

Mesa Suite Version 1.2

Fingerprint Module

Copyright (c) 2003 Mesa Analytics & Computing, LLC

John D. MacCuish and Norah E. MacCuish

www.mesaac.com

support@mesaac.com

The Fingerprint Module in the Mesa Suite

The Mesa Suite is a collection of application modules for research and applications (diversity selection, library analysis, compound acquisition, lead hopping, HTS data analysis, database systems, etc.) for chemical information systems. The modules can be used separately or in tandem to perform a variety of 2D and 3D tasks. Currently there are five modules: Shape, Fingerprint, Grouping, Diversity, and ChemTattoo. Aside from the standalone module applications, a Khoros interface from Khoral Research is also available for each module. Presently we support Linux, Windows, and Irix 6.5 platforms.

The Fingerprint Module is written in C++ and requires OEChem from OpenEye Scientific Software, Inc. (www.eyesopen.com).

Summary

The Fingerprint Module is comprised of two programs:

MACCSKeys164Generator - Generates ASCII 164 bit fingerprints from SMILES strings input

MACCSKeys320Generator - Generates ASCII 320 bit fingerprints from SMILES string input

Each takes Daylight [SMILES](#) strings as input. **MACCSKeysGenerator** outputs 164 bit, and **MACCSKeys320Generator** outputs 320 bit fingerprints in ASCII format. The fingerprints are a public subset of 166 MDL MACCS keys, and recently published MDL MACCS 320 keys.¹ Two key bits have been removed from the original 166 keys in **MACCSKeysGenerator** since they would almost surely be turned off for all pharmaceutical compound data.

Theory Section

Characterizing chemical structures in binary form based on 2D representations facilitates many tasks of cheminformatics. Such binary representations are called *molecular fingerprints*. Molecular fingerprints were first developed by chemical information systems (CIS) companies for efficiency enhancements in chemical database queries. 2D chemical representations are typically stored in chemical databases in a connection table (e.g. SDFfile) or SMILES (Simplified Molecular Input Line ...) format. Exact match, substructure, and superstructure queries against large database systems turns out to be relatively slow if full query searches against each member of a database has to occur for every query. CIS companies designed molecular fingerprints as a means to screen database members and subselect chemical structures from databases that might in fact be a hit. Fingerprints are used to filter out compounds that do not meet certain criteria (e.g. "contains a benzene ring"), thereby avoiding such CPU intensive comparisons for exact match or submatching searches for the majority of a database. Filtering in this way increases search speed by only performing the "expensive" full query searching to a subset of potential answers rather than the whole database.

Several methodologies exist for chemical binary representations. For example, Daylight Chemical Information Systems fingerprint is often referred to as a *path-based* approach. This amounts to a unique subgraph matching of the graph representation of the chemical structure. In the Daylight algorithm the fingerprint is "learned" from the structures themselves. A molecular fingerprint is generated from a hash of all the unique connection paths (subgraphs) up to a maximum size (typically 8) into a fixed length bit string. Fingerprints may be folded to decrease the length and increase the bit density. Typical sizes for Daylight fingerprints are 512 or 1024 bits in length, but any power of two can be generated.²

Molecular Design Limited (MDL), created a key based fingerprint. This fingerprint uses a pre-defined set of definitions and creates fingerprints based on pattern matching of the structure to the defined "key" set. This key based approach relies on the definitions to encapsulate the molecular descriptions *a priori* and does not "learn" the keys from the chemical dataset. The MDL original public key set was 166 keys, and their private key set was comprised of 966 keys. Their recent publication of "drug-like" keys contains a subset of 320 keys from their 966 set.¹ So MDL fingerprints could take on a maximum bit length of 966. No folding occurs with this type of fingerprint.

Barnard Chemical Information Systems (BCI) uses a dictionary approach in which the keys for the fingerprinter are first generated from the set and then implemented in the description. This combines a bit of both of the Daylight and MDL approaches. Typically the BCI dictionary generates thousands of keys, resulting in molecular fingerprint bit lengths on the order of 5,000 bits.

