to belong to more than one SAR table. Additionally, the author used TIBCO® Spotfire® 7.5.0, commercially available visualization software, to create an on-the-fly analysis environment for these structured data. The author also developed a comprehensive analogue search protocol for identifying compounds that correspond to blank cells in SAR tables created from commercially available sources or in-house compound libraries. A case study using data from a screen for ribosomal protein S6 phosphorylation inhibitors (PubChem AID:493208) [6] shows that these tools are useful for generating conventional SAR tables applicable to large-scale data such as HTS.
2. Materials and Methods

2.1 Dataset and software

A set of screening data for ribosomal protein S6 phosphorylation inhibitors (PubChem AID:493208) was retrieved from the PubChem Database [6]. Commercially available compound datasets were obtained from Kishida Chemical Co., Ltd. and Namiki Shoji Co., Ltd.

TIBCO® Spotfire® 7.5.0 was licensed from PerkinElmer Japan Co., Ltd. Pipeline Pilot 2018 ver.18.1.0.1694 was licensed from Dassault Systems BIOVIA K.K.

2.2 Scaffold-based compound classification

Structural classification of compounds was performed using an in-house protocol which aimed to mimic the visual inspection process of medicinal chemists (Figure 1). Compounds were sorted by potency using values such as the pAC50. The most preferable compound was selected as a representative for the molecular scaffold class being considered. Members of each class were selected if they had the same Bemis-Murcko framework or differed in the assembly of one ring compared to the representative compound. Once the members were selected for the current class, the most potent compound from the remaining compounds was selected as the next representative. This iteration continued until all compounds were classified. The procedure was implemented as part of the Pipeline Pilot protocol using the “Custom Manipulator” component written in the Pilot script.

Figure 1. Outline of scaffold-based compound classification

Compounds were sorted by potency using values such as the pAC50. Each compound was classified based on a set of scaffold-like fragments. Each set contained compounds with the same Bemis-Murcko framework and sub-fragments that were generated by removing one ring from the original Bemis-Murcko framework. Compounds were assigned to a class if their fragment sets overlapped in any way with that of the corresponding representative. Due to the structural expression employed, compounds without any ring assemblies were left to the end and combined into one group.
2.3 Pharmacofragment generation

The author proposed a novel substructure concept called Pharmacofragment (PHF) for the automated generation of possible structural cores for SAR tables. PHFs were generated by systematically deleting rotatable bonds attached to cyclic structures or hydrogen donors or acceptors (Figure 2).

The outline of the Pipeline Pilot protocol was as follows:

1. Rotatable bonds were assigned to be deleted if they connected any of following atoms or groups: terminal atom, hetero atom, CO, CS, or ring structure.
2. PHFs were generated by deleting one or two of the above bonds at a time. All combinations were taken into account.
3. R atoms were attached to newly generated terminal atoms.
4. PHFs composed of only carbons, such as alkyl chains or phenyl groups, were excluded.
5. The PHF type was defined as follows:
   i. TERM: Single fragments resulting from one cut; i.e. they have one R-atom each.
   ii. LINK: Single fragments resulting from two cuts; i.e. they have two R-atoms each.
   iii. WTERM: Sets of two terminal fragments resulting from two cuts; i.e. they have two R-atoms for each assembly and one R-atom for each fragment.

For TERM and LINK types, a prefix was attached based on their structural features. Those that were composed of one, more than one, or no rings had SR, MR, or LN added, respectively. Consequently, PHF types included SRTERM, MRTERM, LNTERM, SRLINK, MRLINK, and LNLINK. An example set of fragments is shown in the Supplementary Information (Table S1).

6. The number of bonds along the shortest path between R atoms was defined as the RRPath.
   For LINK, the RRPath was simply measured by counting the number of bonds between R-atoms.
   For WTERM, the RRPath was the same as that of the corresponding LINK.
7. Generated PHFs were merged according to their canonical smiles and RRPath.
8. If a group of compounds had more than one PHF that overlapped structurally, the largest PHF was selected.
9. A unique ID number was assigned to each obtained PHF.
10. The PHF data were stored in two data tables, whose primary keys were the PHF structure ID and the combination of PHF-ID and original compound ID.

