modlab

SmiLib v2.0 - Rapid Assembly of Combinatorial Libraries in SMILES Notation

User Manual

Andreas Schüller, Volker Haehnke
a.schueller@chemie.uni-frankfurt.de

Institute of Organic Chemistry and Chemical Biology
Johann Wolfgang Goethe-University Frankfurt am Main
Siesmayerstraße 70
D-60323 Frankfurt, Germany

www.modlab.de

March 17th, 2003
Last updated July 24th, 2006

If you wish to publish results obtained with SmiLib, please cite:

A. Schüller, G. Schneider, E. Byvatov; SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation, *QSAR & Combinatorial Science* **2003**, 22, 719-721.

Table of Contents:

1. Introduction

SmiLib [1] is a platform independent command line and graphical user inferface software tool designed to rapidly create combinatorial libraries in SMILES format. This document describes the command line version of SmiLib only, yet a graphical user interface version is available at www.modlab.de.

A combinatorial library is a set of molecules created by virtual reactions of so-called building blocks with scaffold molecules. Scaffolds are represented by Markush structures of molecules that contain R-groups – sites of variability. Those scaffolds undergo virtual reactions with building blocks – small Markush molecules themselves – yielding newly synthesized reaction products.

The combination of scaffolds with building blocks may be determined by a reaction scheme that selects only a subset of all possible reaction products. The resulting libraries are called combinatorial libraries, in contrary to virtual libraries which comprise a full enumeration of all scaffolds with all building blocks.
The dimensions of combinatorial libraries easily get very high. The complete enumeration of a virtual library with 10 scaffold molecules containing three sites of variability and a set of 100 building blocks would result in $10 \times 100 \times 100 \times 100 = 10.000.000 = 10^7$ virtual reaction products.

2. SMILES

SMILES (Simplified Molecular Input Line Entry System) is a simple yet comprehensive chemical nomenclature, a *line notation* (a typographical method using printable characters) for entering and representing molecules developed by David Weininger [2]. In general, SMILES use characters and numbers arranged in a string, a line of text, to describe molecules. The SMILES format is a very easy to handle concept for bioinformatical purposes and computer data exchange.
In SMILES notation atoms are written one by one starting at one point of the molecule, omitting H-Atoms. Atoms of the “organic subset” are specified as B, C, N, O, P, S, F, Cl, Br, I, using their common valences. All other atoms are specified enclosed in square brackets, e.g. [H] for a single H-Atom. By default, single bonds are assumed unless otherwise specified by an equation sign ‘=’ for double bonds and a hash sign ‘#’ for tipple bonds. Aromatic molecules can either be specified in their Kekulé notation or using lower case atom labels, indicating mesomerism. Branching of molecules is indicated by the branch being enclosed in round braces. Ring closure bonds are specified by adding a number from 1 to 9 to the right of either connected atom.
A brief and introductory tutorial on SMILES can be found on the Daylight Chemical Information Systems, Inc. homepage at http://www.daylight.com/smiles/f_smiles.html [3].

3. SmiLib – Concepts

SmiLib does not perform actual chemical reactions. It does not take into account chemical or physicochemical characteristics of chemical reactions but rather simply concatenates scaffold molecules and building blocks with single bonds, whether or not this could be actually done in a laboratory or in nature. This should be kept in mind when creating combinatorial libraries with SmiLib.

The main advantages of SmiLib are its simplicity to use, high flexibility in constructing combinatorial libraries (exact subset of molecules for virtual synthesis can be specified) and high speed of library construction. The combinatorial library is created by concatenating SMILES strings.

