COMMENTARY

Single-Mode Compound Retrieval for QSAR, QSPR Data Sets, and Batch Mode Exact Structure Searching

CHRISTOPHER A. LIPINSKI

Pfizer Global Research and Development, Groton Laboratories, Eastern Point Road, Groton, Connecticut 06340

This commentary describes two simple procedures using commercially available software packages that greatly facilitate the creation of and replication of data sets intended for quantitative structure activity relationship (QSAR) and quantitative structure property relationship (QSPR) studies. Used properly, the procedures allow the capture of individual chemical structures from the Chemical Abstracts Service (CAS) SciFinder® software in a computer readable format that is recognized by most chemical database and computational calculation software packages. The researcher need not draw in a chemical structure to create a Molecular Design Limited (MDL) mol file, the 2D connection table format most commonly used to create the chemical depiction of compound or drug. The MDL mol format is needed so that properties can be calculated from the chemical structure alone. All that is required is that the compound or drug be located in SciFinder. The procedures are described in considerable detail because the key procedures for capturing structures from Chemical Abstracts Service (CAS) SciFinder through the use of Accelrys'Accord for Excel software are undocumented in either software.

Also described is a batch procedure that allows search of CAS SciFinder for the exact chemical structure of up to 25 compounds. Without use of this procedure, Scifinder can only be searched for an exact chemical structure a single compound at a time using a query consisting of a drawn in structure. Both the single-mode structure retrieval and batch-mode compound search procedures result in very significant time savings to the researcher creating or replicating QSAR/QSPR data sets and likely may enable structure searches that previously might not have been attempted because of researcher time constraints. These procedures do not affect positively or negatively the cost to the user of the searches against the SciFinder software. These costs are determined by CAS policy, and depend on the numbers of structures/compounds searched.

Locating a compound or drug in SciFinder is most accurately done using the CAS Registry Number®. The CAS Registry Number uniquely identifies a specific compound and salt form. Older deleted CAS Registry Numbers for a specific compound may be encountered, but a search on the current (or older) CAS Registry Numbers will always bring up the correct compound. If different salt forms of the same compound exist in the CAS databases, they will have different CAS Registry Numbers. By contrast, a search by compound or drug name may fail. IUPAC names for compounds are of no value in searches against SciFinder because IUPAC Names are not listed in the records. Commonly used drug names work fairly well. However, it is frequent to find variant or misspelled drug names in the scientific literature. A deviation between a search input name and the names stored in CAS databases in only a single letter or number will result in a search failure. Searches using drug trade names (as in very recent drugs) or company code numbers (as in early discovery stage compounds) fail more frequently than searches using common names.

SINGLE-MODE COMPOUND RETRIEVAL FOR QSAR, QSPR DATA SETS BY CAPTURING INDIVIDUAL STRUCTURES FROM SCIFINDER TO ACCORD FOR EXCEL

To retrieve a chemical structure in MDL mol format from SciFinder the researcher needs valid software licenses to both SciFinder and the Accord for Excel software available from Accelrys. The most basic (and least expensive) version of the Accord for Excel software is sufficient for the following procedure, which can also be performed with any of several other Accord for Excel variants offering other types of increased functionality. In essence, SciFinder provides the data source for structural, name, and property information. Accord for Excel software acts to translate a proprietary SciFinder CXF chemical depiction format into the MDL mol format, which is compatible with most chemistry database and software calculation programs.

To prepare an Accord for Excel spreadsheet for capture of chemical structures from SciFinder do the following. Enter CHEMISTRY into the first cell in column A. This prepares this column for chemical structures. To capture a structure from SciFinder following a successful search by CAS Registry Number or name do the following. Within SciFinder, click the microscope icon button at the top left of the first structure display window to open the Detail of Substance window. Highlight just the CAS Registry Number in the Registry Number: field. Be sure that there are no leading spaces copied either at the front or end of the registry number numerical string. Copy to the clipboard using the SciFinder Edit menu Copy function or use the control C keyboard equivalent. Navigate to an open Accord for Excel spreadsheet and use the Paste Chemistry command from the special Accord for Excel Chemistry menu to paste the structure into a cell in column A (somewhere below the CHEMISTRY header cell). The act of pasting the CAS Registry Number into a cell in column A in the Accord for Excel spreadsheet automatically converts the SciFinder proprietary chemistry format into an MDL mol file format. The pasted cell looks like a CAS Registry Number, but it is actually an MDL mol file connection table. Double clicking the pasted cell displays the chemical structure. Double clicking again collapses the structure back to the visual appearance of a CAS Registry Number. Both the CAS Registry Number as well as the chemical structure can be captured to the Accord for Excel worksheet if the clipboard contents are first pasted using the standard software Edit Paste command or control V keyboard equivalent into a noncolumn A cell, for example, the cell in column B adjacent to the cell in column A, which is to contain the chemical structure. This step is followed by the Chemistry Paste Chemistry command to paste the clipboard to the cell in Column A. Reversing this procedure will paste the smiles string depiction of the chemical structure rather than the AS Registry Number into cell B.

