A Convenient and Free Method for Adding Chemical Structures to CDD Datasets

Richard L. Apodaca

Metamolecular, LLC

8070 La Jolla Shores Drive #464
La Jolla, CA 92037
858.658.0860
metamolecular.com

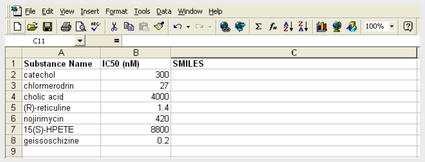
Abstract

Chemical structures are essential during the optimization of structure-activity relationships. Ideally, the structure of every sample tested during screening would be recorded and linked to each measurement. But sometimes this isn't feasible - the only available sample information might consist of a substance name, synonym, or CAS Number.

Manually adding structures to datasets and upload files is expensive and tedious. This poster describes an automated alternative based on a a free Microsoft Excel macro. Procedures for adding chemical structures to existing CDD datasets and upload files with this new macro will be illustrated.

The Problem

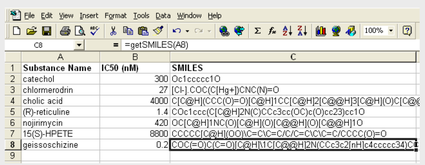
Adding structures to CDD datasets is simple - given SMILES representations in an upload file. But how can get these SMILES representations be conveniently obtained?



	A	B	C
1	**Substance Name**	**IC50 (nM)**	**SMILES**
2	catechol	300	
3	chlormerodrin	27	
4	cholic acid	4000	
5	(R)-reticuline	1.4	
6	nojirimycin	420	
7	15(S)-HPETE	8800	
8	geissoschizine	0.2	
9			

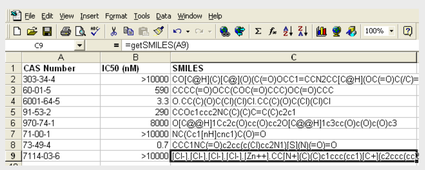
ChemCell as a Solution

SMILES can be added to an Excel file using a simple formula:



	A	B	C
1	**Substance Name**	**IC50 (nM)**	**SMILES**
2	catechol	300	Oc1ccccc1O
3	chlormerodrin	27	[Cl-].COC(C[Hg+])CNC(N)=O
4	cholic acid	4000	C[C@H](CCC(O)=O)[C@H]1CC[C@H]2[C@@H]3[C@H](O)C[C@@
5	(R)-reticuline	1.4	COc1ccc(C[C@H]2NCCc3cc(OC)c(O)cc23)cc1O
6	nojirimycin	420	OC[C@H]1NC(O)[C@H](O)[C@@H](O)[C@@H]1O
7	15(S)-HPETE	8800	CCCCC[C@H](OO)\C=C\C=C/C/C=C\C\C=C/CCCC(O)=O
8	geissoschizine	0.2	COC(=O)C(C=O)[C@H]1C[C@@H]2N(CCc3c2[nH]c4ccccc34)C1
9			

The same formula works with Chemical Abstracts Service (CAS) Registry Numbers:



	A	B	C
1	**CAS Number**	**IC50 (nM)**	**SMILES**
2	303-34-4	>10000	CO[C@H](C)[C@](O)(C(=O)OCC1=CCN2CC[C@H](OC(=O)C/C=
3	60-01-5	590	CCCC(=O)OCC(COC(=O)CCC)OC(=O)CCC
4	6001-64-5	3.3	O.CC(C)(O)C(Cl)(Cl)Cl.CC(C)(O)C(Cl)(Cl)Cl
5	91-53-2	290	CCOc1ccc2NC(C)(C)Cc2c1
6	970-74-1	8000	O[C@@H]1Cc2c(O)cc(O)cc2O[C@@H]1c3cc(O)c(O)c(O)c3
7	71-00-1	>10000	NC(Cc1[nH]cnc1)C(O)=O
8	73-49-4	0.7	CCC1NC(=O)c2cc(c(Cl)cc2N1)S(N)(=O)=O
9	7114-03-6	>10000	[Cl-].[Cl-].[Cl-].[Cl-].[Zn++].CC[N+](C)(C)c1ccc(cc1)[C+](c2ccc(cc

Recall and Accuracy

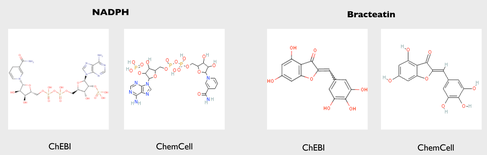
Evaluating the recall and accuracy of ChemCell required a benchmark containing name to structure mappings. We chose the ChEBI 3-star dataset,[1] which contains 22,749 human-curated name to structure mappings and is freely downloadable.