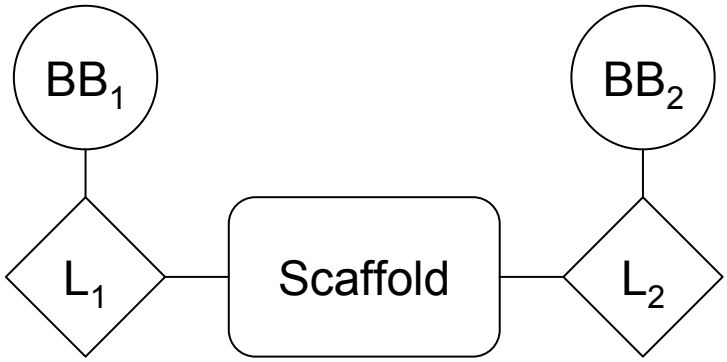

Figure 1: In virtual reactions SmiLib enumerates combinatorial libraries by means of scaffolds, linkers (L_1 and L_2) and building blocks (BB_1 and BB_2).

SmiLib deals with three classes of molecules: scaffolds, linkers and building blocks (Fig. 1). Linkers function as junctions between scaffolds and building blocks. The resulting molecule is a fusion product of its single molecule components. Thus SmiLib concatenates scaffolds with linkers and those with building blocks.

The SMILES format is used as both input and output format. Concatenation of SMILES is performed with a special technique, using “unsatisfied” ring closures [4]. In SMILES notation, parts of molecules that are not physically attached to each other, like salts, are notated with a separating point ‘.’, e.g. Sodium chloride: [Na+].[Cl-]. In combination with ring closures, this feature is used to create unconventional, yet valid SMILES. Ethane, for example, can be denoted by ‘CC’ or ‘C1.C1’. In the latter case, the bond between the two C-atoms is formed by a ring closure, rather than by a normal single bond.
In order not to be limited by the number of concatenations that can be performed, SmiLib uses an alternative syntax for ring closures, i.e. any two-digit number preceded by a percentage sign. In summary: ‘C%10.C%10’ ≡ ‘C1.C1’ ≡ ‘CC’ (Ethane C_2H_6).

In our implementation SMILES strings for scaffolds, linkers and building blocks have a slightly enhanced SMILES notation. In addition to normal SMILES [R1], [R2], [R3], etc. are used as labels for sites of variability and [A] is used as label for attachment sites (Fig. 2). An attachment site is part of the molecule, which is to be attached to a

scaffold or a linker. Building blocks are attached to linkers by concatenating the [A]-attachment site of the building block with the [R1]-site of variability of the linker. Linkers are fused with scaffolds by concatenating the [A]-attachment site of the linker with a site of variability ([R1], [R2], etc.) of the scaffold. During concatenation, the special atom labels like [R1] and [A] are removed from the final product.

In order to be able to create only a subset of all possible reaction products of a virtual library, a reaction scheme is used that exactly defines the composition of each reaction product. Similar to a connection table, the reaction scheme is comprised of index numbers that refer to scaffold, linker and building block molecules.

