PubChem Fingerprints

PubChem describes the 2-D structure of chemicals using a ‘CACTVS Substructure Key Fingerprint’, an fingerprint stored in a base64 encoding. The fingerprint is in section 4.1 ‘Computed Properties’ of section 4 ‘Chemical and Physical Properties’ on a compound’s PubChem page. CACTVS is a product of xemistry.com and is available in a free academic version. According to <http://dalkescientific.com/writings/diary/archive/2011/01/20/implementing_cactvs_keys.html> , the PubChem keys are ‘property E-SCREEN in [CACTVS] version 1.0, with non-default parameter “extended” set to 2’.

A description of the descriptors in the fingerprint can be found at <ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf> , reachable from the PubChem Help page. However, the description is not quite accurate. The current versions of fingerprints are 152 characters long (NOT 156 as described in the pdf file). This corresponds to 912 (6 x 152) binary digits. Also, the documentation says that there are 881 descriptors but examination of some examples chemicals (below) suggests that there are only 880.

Code to work with the fingerprints can be found in the R packages ‘rcdk’ and ‘fingerprint’; ‘fingerprint’ is installed when you install ‘rcdk’. As examples of working with fingerprints, I will use two pairs of similar chemicals: ethane and ethanol, and 1-bromo-2-chloro-cyclopentane and 1,2-dibromocyclopentane. The second pair is of interest because the last 2 descriptors in the PubChem fingerprint document are also valid SMILES codes for those two chemicals. These are descriptors 879 and 880, since the descriptor numbering starts at 0 and goes to 880. I look at these chemicals in pubchem\_fingerprint\_example.R, using some functions I wrote and put in pubchemfunctions.R.

PubChem pages for the four chemicals:  
Ethane:

<https://pubchem.ncbi.nlm.nih.gov/compound/6324>

Ethanol:

<https://pubchem.ncbi.nlm.nih.gov/compound/702>

1-bromo-2-chloro-cyclopentane:

<https://pubchem.ncbi.nlm.nih.gov/compound/26654>

1,2-dibromocyclopentane:

<https://pubchem.ncbi.nlm.nih.gov/compound/139113>

In pubchem\_fingerprint\_example.R, I convert SMILES codes for the four chemicals to PubChem fingerprints using functions from ‘rcdk’ and ‘fingerprint’, as implemented in pubchemfunctions.R. I also convert the fingerprints for those chemicals, as found on PubChem, from the base64 encoding using functions from pubchemfunctions.R. Then I compare the two sets of descriptors. The descriptors produced by ‘rcdk’ match those from PubChem, with one exception. 1,2-dibromocyclopentane should have descriptor 880 (the 881st descriptor in the list), according to the descriptions in the PubChem documentation. That descriptor shows up in the list of descriptors produced by the ‘rcdk’ package but not in the descriptors from PubChem. However, descriptor 879 shows up in both the ‘rcdk’ set and the PubChem set for 1-bromo-2-chloro-cyclopentane.

I also calculate Tanimoto similarity for the two pairs of chemicals using the ‘tanimoto\_binaryvectors’ function I wrote (in pubchemfunctions.R). There are some other difference/similarity metrics in the ‘fingerprint’ package.

Other notes

PubChem also calculates 3-D similarity for some pairs of chemicals; 2-D and 3-D similarity do not necessarily overlap, as shown in a paper by Kim et al. : ‘Similar compounds versus similar conformers: complementarity between PubChem 2-D and 3-D neighboring sets’, J Cheminform (2016) 8:62.