J Comput Aided Mol Des (2006) 20:179–190
DOI 10.1007/s10822-006-9044-4

Permuting input for more effective sampling of 3D conformer space

Giorgio Carta · Valeria Onnis · Andrew J. S. Knox · Darren Fayne · David G. Lloyd

Received: 21 December 2005 / Accepted: 18 March 2006 / Published online: 14 July 2006

Abstract SMILES strings and other classic 2D structural formats offer a convenient way to represent molecules as a simplistic connection table, with the inherent advantages of ease of handling and storage. In the context of virtual screening, chemical databases to be screened are often initially represented by canonicalised SMILES strings that can be filtered and preprocessed in a number of ways, resulting in molecules that occupy similar regions of chemical space to active compounds of a therapeutic target. A wide variety of software exists to convert molecules into SMILES format, namely, Mol2smi (Daylight Inc.), MOE (Chemical Computing Group) and Babel (Openeye Scientific Software). Depending on the algorithm employed, the atoms of a SMILES string defining a molecule can be ordered differently. Upon conversion to 3D coordinates they result in the production of ostensibly the same molecule.In this work we show how different permutations of a SMILES string can affect conformer generation, affecting reliability and repeatability of the results. Furthermore, we propose a novel procedure for the generation of conformers, taking advantage of the permutation of the input strings—both SMILES and other 2D formats, leading to more effective sampling of conformation space in output, and also implementing fingerprint and principal component analyses step to post process and visualise the results.

Keywords Conformers · Docking · Drug discovery · Fingerprints · Scoring · SMILES · Virtual screening

G. Carta · V. Onnis · D. Fayne · D. G. Lloyd (✉)
Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland
e-mail: lloyddg@tcd.ie

A. J. S. Knox
School of Pharmacy and Pharmaceutical Sciences, Trinity College Dublin, Dublin 2, Ireland

Introduction

The advent of drug design, concurrent with the accelerated evolution of computational power, allows us to manipulate and store large amounts of data in different formats, for pre-processing, substructure searching, 3D-matching and finally for creation of a combinatorial library [1–3]. A widely used method to handle compound structures is to represent them in SMILES format. Simplified Molecular Input Line Entry Specification (SMILES), proposed by Weininger in 1988 [4] is a simple linear chemical language for specifying molecules or molecular fragments. Compared with most file formats, SMILES is broadly used in computational chemistry and is supported by most major software tools in the field, reducing data exchange due to its compact nature. SMILES strings have been extensively used in both pharmaceutical and academic settings to uniquely code a 2D structure of a molecule with correct stereochemical and protonated states represented.

However, a molecule although uniquely represented, can be described by differently ordered SMILES strings. Rules can be applied that specify the generation of SMILES but, intriguingly, we show that different software packages produce different SMILES

strings for the same encoded molecule. If the SMILES strings are used for representing the 2D structure of the molecules, possible permutations of the string should not affect the generation of the 3D structure.

However, when the SMILES strings are used as input for the generation of conformers, it becomes clear that different permutations of the same string can generate diverse sets of conformers. This has a large impact for all applications and validations where a dataset was initially represented by SMILES strings. Our group has previously shown the impact that different SMILES strings representations have on the docking process in terms of enrichment (E) and false positive (FP) rate calculations [5].

To examine this concept we carried out a study on the permutation of SMILES strings using data from a set of 15 crystal structures extracted from a recent study by Vigers et al. [6] and also 2 X-ray structures of interest within our group [7,8] (Table 1). Our aim was to quantitively show the influence that each permutation of a SMILES string, generated by several software programs, has on the production of conformers. We also show how this phenomenon can be used, in some cases, to increase the rate of production of a diverse set of conformers.

Figure 1 illustrates the process whereby a single PPAR agonist 3D structure was reduced to SMILES format using Mol2smi (Daylight Chemical Information Systems, Inc.) [9], MOE (Chemical Computing Group) [10], Babel (Openeye Scientific Software) [11] and ChemSketch (Advanced Chemistry Development) [12]. The SMILES string produced by each software package, as shown in Fig. 1, are different for the same molecule.

As SMILES notation represents a linear form of a connection table, we also wanted to investigate if the concept was extendible to more elaborate connection tables, as those found in formats such as sdf.

Permuted SMILES strings were generated for ligands extracted from 17 crystal structures in the Protein Data Bank (PDB) and then used as input for different conformer generating software packages (Omega v1.8.1, Corina v3.6, Rubicon and Catalyst v4.9.1).