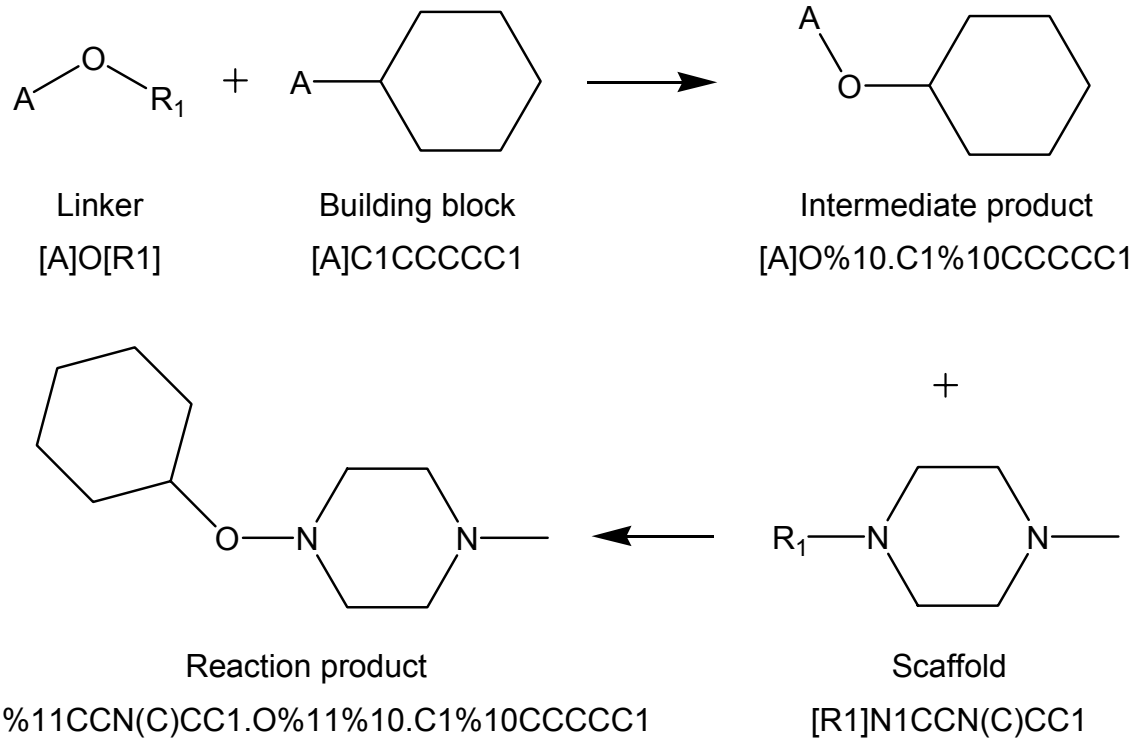

Figure 2: A virtual reaction along with corresponding SMILES: The building block is connected to a linker having the building block's A-group react with the linkers R1-group forming an intermediate product and a virtual A-R1 by-product (neglected). The intermediate product's A-group then undergoes reaction with the scaffolds R1-group yielding the final reaction product and a second A-R1 by-product (also neglected).

Each virtual reaction product is assigned its own unique identifier during the enumeration. The identifier is created based on the identifiers of the original fragments or their indices (line numbers) in the source files. This way the identifiers directly relate to the building blocks each virtual reaction product was synthesized of. The identifier meets the following definition: scaffold.linker1_bblock1.linker2_bblock2 … Scaffold id and linker/building block groups are separated by '.', linker and building block ids are separated by '_'.

4. SmiLib – Usage

```
Abstract: java –jar SmiLib.jar –s <scaffolds.smi> -l <linkers.smi>
          -b <building_blocks.smi> -f <library.smi>
          -r <reaction_scheme.txt> <options>
```

As a platform independent program written in Java, SmiLib is run with help of the Java virtual machine with "java –jar SmiLib.jar". Additionally SmiLib takes up to nine command line parameters.

Mandatory inputs of SmiLib are three ASCII files containing all scaffold, linker and building block molecule fragments in SMILES format (command line parameters -s <scaffolds.smi>, -l <linkers.smi>, -b <buildingblocks.smi>). Each line may contain exactly one molecule. The molecule SMILES string may be preceded by a tab separated identifier or name of this fragment which in turn will be used for the identifier of the resulting reaction product. No blank lines should be included.
R-groups in scaffold SMILES should be specified as a capital R followed by a number starting from 1 and increasing by one for each following R-group. In compliance to the SMILES notation, R-groups are to be enclosed by square brackets, e.g.: [R1], [R2], [R3], etc. Linkers have exactly one site of variability specified by either [R1] or [R].
Linkers and building blocks use a second type of special atoms: exactly one attachment side defined as [A]. That part of the linker molecule connected with the attachment site will be attached to a scaffold molecule in the combinatorial reaction. That part of the linker molecule connected to the R-group will be attached to a building block. If no linker should be used at all in a virtual reaction, a dummy linker [A][R] can to be used.