Figure 2. Basic concept of Pharmacofragments (PHFs)
A compound to be processed is schematically shown at the top. Solid lines represent rotatable bonds. Circles, rectangles, and polygons represent ring assemblies and/or those connected to N, O, S, CO, or CS. As PHF is designed for 2D-SAR analysis, at most two bonds are broken at a time. All possible combinations of broken bonds are considered. Consequently, three types of fragments are generated: TERM, LINK, and WTERM. Details are described in the text.
2.4 Exhaustive SAR table generation

A SAR table contains information on a common substructure, i.e. a core, and its substituents. Cores for SAR tables were selected from the PHFs generated for hit compounds if their enrichment factor calculated using Scheme 1 was greater than 10. The core data were stored as the “PHF Table” (Figure 4). R-deconvolution was performed using the Pipeline Pilot component “Generate SAR Information”. Substituents for each of the PHFs were structurally classified using the Pipeline Pilot component “Cluster Molecules” with FCFP_4 [7] as the distance property. The SAR data were merged and stored as the large table “SAR Table” (Figure 4).

$$[\text{Enrichment Factor}]_{PHF_i} = \frac{[N_{Active}]_{PHF_i}/[N_{Tested}]_{PHF_i}}{[N_{Active}]_{all}/[N_{Tested}]_{all}}$$ Scheme 1

where [N_Active] and [N_Tested] are the number of active and tested compounds, respectively.

2.5 Comprehensive analog search

A number of algorithms were used to comprehensively search for structural analogs (Figure 3). The results were merged with their structures using flags to indicate the algorithm used for their identification. The whole procedure was implemented as a Pipeline Pilot protocol.

1. Similarity search with simple queries
   A standard similarity search was performed using the Pipeline Pilot component “Molecular Similarity”, where a functional-class fingerprint (FCFP_4) and Tanimoto coefficient (Tc) [7] were employed as the structural descriptor and similarity coefficient, respectively. Tc greater than or equal to 0.85 was employed as the similarity criteria.

2. Similarity search with secondary queries
   Virtual molecules derived from the original queries were employed as secondary queries for the FCFP_4-based similarity search described above. The virtual molecules were generated using the four algorithms below. Recursively generated query compounds were removed from all algorithms; that is, only novel molecules were considered.
   i. Bioisosteric transformation
      Virtual molecules were generated using the Pipeline Pilot component “Enumerate bioisosteres”, which applied classic bioisosteric transformation to the input molecules.
   ii. Matched molecular pair (MMP)-based recombination
      MMPs were generated for all combinations of query compounds using the Pipeline Pilot component “Matched molecular pair” with one bond break at a time [8]. Each compound in the MMP was expressed as a set of fragments; that is, a common core and its corresponding fragment. Compounds were grouped based on common cores in their Bemis-Murcko framework. Virtual molecules were generated by recombination of the common cores and fragments in the group.
   iii. PHF-LINK-fixed recombination
      Compounds were grouped based on LINK-type PHFs defined above. Virtual molecules were generated by recombination of the R1 and R2 substituents in the group.
   iv. PHF-WTERM-fixed recombination
      Compounds were grouped based on RRPaths for WTERM-type PHFs defined above. Virtual molecules were generated by recombination of the WTERM-type PHFs and the LINK-type PHFs in the group.
3. Automated substructure search
The substructure search was performed using PHFs selected for the SAR tables described above. TERM and LINK-type PHFs were employed as R-position fixed substructure queries. WTERM-type PHFs were structurally modified with virtual linkers of the same length as corresponding RRPaths, allowing all types of atoms and bonds. These substructure searches were applied to the compounds selected using the FCFP_4-based standard similarity search described above but with a Tc range from 0.6 to 0.85.

**Figure 3.** Outline of the comprehensive analog search
An outline of the comprehensive analog search is shown using an example query molecule. Details are described in the text.

2.6 Visualization

Four data tables shown in Figure 4 were read into TIBCO®Spotfire®7.5.0. They were linked to each other for interactive analysis by the user and presented in a set of predefined analysis pages, shown in Figure 4. The overview page was designed to have 3 visualization sections: a spreadsheet presentation of the assay results table and two scatter plots based on the tables containing assay results and PHFs (Figure 8). The SAR analysis page was designed to present either a 1D table for TERM- or WTERM-type PHFs or a 2D table for LINK-type PHFs (Figure 8). As these data tables were linked, data selection in one plot allowed linked data in the other plot to be highlighted and displayed in the spreadsheet and 1D- or 2D-SAR table.
Figure 4. Data tables read into the interactive analysis environment

Four data tables were read into TIBCO® Spotfire® 7.5.0. The “Assay Results” table contains data on all hit compounds. The “Compounds not tested” table contains the molecular structures of compounds that might be useful for further screening, including data such as the similarity search results of the hits. The “SAR Table” and “PHF Table” contain outputs from the protocols described in the text. The key properties used for the links are connected by solid lines.

3. Results and Discussions

A set of HTS data, PubChem AID:493208, was analyzed as a test case. The data set contained assay results for 43,989 compounds, including 342 compounds with activity outcomes annotated as “active”. For descriptive purposes, tested compounds other than those annotated as “active” and those obtained using the analogue search were annotated as “INACTIVE” and “NOT_TESTED”, respectively.