Principal component analysis (PCA) of several 3D descriptors and potential energy was employed to investigate and compare the conformational space explored by the same output quantity of conformers arising from a permuted set of a SMILES, in comparison to conformational space sampled when using a single unique canonical SMILES string.

Methods

Seventeen crystal structures from the PDB were selected. Most of the structures were selected from the list used to validate FlexX [13]/Gold [14] and also MASC [6]. All structures were solved at a resolution ≤ 2.5 Å, except for 1P93 (2.7 Å) the glucocorticoid complex) and they represent a wide variety of different structural characteristics such as size, number of

Table 1 List of ligands extracted from the PDB and their associated resolutions

	PDB code	Ligand–protein complexes	Resolution (Å)	Perm[a]	P. Sel.[b]
1	1AQ1	CDK	2.2	35	6
2	1D1Q	Tyrosine phosphatase	1.7	14	3
3	1GLQ	Glutathione-*S*-transferase	1.8	30	6
4	1ABE	Arabinose-binding protein	1.7	10	2
5	1AZM	Carboic anidrase	2.0	14	3
6	1CBX	Carboxypeptidase	2.0	15	3
7	4DFR	DHFR	1.7	33	7
8	1EBP	Retinoic acid binding protein	2.2	22	4
9	1HYT	Thermolysin	1.7	15	3
10	1MRK	Ribosome inactivating protein	1.6	19	4
11	1P93	Glucorticoid	2.7	28	5
12	1PHF	Cytochrome p450-cam	1.6	11	3
13	4PHV	HIV-protease	2.1	46	9
14	1I7I	PPAR gamma	2.35	28	6
15	1POC	Phospholipase A2	2.0	31	6
16	1SRJ	Streptavidin	1.8	22	4
17	1TPP	Trypsin	1.4	15	3

[a]Specifies the number of permutations generated by 'Permsmi' (Daylight Inc.)

[b]Indicates the number of strings selected to carry out the analysis, including the canonical SMILES string, as generated by Mol2smi (daylight)

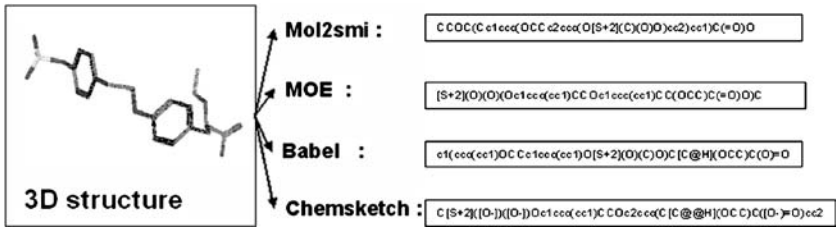

Fig. 1 Classification of conversion programs applied and their associated SMILES strings generated

rotatable bonds and features. The list of compounds studied is shown in Table 1.

Preparation of ligands

Compounds were downloaded and visually inspected within MOE [10]. Ligands were extracted from the complex and all hydrogen and coordination atoms deleted. Mol2smi was utilised to reduce the 3D representation to an initial canonical SMILES version.

Permsmi [9] was subsequently used to produce permutations of the SMILES string, starting with each of the atoms in the input SMILES in turn. The set of permutations of the molecule are written out in pseudo-canonical order (i.e. ring closures are renumbered). Thus, the longer the string, the higher the numbers of permutations 'Permsmi' is able to generate. An average of 22.8 permuted SMILES strings were created for each molecule. We chose to use every fifth permuted SMILES string of a molecule as we deemed it to be representative of the set of five making sure to include the canonical SMILES strings initially produced by mol2smi. This procedure narrows down to an average of 5.5 permutation strings per molecule, which were used as input for generating 3D conformers as illustrated in Fig. 2.

Conformer generation

Corina

Corina v3.6 [15] produces 3D conformations by combining monocentric fragments of standard bond lengths, appropriate angles and dihedral angles. To incorporate appropriate torsion angles, Corina uses a set of rules derived from statistical analysis of the Cambridge Structural Database (CSD) [16]. Furthermore, Corina is able to generate only a single conformer per SMILES string.

Corina, retaining default settings, was used to generate conformers of the permuted list and the output was saved in mol2 format. One conformer per SMILES string was created for a total of 94 conformers.