The forth parameter is the output format. One can choose whether the library shall be written in a file specified by the option "-f <library.smi>". If option "-f" is omitted SmiLib by default prints the generated combinatorial library to standard out.
The file extension of the library file specifies the file format. A library file name with the extension ".sdf" saves the combinatorial library in an SD file, any other file extension saves the library in an ASCII file in SMILES notation.

A reaction scheme file for the enumeration of the combinatorial library can be specified with option "-r <reaction_scheme>". The reaction scheme defines how the concatenation of scaffolds, linkers and building blocks is performed and contains tab separated index numbers (line numbers) of the SMILES input files in ASCII format.

Each row in the reaction scheme refers to one or more scaffolds in the scaffold SMILES source file. The first column of the reaction scheme specifies the scaffolds that shall be used in a virtual reaction. The following columns alternating specify the linkers and building blocks used for each side of variability of the scaffold. Column 2 specifies the linkers to be used on the first side of variability ([R1]), column 3 the building blocks to be concatenated to those linkers on the first side of variability, column 4 defines the linkers to be used on the second side of variability ([R2]), column 5 the building blocks that shall be concatenated to those linkers on the second side of variability and so on.
In summary the format of one line of a reaction scheme file in a two R-group example

is:

```
<scaf idx> <linker idx [R1]> <bb idx [R1]> <linker idx [R2]> <bb idx
[R2]>
```

with scaf: scaffold, bb: building block, idx: index number.

For ease of use, ranges can be defined with the '-'-operator, e.g. "1-5" for 1, 2, 3, 4, 5 and discontinuous ranges can be defined with the ';'-operator for e.g. "5;10" for fragment 5 and fragment 10. Both operators may be combined: the range "1-5;15;20-23" defines the use of molecular fragments 1, 2, 3, 4, 5, 15, 20, 21, 22 and 23. Ranges can be defined for each kind of fragment – scaffolds, linkers and building blocks.
If more than one scaffold in a line of the reaction scheme is defined using ranges, the same combination of linkers and building blocks will be attached to their sides of variability. So those scaffolds must have exactly the same number of variable side chains.

SmiLib by default performs syntax and validity checks on SMILES to ensure their compliance to SmiLib restrictions. To gain more flexibility, those checks can be deactivate with option "-c".

If the combinatorial library is saved in SD format, implicit hydrogens can be added with option "-y". This will slow down the enumeration of the combinatorial library due to the higher number of atoms per molecule.

Option "-u" starts SmiLib in interactive graphical user interface mode.

An unlimited number of scaffolds, linkers and building blocks may be used. The reaction scheme matrix size and the number of resulting reaction products is not limited either. All those parameters are only limited by amount of available memory. The number of definable R-groups is limited to 90 due to use of SMILES notation (percentage sign two-digit number ring closure notation).

5. Usage examples

Sample scaffolds file:

```
                  Line
                  number
scaffolds.smi:       1   | N2([R1])CCN([R2])C1=CC=CC=C1C2
                     2   | N([R2])(C3=C2C=CC=C3)CC12CCN([R1])CC1
                     3   | N1([R1])CCN([R2])CC1
```

Sample linkers file:

```
                  Line
                  number
linkers.smi:         1   | [A][R1]
                     2   | [A]C(=O)[R1]
                     3   | [A]O[R1]
                     4   | [R1]S([A])(=O)
```

Sample building blocks file:

```
                  Line
                  number
building_blocks.smi: 1   | [A]C
                     2   | [A]CC(C)C
                     3   | [A]CCC(C)C
                     4   | C(C(C)C)(C[A])C
                     5   | [A]CC1=CC=CC=C1
                     6   | [A]CCC1=CC=CC=C1
                     7   | C(C[A])(C)(C)C
                     8   | C1CC1C[A]
                     9   | [A]CC1CCCCC1
```