3.1 Scaffold-based compound classification

Scaffold-based classification of the 342 active compounds resulted in 149 structural groups including 87 singletons (Figure S1). Groups other than singletons were inspected based on their internal similarity to the FCFP_4-based Tc toward the respective representative compound. Most of the groups contained compounds with Tc below 0.85, with some with Tc below 0.60 (Figure S2). As this Tc range is often used as criteria for the ordinary similarity search, this structural classification should be unique to classification methods using conventional structural descriptors.

For example, one group contained CID 16013351 as its representative. Some of the members of this group contained Bemis-Murcko frameworks that differed from that of the representative compound and had a FCFP_4-based Tc toward the representative of less than or equal to 0.60 (Table 1).

3.2 Comprehensive analog search

A total of 342 active compounds were used as queries for the comprehensive analog search. A total of 9,803 compounds were identified as structural analogues from about 7,000,000 commercially available compounds (Table S2). This represented about 3 times as many compounds as that obtained using the simple query search, and about 60% of these were identified using only a single search method. This shows that the methods complemented each other and enhanced the efficiency of the search.
Table 1. An example of the scaffold-based classified compound groups
A group of compounds containing the most potent compound, PubChem CID 16013351, is shown. The color of the "structure" column is based on the Bemis-Murcko framework. FCFP_4 based Tanimoto coefficients (Tc) toward PubChem CID 16013351 are indicated. Details are in the text.

| PubChem CID | Structure | pAC50 | Tc  | PubChem CID | Structure | pAC50 | Tc  |
|-------------|-----------|-------|-----|-------------|-----------|-------|-----|
| 16013351    | ![Structure](image1) | 6.88  | 1.00| 16013393    | ![Structure](image2) | 5.68  | 0.60|
| 16013374    | ![Structure](image3) | 5.58  | 0.97| 6013392     | ![Structure](image4) | 6.23  | 0.56|
| 16013369    | ![Structure](image5) | 6.13  | 0.89| 16013375    | ![Structure](image6) | 6.38  | 0.51|
| 16013368    | ![Structure](image7) | 5.53  | 0.89| 16013390    | ![Structure](image8) | 5.93  | 0.48|
| 16013352    | ![Structure](image9) | 6.33  | 0.83| 3426438     | ![Structure](image10) | 5.53  | 0.45|
| 16013370    | ![Structure](image11) | 5.38  | 0.81|              |           |       |     |
| 16013372    | ![Structure](image12) | 6.08  | 0.81|              |           |       |     |
| 16013371    | ![Structure](image13) | 6.28  | 0.80|              |           |       |     |

3.3 Exhaustive SAR table generation
All tested compounds and those from the analogue search, totaling 53,792 compounds, were subjected to the exhaustive SAR table generation protocol. The protocol identified 1,624 PHFs as core fragments for SAR tables and found 29,155 PHF-compound pairs. Out of 342 active compounds, 341 compounds were assigned to the SAR tables. Most of the compounds were found to have more than one PHF for SAR analysis (Figure S3). This would allow users to investigate the SAR of the compound from multiple perspectives.

3.4 SAR analysis of compound CID 16013351 and its analogues
The author illustrates the utility of the SAR analysis tools using the most potent compound, CID 16013351, and its structural analogues as an example. First, the compound can be visually inspected to determine the scaffold-based structural group to which the compound belongs. Compounds in the group can be displayed in the spreadsheet (Figure 5A) by highlighting the corresponding points in the scatter plot (Figure 5B). The group containing compounds of interest contains 13 compounds including CID 16013351. The structural features for active analogues other than CID 3426438 can be depicted as shown in Figure 6.
Figure 5. A snapshot of the data overview page
A. A spreadsheet presentation of the assay results. Data displayed correspond to the points highlighted in the scatter plot in B. Cells containing pAC50 values are colored based on the binned value. B. A scatter plot of the assay results. Structural group IDs derived from the scaffold-based classification and pAC50 values are presented on the x- and y-axes, respectively. Points are colored based on their structural group ID. C. A scatter plot based on data in the PHF table. The confirmation rate defined as the numerator in Scheme 1 and maximum activity are presented on the x- and y-axes, respectively. Points are colored based on the PHF type and sized based on the number of compounds tested. The highlighted points correspond to those highlighted in the scatter plot in B.