Omega

Omega v1.8.1 [17] generates conformers in three separate steps. First, a generic 3D structure is built. This primary conformation can be minimized with the MMFF force field. In the second stage, in order to generate different 3D conformations, Omega applies a depth-first divide and conquer algorithm. More simply, Omega disassembles the molecules into small

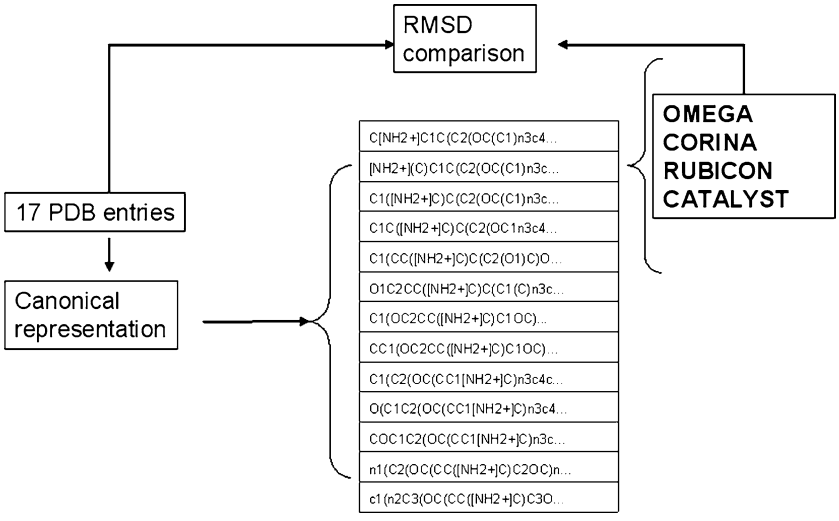

Fig. 2 Protocol for the evaluation of SMILES string permutations

fragments, and calculates the fragments energy using a Deidral FF, and finally reassembles the fragments in order. The generated conformations are limited by their total energy (e-windows) (Deidral force field, no electrostatic interaction) and molecules whose RMSD values are too similar (RMSD value cut off) are discarded. The third and final step offers the possibility to refine conformations with standard MMFF to avoid the presence of high energy unrealistic conformations.

Omega was set to generate 50 conformers per SMILES string. The output file was set to save as mol2 format and the number of rotatable bonds increased to 30, to allow all molecules to pass the Omega filter. All other remaining parameters were kept as default. 1 azm was discarded by Omega filter due to the presence of a zinc atom. A total of 2,406 conformers were generated.

Catalyst (fast)

Catalyst [18] uses a poling algorithm for the generation of conformers with the CHARMM force field, which does not include electrostatic terms. There are two integrated algorithmic methods of conformation creation to choose from: the BEST and FAST algorithm, of which FAST was the method of choice for this study. Its conformational search algorithm is based on a grid search of torsion values in combination with some energy minimisation to resolve high-energy configurations, such as overlaps of van der Waals radii.

A maximum number of 50 conformers per each SMILES string were created with the output saved in sdf format. Again, 1 azm was discarded, and a total of 3,623 conformers were generated.

Rubicon

Employing Rubicon [19], conformers are initially generated with a distance-geometry method that samples random conformations—this is by default a non-stochastic process, yielding reproducible output for a given input string in separate runs. These 3D structures are then minimized for distance and volume bounds violations. Only one minimizer algorithm is provided with version 4.3 of Rubicon [20].

SMILES strings were first converted into tdt format by smi2tdt [9] since Rubicon does not read the SMILES format and the output set to pdb format. A total number of 3,845 conformers were generated.

The default settings were used, except where noted, when implementing the conformer generation software packages. All conformations were superimposed by matching non-hydrogen atoms, to the corresponding unmodified X-ray ligand structure. To remove artificial differences resulting from RMSD deviations attributed to automorphisms [21,22], the program RMSD (Openeye Scientific Software) was used for analysis. The minimum RMSD values (the closest conformers to the crystal structure) and the standard deviations were then calculated for each set.

As it is clear that sampling conformer space with an ensemble size of 50 output molecules might be not sufficient to allow the software to produce a reasonable and diverse ensemble of conformers, two further experiments were performed. Thus, a subsequent RMSD analysis was carried out allowing the generation of a larger number of conformers (sets A and B described below).