Any desired field can be copied and pasted between the SciFinder window and the Accord for Excel window. For example, SciFinder 2001 now contains calculated properties for H donors, H acceptors, Molecular Weight, log D, and Molar Solubility Calculated using Advanced Chemistry Development (ACD) Software. By alternating between SciFinder and Accord for Excel it is relatively easy (although boring) to capture all the compounds and some data required to create a moderate sized QSAR/QSPR table. With some practice the structure, CAS Registry Number and name can be captured at about a compound a minute, which is certainly much faster than if the structures had to be drawn by hand. Moreover, this procedure is capable of handling the capture of complex structures up to several thousand Daltons in molecular weight, thus eliminating the possibility of a drawing error. Accord for Excel allows all the usual operations possible in the native Excel plus additional operations pertinent to chemistry structures. Most germane to this discussion is that multiple structures and data can be exported as an MDL format structure data file (sdf) file. Most of the chemistry manipulation features of Accord for Excel are documented in the Accord for Excel help files. The import of structures from CAS SciFinder and background translation to MDL mol format is undocumented.

BATCH MODE EXACT STRUCTURE SEARCHING BY CONVERTING AN SDF FILE TO A BATCH CAS INDEX NAME FILE AND SEARCHING IN SCIFINDER

To perform batch mode exact chemical structure searches in SciFinder the researcher needs valid software licenses to both SciFinder and the Advanced Chemistry Development (ACD) Naming software. ACD/Name software comes in both basic and batch mode versions. The basic version is sufficient for the following procedure. In essence, the ACD/-Name software is used to generate a batch file of CAS-type Index names from an input file of 25 compound structures in MDL sdf file format. The batch file of CAS-type Index names can be pasted into the Explore by Substance Identifier box in SciFinder to permit an exact compound search on up to 25 compounds without having to individually draw in chemical structures. A very nice feature of this method is that the sdf file can come from any source. For example, it might come from a corporate proprietary database, a commercially available database, or from a file of compounds enumerated from a virtual library.

ACD provides systematic nomenclature software that can generate IUPAC names and CAS-type index names from a chemical structure input, and can operate in the reverse mode and produce a chemical structure from an input of CAS-type index name, an IUPAC name or from a trivial name (from a dictionary of 84,000 names in Version 5.0). An IUPAC name is essential for use in patent applications, but generally is not useful in SciFinder searches. It is the CAS-type Index name, which is useful in searching against SciFinder. Following the instructions in the ACD/Name help files the standard nonbatch ACD/Name software is set to operate in CAS-type index name mode and an sdf file of up to 99 compounds is imported. A history window displays the results from processing up to 99 compounds to 99 CAS index names. In the standard ACD/Name software the name fields must be individually copied and pasted to a flat text output file. The compound identifiers are lost in the sdf file import process, but the compound order is maintained and there are no skipped entries in the history window even if there is a problem in processing the structure to a name. So it is relatively easy to keep track of which name corresponds to which structure. As mentioned, the ACD Naming software can process up to 99 compounds at a time. The 25 compound exact structure search limit comes from the maximum number of names that SciFinder can process at a time. If the ACD batch name software is used, batch files of very large numbers of CAS-type index names can be created so there is a time saving from elimination of the name copying and pasting process. Even if the nonbatch ACD/Name software is used the time saved by the researcher is considerable. It is much faster to copy and paste 25 names than it is to draw in 25 chemistry structures.

In my experience the use of ACD/Name for batch exact chemistry structure searches is about 85–90% successful in the sense that 85–90% of chemical series are 100% successfully processed and perhaps 10% of series are processed with a high failure rate. Most of the failures seem to occur with complex carbohydrate structures with ambiguously drawn stereochemistry. There are some failures with complex peptidic compounds and the occasional highly symmetrical compound or when the CAS-type index name does not match the correct CA Index name, as issued by Chemical Abstracts Service (CAS) (Anyone requiring official CA Index Names or Chemical Substance names for reporting to regulatory agencies, should obtain them directly from CA at 1-800-753-4227 (toll free in North America) or 614-447-3613)

It should be noted that the CAS SciFinder License Agreement limits the amount of CAS information that can be downloaded, and restricts its redistribution. Authorized Users of SciFinder may not transmit or deliver SciFinder information in any form to any third party, with the exception of the following: (a) in copyrighted scientific publications when the search results are incidental to the publication, and (b) in reports to a Government Agency that are required by law or by administrative rule. Further, the SciFinder License Agreement includes a "reasonableness of use" section that limits excessive searching and downloading. *Please review your SciFinder License Agreement for details, or call CAS at 614-447-3613 for specific questions*.