Sample reaction scheme file (for 2 R-groups):

```
                  Line
                  number
reaction_scheme.tpl: 1   | 1    1    1    1    1
                     2   | 2    4    4    2    3
                     3   | 4    3    7    4    6
                     4   | 1    1    9    2    3
                     5   | 3    2    3    3    5
                     6   | 2    1    5    2    9
                     7   | 2    2    4    1    3
                     8   | 1    1    1    1    2
                     9   | 1    1    1    1    3
```

Sample output:

```
Line
number
   1 | N2%11CCN%13C1=CC=CC=C1C2.C%11.C%13
   2 | N%13(C3=C2C=CC=C3)CC12CCN%11CC1.S%10%11(=O).C(C(C)C)(C%10)C.C%13%12(=O).C%12CC(C)C
   3 | N1%11CCN%13CC1.O%11%10.C(C%10)(C)(C)C.S%12%13(=O).C%12CC1=CC=CC=C1
   4 | N2%11CCN%13C1=CC=CC=C1C2.C%11C1CCCCC1.C%13%12(=O).C%12CC(C)C
   5 | N1%11CCN%13CC1.C%11%10(=O).C%10CC(C)C.O%13%12.C%12C1=CC=CC=C1
   6 | N%13(C3=C2C=CC=C3)CC12CCN%11CC1.C%11C1=CC=CC=C1.C%13%12(=O).C%12C1CCCCC1
   7 | N%13(C3=C2C=CC=C3)CC12CCN%11CC1.C%11%10(=O).C(C(C)C)(C%10)C.C%13CC(C)C
   8 | N2%11CCN%13C1=CC=CC=C1C2.C%11.C%13C(C)C
   9 | N2%11CCN%13C1=CC=CC=C1C2.C%11.C%13CC(C)C
```

6. Restrictions

SmiLib is subject to a few limitations that are important to know.
First, as already stated above, SmiLib does not perform actual chemical reactions. It does neither take chemical nor physicochemical characteristics of chemical reactions into account.

Second, the number of definable R-groups is limited to 90 due to use of SMILES notation (percentage sign two-digit number ring closure notation).
However an unlimited number of scaffolds, linkers and building blocks may be used. Also the reaction scheme matrix size and the number of resulting reaction products is not limited either.

Furthermore a few molecule limitations need to be considered. R-groups and attachments sites may not be connected to more than one atom. The following examples are **invalid**:

```
C[R1]C
C([R1]C)
C1CC[R1]1

C[A]C
C([A]C)
C1CC[A]1
```

R-group and attachment sites must be connected to the next atom by a single bond. The following examples are **invalid** (only shown with [R1]):

```
C=[R1]
[R1]=C
C#[R1]
[R1]#C
C(=[R1])C
```

Due to design limitations SmiLib can't handle SMILES with R-groups connected to atoms with specification of E/Z isomerism. Therefore SMILES are not allowed to include the symbols '/' and '\' (this behavior may be disabled with command line option '-c'). The following examples are **invalid**:

```
[R1]/C=C(F)/I
Br/C(Cl)=C(O/C=C/F)/[R1]
Br/C(Cl)=C(F)/[R1]
```

7. Compatibility

SmiLib has shown to produce molecules in SMILES format that are compliant with ChemDraw Ultra [5], Daylight SMILES Depict Service [6], Molecular Operating Environment [7], CORINA 3D Structure Generator [8], the Open Babel Package [9] and the Chemistry Development Kit [10].

We found that special care needs to be taken when working with chiral SMILES with dot-disconnected unsatisfied ring closures. Among the programs tested for compatibility, the Daylight SMILES Depict Service [6], the Molecular Operating Environment [7], and the CORINA 3D Structure Generator [8] worked flawlessly; ChemDraw Ultra [5] depicted some stereo centers incompletely configured, OpenBabel [9] needed special command line parameters for some output formats[1], and the Chemistry Development Kit [10] ignored stereo configurations in SMILES.