Set A: From the original set of crystal structures we selected a molecular ''probe'' characterized by an average number drug like rotable bonds (1DFR, 10 RBs)—representative of the typical drug-like compounds encountered in commercial screening datasets used in virtual screening. The canonical string (cansmi) and subsequently eight permutations (Permsi) were generated as described above. In this study case the conformers were built using OMEGA. The 1DFR canonical SMILES string was first used and highest number of conformers with the default setting was obtained (a total 280 conformers with *max_confs* flag set to 500). To produce a comparable set of conformers the same number of output structures were replicated for the permuted string setting the *max_conf* flag to 35 (35 confs by 8 permutations). The results were then imported in MOE and 3D potential energy descriptor (MMFF94), PCA analysis were then calculated and plotted.

Set B: The same canonic SMILES string and its eight permutations were then used for a subsequently run. OMEGA was used to generate conformers from the canonic string with a *RMSD (rms* flag) cut off of 0.5 (default = 1) and an increased *energy window* (*ewindow* flag) to reflect the energy window span demonstrated by the conformers previously generated in A in terms of potential energy. (about 60 kcal/mol calculate with MMFF94 force field). *Max_confs* flag was increased to 1,000,000 to remove any restriction in the number of output structures. This produced an upper ensemble of **X** conformers from the single input string. As in experiment *A* a similar final number of conformers were produced from the permuted strings through setting *max_confs* to 310 and leaving the *energy window* and the *RMSD* cut off at the default settings. Again, the sets generated were imported in MOE and 3D potential energy descriptor and PCA analysis were carried out and plotted in graphs.

Fingerprint analysis

Conformers generated from the ensemble of permutations of a single molecule were then considered. For example, a set was represented as a collection of conformers produced from the six permutations of the PPAR ligand SMILES string. To measure the number of duplicate conformations presents in the set we imported the conformational file into MOE and an eigenvalue spectrum shape fingerprint ESshape3D 10 analysis was undertaken. Each fingerprint is of fixed length and allows for a comparison of 3D molecular shapes considering the heavy atoms of a molecule. Duplicate conformers appeared in the database with identical fingerprints, allowing us to calculate the percentage of unique conformers.

PCA analysis

In order to investigate the chemical space explored by the set generated for the set A and B previously described was analysed with a principal component analysis. A total of six MOE 3D descriptors (all x3D descriptors, *x*–*y*–*z* dipole moment, and principal moment of inertia *x*–*y*–*z*) were selected and computed for all the conformers generated to reflect the different topological characteristics of the 3D molecules in the output datasets. The aim of PCA is to reduce the dimensionality of a set of descriptors by linearly transforming the data. In detail, in our analysis six vectors have been transformed into a manageable 3D space described by three principal component vectors, where each of the 3D vectors is a combination of the selected descriptors.

To achieve this result the number of principal components were limited to three and all the other options in the MOE PCA module were run as default settings, without applying specific weighting to any of the 30 descriptors selected. The three principal components were then plotted as *XYZ* coordinates respectively for each compound. At the end of the process each molecule in the database is described by a set of coordinates, and may be plotted relative to all other members in the three-dimensional space.

Alternative 2D representations—connection tables

The connection table is a common and predominant way to simply represent a chemical structure—in its simplest incarnation we encounter SMILES strings. A typical file extension that contains a more complex connection table describing the organization of the atoms within a molecule is the SDF file format. Typically an SDF file includes two essential main sections: the first describing the 3D coordinates of the individual atoms in the molecules, the second depicting the connections of each atom with the others and the bond order present (1 = single bond, 2 = double bond) [23].

Section 1 (3D coordinates)

54	57	0	0	0	0	0	0	0	0	0							
	−0.3250			0.1625			0.0000			C	0	0	0	0	0	
	2.6042			1.8667			0.0000			C	0	0	0	0	0	
	−0.3250			−0.9917			0.0000			C	0	0	0	0	0	
	0.6458			0.7083			0.0000			C	0	0	0	0	0	
	2.6042			0.7083			0.0000			C	0	0	0	0	0	
	1.6250			0.1625			0.0000			C	0	0	0	0	0	
	....																
	....																

Section 2 (Connection table)

3	1	1	0	0	0
4	1	1	0	0	0
12	1	1	0	0	0
1	22	1	0	0	0
2	9	1	0	0	0
8	2	1	0	0	0
2	23	1	0	0	0

For this part of our study a bespoke perl utility was written to produce random permutation of the connection table. In detail, the utility randomises the position of the first section of the file and, while keeping trace of movement, it rewrites a new different connection table underneath. To avoid the production of duplicates an adjacency matrix is first calculated from the connection table and subsequently its connectivity stack [24].