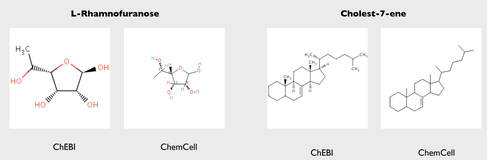
A random subset of 1,000 ChEBI 3-star entries was selected. Standard InChI Keys were generated from connection tables using IUPAC InChI v1.03.[2] These Keys were compared to the Standard InChI Keys generated by ChemCell.

Recall[a]	696/1,000	70%
Accuracy[b]	532/696	76%
Constitutional Accuracy[c]	636/696	91%

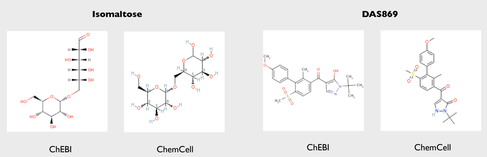
Notes: (a) Number of InChI Keys found; (b) InChI Key exact matches; (c) Atom connectivity InChI Key matches found by comparing first 14 characters of InChI Keys.

Many structure mismatches were due to over- or underspecified stereochemistry:





Some structure mismatches were due to slight constitutional differences:



How it Works

ChemCell invokes the Chemical Structure Lookup Service[3] over HTTP using Visual Basic running in Excel:

Excel — **Chemical Structure Lookup Service**

http://cactus.nci.nih.gov/chemical/structure/

Main Visual Basic Function:

```
Public Function getSMILES(ByVal name As String) As String
    Dim XMLhttp: Set XMLhttp = CreateObject("MSXML2.ServerXMLHTTP")
    XMLhttp.setTimeouts(2000, 2000, 2000, 2000)
    XMLhttp.Open("GET", "http://cactus.nci.nih.gov/chemical/structure/" +
                                name + "/smiles", False)
    XMLhttp.send

    If XMLhttp.Status = 200 Then
       getSMILES = strip(XMLhttp.responsetext)
    Else
      getSMILES = ""
    End If
End Function
```

Variations

Chemical Structure Lookup Service supports several conversions besides name to SMILES. The following are also supported and could be integrated into ChemCell with minimal effort:

Standard InChI
Standard InChI Key*
Structure Image
Molfile
Hydrogen bond donor/acceptor count
XLOGP2
Rule of 5 Violation Count
Molecular Weight and Monoisotopic Mass
Molecular Formula
IUPAC Name

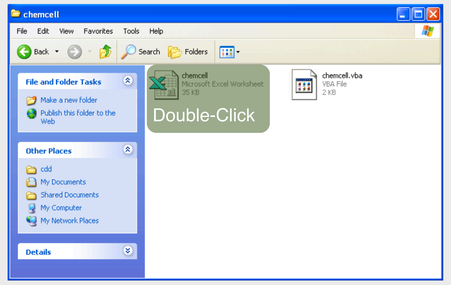
* Currently supported by ChemCell.

Download and Install

ChemCell can be downloaded from the CDD GitHub repository at:

http://github.com/cdd/chemcell

Installation:



Conclusions

The combination of Visual Basic and cheminformatics Web services such as Chemical Structure Lookup Services offers many possibilities to automate chemistry data workflows within Microsoft Office.

Acknowledgements

The Financial Support of Collaborative Drug Discovery is gratefully acknowledged.

References

1. ChEBI Database
   http://www.ebi.ac.uk/chebi/
2. IUPAC InChI Software
   http://www.iupac.org/inchi/
3. Chemical Structure Lookup Service
   http://cactus.nci.nih.gov/chemical/structure/