Figure 6. SAR for CID 16013351 and its active analogues
CID 16013351 and its active analogues are depicted as substituted oxazoles with 5 substitution points with the substituents.
Figure 7. PHFs derived from CID 16013351
A. The most potent compound, CID 16013351, is highlighted in a scatter plot of the assay results. The plot is the same as that presented in Figure 5B. B. PHFs derived from CID 16013351 are highlighted in a scatter plot of PHFs. The plot is the same as that presented in Figure 5C. Example PHFs are numbered for later discussion: ①LINK, ②TERM, and ③WTERM. Details of each PHF type are provided in the text.
Figure 8. SAR table examples
Example snapshots of the SAR table that emerged on the fly. Details of the compound annotations are in the text. Substituents and linkers are categorized using the FCFP_4-based classification so that similar substituents are clustered together. The numbers above columns and beside rows are the substituent group IDs. A. 2D-SAR table for LINK-type PHFs (① in Figure 7B). The columns and rows represent R1 and R2 substituents, respectively. B. 1D-SAR table for TERM-type PHFs (② in Figure 7B). The rows represent R1 substituents. C. 1D-SAR table for WTERM-type PHFs (③ in Figure 7B). The rows represent linkers. The corresponding snapshots of the data overview page are shown in Supplementary Information (Figure S4).
Alternatively, the SAR can be inspected in detail together with all tested compounds and commercially available analogues. Highlighting CID 16013351 in the scatter plot of the assay results (Figure 7A) reveals that the compound has 12 PHFs (Figure 7B). The PHFs plotted in the upper right region are preferable because of their high confirmation rates and maximum potencies for the respective group of compounds. Once the PHF of interest is selected, the corresponding SAR table instantly opens in a separate page. As further examples, 3 PHFs in the upper right region (numbered ① to ③ in Figure 7B) are investigated below.

The SAR tables for LINK-type PHFs provide 2D representations of substituents R1 and R2. The substituents were categorized by FCFP_4-based classification so that similar substituents were clustered together. For the first example PHF (① in Figure 7B), the corresponding table indicates that alkyl chains and o-substituted phenyl groups are preferred substituents for R1 and R2, respectively (Figure 8A). This coincides with the SAR obtained from the structural classification (Figure 6).

The SAR tables for TERM-type PHFs represent sets of matched molecular pairs. For the second example PHF (② Figure 7B), the table clearly indicates that alkyl chains are the preferred substituent at the R1 position (Figure 8B).

The SAR tables for WTERM-type PHFs are designed to compare linkages between respective pairs of molecular fragments. The SAR table for the third example PHF (③ in Figure 7B) suggests that substituted benzenesulphonyl oxazoles are preferred as linkers (Figure 8C).

These SAR tables can also provide additional suggestions for further SAR exploration. As they indicate preferable substituents and/or linkers, they provide promising candidate substructures for novel compounds. These compounds can be procured or newly synthesized if they are annotated as “NOT_TESTED” or blank in the SAR tables.

4. Conclusion

One of the conventional approaches in SAR analysis is visual inspection of SAR tables by medicinal chemists. In this article, the author suggests that the same approach can be applied to HTS data by virtue of computational data preprocessing and interactive visualizations. Our SAR analyzing tools are designed to provide a more comprehensible overview of HTS data and can be used to suggest directions for further exploration of the chemical space of interest if applied together with yet-to-be-assayed structural analogues. Recently various computational approaches including state-of-the-art AI and machine learning techniques have become available for exploration of drug-like chemical spaces [9]. Some of these are used to screen compounds \textit{in silico} and to provide lists of recommended compounds for testing. As is the case with HTS, the tools can also be used to handle these lists and help medicinal chemists to comprehend the rationale based on the conventional SAR approach. In turn, this will hopefully lead to improvement in computational approaches with practical feedback from medicinal chemists.

Supplementary Information

Figures S1-S3, Table S1-S2 are in the supplementary materials (1) available at: https://doi.org/10.1273/cbij.19.19.
Acknowledgements

The author is deeply grateful to Ms. Rika Mutou and Dr. Satoru Shimizu for their dedicated technical assistances. The author also thanks Dr. Ayako Moritomo for technical support with accessing the commercially available compounds databases and Mr. Kyoichi Maeno and Mr. Ryushi Seo for their valuable discussions.

References

[1] http://www.diffen.com/difference/Data_vs_Information (accessed Feb 9, 2019).
[2] Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. Med Res Rev. 1996, 16(1), 3–50. DOI:10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6
[3] Mullard, A. The drug-maker's guide to the galaxy. Nature 2017, 549(7673), 445–447. DOI:10.1038/549445a
[4] Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov. Today 2010, 15(15–16), 630–639. DOI:10.1016/j.drudis.2010.06.004
[5] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. J. Med. Chem. 1996, 39(15), 2887–2893. DOI:10.1021/jm9602928
[6] https://pubchem.ncbi.nlm.nih.gov/ (accessed Jul 11, 2018).
[7] Hassan, H.; Brown, R. D.; Varma-O’Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. Mol. Divers. 2006, 10, 283–299. DOI:10.1007/s11030-006-9041-5
[8] Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J. Chem. Inf. Model. 2010, 50(3), 339–348. DOI:10.1021/ci900450m
[9] Lo, Y.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. Drug Discov. Today 2018, 23(8), 1538–1546. DOI:10.1016/j.drudis.2018.05.010