Mesa A&C uses the 320 "drug-like" published by MDL to generate 320 bit string representations as well as the 166 bit string representations based on

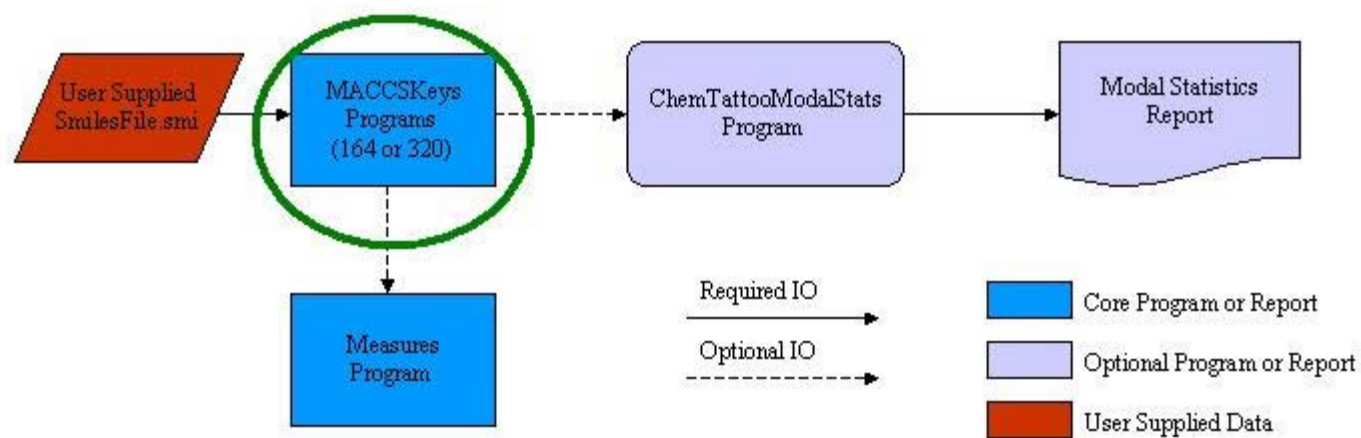
MDL's original public dataset. The keys are generated from SMARTS pattern matching against the chemical dataset using the SMARTS matching algorithm in OEChem from OpenEye Scientific Software.

The first three approaches mentioned above all have their advantages and disadvantages. In the Daylight case, learning the paths from the dataset enables new chemistries to be encapsulated in the fingerprints. Novel unique paths in the dataset will be encoded and input into the fingerprint. Searches against such databases will result in structure hit lists that contain this new chemistry. The disadvantage of the Daylight fingerprinting approach is that in some cases the unique paths do not encode symmetric systems well, it may not be able to distinguish between a monomer and a dimer, for example. Multiple counts of an identical path are not included in the description. In the MDL case, the user is dependent on the key-set created by MDL to encapsulate all of the chemistry that user has in their database or chemical dataset. The keys do take into account multiple counts of some features, which can be an advantage over the Daylight approach, but may not be able to uniquely describe a chemical dataset, if the keys are missing some of the chemical features in the dataset. BCI does seem to combine the best of both strategies, but at the expense of generating very large fingerprints. Mesa's approach is identical to the MDL approach so our fingerprints will have the same advantages and disadvantages. The decision as to which is the "best" fingerprinter is a decision one should not take lightly, and Mesa A&C believes the choice of a fingerprinter should be dictated by the data under study. For example, if one has an acquisition database full of inorganic substances that one would like to cluster, using our fingerprinter would be a "bad" choice.

Example Applications and Flow Diagrams

1. The Fingerprint Module programs are typically used to generate fingerprints for input into the **Measures** or **ChemTattooModalStats**. The **Measures** program generates a similarity or dissimilarity matrices which are necessary input for the **russianDollTransformation**, **SimilarityOutput**, or the **Clustering** program in the Grouping Module. The **ChemTattooModalStats** program returns the modal fingerprint at a threshold and the frequency fingerprint of the set. For more information on these programs please see the **Grouping Module Manual** and the **ChemTattoo Manual**.
2. The Fingerprint Module programs can generate binary string output for use with any non-Mesa program which requires such representation.