[1] E.g. to correctly convert from dot-disconnected SMILES to conventional SMILES use:
`./babel -ismi library.smi -osdf -x3 | ./babel -isdf -osmi`

8. License

SmiLib

9. Portions copyright

The following programs and libraries are included in the SmiLib v2.0 distribution:

AbsoluteLayout
Copyright 1997-2000 Sun Microsystems, Inc. All Rights Reserved.
Distributed under the terms of the Sun Public License Version 1.0.
http://www.netbeans.org/

BrowserLauncher2 version 1.0 rc4
Distributed under the terms of the GNU Lesser General Public License Version 2.1.
http://browserlaunch2.sourceforge.net/

The Chemistry Development Kit release 20060714
Copyright (C) 2000-2006 The Chemistry Development Kit (CDK) project.
Distributed under the terms of the GNU Lesser General Public License Version 2.1.
http://cdk.sourceforge.net/

Jakarta Commons CLI library version 1.0
Copyright (c) 1999-2001 The Apache Software Foundation. All rights reserved.
Distributed under the terms of The Apache Software License, Version 1.1.
http://jakarta.apache.org/commons/cli/

JGoodies Looks version 2.0.3
Copyright (c) 2001-2006 JGoodies Karsten Lentzsch. All rights reserved.
Distributed under the terms of the BSD License for the JGoodies Looks.
http://www.jgoodies.de/

Swing Layout Extensions version 1.0
Copyright (C) 2005 Sun Microsystems, Inc. All rights reserved. Use is subject to license terms.
Distributed under the terms of the GNU Lesser General Public License Version 2.1.
http://swing-layout.dev.java.net/

10. References

[1] A. Schüller, G. Schneider, E. Byvatov; SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation, QSAR & Combinatorial Science **2003**, 22, 719-721.

[2] David Weininger, "SMILES 1. Introduction and Encoding Rules", J. Chem. Inf. Comput. Sci., **1988**, 28, 31.

[3] Daylight Chemical Information Systems, Inc., 27401 Los Altos - Suite 360 Mission Viejo, CA 92691, http://www.daylight.com/.

[4] J. M. Barnard, *Reactions to Markush*, presentation at MUG 2000 (Daylight User Group Meeting), Santa Fe, NM, 24 Feb 2000.

[5] ChemOffice Ultra 7.0.1, CambridgeSoft Corporation, 100 CambridgePark Drive, Cambridge, MA 02140 USA, http://www.cambridgesoft.com/.

[6] a) D. Weininger*, J. Chem. Inf. Comput. Sci.* **1990***, 30,* 237-243; b) Daylight SMILES Depict Service, Daylight Chemical Information Systems Inc., 120 Vantis - Suite 550, Aliso Viejo, CA 92656 USA, http://www.daylight.com/daycgi/depict.

[7] Molecular Operating Environment (MOE), version 2005.06, Chemical Computing Group, 1010 Sherbrooke St. West, #910, Montreal, Canada, H3A 2R7, http://www.chemcomp.com/.

[8] a) J. Gasteiger, C. Rudolph, J. Sadowski. *Tetrahedron Comp. Meth*. **1990**, *3*, 537-547; b) CORINA 3D Structure Generator, version 3.20, Molecular Networks GmbH, Nägelsbachstraße 25, 91052 Erlangen, Germany, http://www.molecular-networks.com/.

[9] a) R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner, E. Willighagen, *J. Chem. Inf. Model.* **2006**, *46*, 991-998; b) The Open Babel Package, version 2.1.0b1, http://openbabel.sourceforge.net/ (accessed June 2006).

[10] a) C. Steinbeck, Y. Q. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, *J Chem. Inf. Comput. Sci.* **2003**, *43*, 493-500; b) The Chemistry Development Kit, release 20060714, http://cdk.sourceforge.net/ (accessed July 2006).