Fingerprint Module Flow



Detailed Summary of Programs with both I/O and Commandline Examples, and Specific References

The Fingerprint Module: **MACCSKeys164Generator** and **MACCSKeys320Generator**

The **MACCSKeys164Generator** takes as input a file of Daylight Smiles (.smi file) in single column format. E.g.

```
CCC(Br)CCI  
CCCC(Br)CCI  
CCCCC(Cl)C=C
```

This program is a MACCS (MDL) key 166 fingerprint generator that uses SMARTS matching provided in OEChem from [OpenEye Scientific Software, Inc.](#) It generates the respective binary fingerprints with just 164 of the 166 MACCS keys from MDL, using OEChem SMARTS matching from [OpenEye Scientific Software](#)

e. Two of the 166 MDL keys were removed as not needed for a typical compound library for drug discovery applications. As a default bit strings are output in column form without spaces. E.g.,

```
10010100...  
01010100...  
10010011...  
.  
.  
.
```

However, the programs contain the option of outputting fingerprints as a set of unsigned long integers, hexadecimal characters, or raw character bytes. A SMILES data file is needed for input, where the first column in the file contains just SMILES strings. The file may contain additional columns of associated data providing the columns are space delimited. The user also has the option of outputting just the fingerprints or the fingerprints and the SMILES plus any associated data. If a SMILES cannot be parsed this error is logged in an error log file with the offending SMILES and its index into the original SMILES data file. If the option to return all associated data is on, the error file will also return the associated data as well.

MACCSKeys320Generator is the same as *MACCSKeys164Generator* except that it has 320 keys.

Example Usages:

```
./MACCSKeys164Generator SampleColl1.smi -A -F -T >SampleFingerprints164.txt  
./MACCSKeys320Generator SampleColl1.smi -U -T -F >SampleFingerprints320.txt
```

Options explained:

The first option has 4 settings, -A (the default) ASCII 1s and 0s, -U unsigned long integers (space delimited), -C raw character bytes (dangerous as a text file, as these contain newlines, unprintable characters, etc.), -H hexadecimal character string. The second option toggles the output of the SMILES and any associated data (-T) for both the output and the error log file. The last option toggles whether the error log file is created or not. Note that the input SMILES data file can be input from stdin using '-' rather than the file name. E.g.,

```
./MACCSKeys320Generator - -U -T -F <SampleColl.smi>SampleFingerprints320.txt
```

References

1. *Reoptimization of MDL Keys for Use in Drug Discovery*, J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, JCICS, 2002, 42 (6), 1273-1280.
 2. Daylight Chemical Information Systems, Inc. [Daylight Clustering Manual](#).
 3. [MDL Information Systems, Inc.](#)
-
-

Example Script

Below is an example script that is included in the program directory, called TestScript. It contains the commandline interface details.

```
#TestScript for Cluster Module
#Copyright (c), 2002,2003, Mesa Analytics & Computing, LLC

# ./MACCSKeysGenerator 164 keys
# ./MACCSKeys320Generator 320 keys
#
# Usage: ./MACCSKeysGenerator filename.smi
# Or similarly,
# Usage: ./MACCSKeys320Generator filename.smi
#
# Note the .smi file is a single column of Smiles (one per row).

echo Number of lines =
wc -l SampleColl.smi
echo This is the number of compounds or "Size"

echo Generate Binary Fingerprints with 164 keys
./MACCSKeysGenerator SampleColl.smi -A -F -T > SampleFingerprints.txt
echo Generate Binary Fingerprints with 320 keys
```

```
./MACCSKeys320Generator SampleColl.smi -A -F -T > SampleFingerprints320.txt
```