The connectivity stack of a matrix is a sequence of the adjacency matrix values $s_k = a_{ij}$:

$$S = (s_1, s_2, s_3 \ldots s_k \ldots s_n)$$

where the position of the *k*th element is given by

$$k = \frac{i + (j\ 1)(j-2)}{2} \quad \text{with } (i<j)$$

Identical stack matrixes correspond to identical adjacency matrix and, thus, identical connection tables.

To investigate if permutation of connection tables in more complex formats such as SDF resulted with the same effect observed for the permuted SMILES strings on conformer generation, we first generated an SDF file from the ligand co-crystallised in PDB entry 1DFR [25]. OMEGA was then used to generate 280 conformers (*max_confs* set to 500 and remaining parameters as default). As described above, the perl script was used to produce 10 permutations of the crystal structure sdf file.

Finally, to reproduce the same amount of conformers as above, *max_conf* was set to 10 to furnish 280 conformers from the list of permuted connection tables of 1DFR. RMSD calculation was then carried out between the resulting conformers and the co-crystal ligand in 1DFR using OpenEye's RMSD [26].

Results and discussion

Two categories of data are reported in the results: a quantitative result taking into account the number of conformers generated by each software package and a qualitative result, where we considered the characteristics of the conformers generated. From a quantitative point of view, some software tools show interdependence between string permutations and the numbers of conformers generated. In others words, a different number of conformers were generated when starting from a different permutation of the same SMILES string.

In particular, Omega and Rubicon exhibited this behaviour, as although the maximum numbers of conformers to be generated was set to 50, this number was not always constructed. For example, Rubicon generated 27 conformers from the initial canonical SMILES string and up to 32 for one of its permutations for the ligand extracted from 1DFR. Omega produced only 9 conformers for the canonical string and 14 for a permutation of it in the case of the X-ray ligand taken from 1MRK. Less affected was Catalyst, which in the worst case (1CBX) generated 39 conformers for the canonical and 36 for most other permutations. The full set of results is shown in Tables 2–8. Results in Tables 2, 3, 4 report the RMS standard deviation values calculated for canonical SMILES strings and their permutations. Cells labelled with 'D' were molecules discarded by the software in the cases of both Omega and Rubicon. Label 'A' indicates that standard deviation was not computable because just one conformer per string was produced. Empty cells indicate that no permutation was produced for that string.

It is clear from Table 2 that each set of the permuted SMILES string generated molecules exhibits a different SD of the RMSD values for each set of conformers compared to the X-ray structure. That suggests that each set includes a different pool of RMSD values when compared with their respective crystal structures. Similar results generated by Omega are also observed in the case of Rubicon, with different RMSD values also being produced (Table 3). Table 4, however, shows that Catalyst is not as affected as Omega or Rubicon, in the conformers being generated from the permutations of the canonical SMILES string. Approximately the same SD values appear for the set indicating that identical conformers were created by each permutation. In the case of Corina, no SD calculations could be carried out as only one conformer per SMILES string could be produced.

We next sought to assess the quality of the conformers produced by each software package with respect to the different permutations of a SMILES string. Tables 5, 6, 7, 8 show the minimum RMSD values retrieved from each sets. Label D and empty spaces have the same significance as in the previous tables.

Table 2 depicts that minimum RMSD values (the closest conformers to the crystal structure) are derived from a permuted string rather than the canonical one which would be normally used in a virtual screening process. Importantly, in some cases, the lowest RMSD value is nearly half that observed with the canonical SMILES string (eg.1POC, 4DFR, 1ABE). As before, we see that the same holds for Rubicon as in the case of Omega, where the minimum values retrieved in the permuted sets are lower than those of the canonical string (Table 6). A less

Table 2 RMSD standard deviation for Omega

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	A	0.21	0.407	0.28	D	0.353	0.291	0.369	0.36	0.426	0.12	0	0.717	0.434	0.417	0.417	0.516
P1[b]	0.007	0.021	0.398	0.249	D	0.258	0.53	0.355	0.269	0.394	0.184	0.304	0.741	0.666	0.342	0.698	0.516
P2	A	0.2	0.456	0.224	D	0.344	0.316	0.308	0.135	0.445	0.163	0	0.568	0.888	0.324	0.85	0.6
P3	A	0.177	0.461		D	0.338	0.365	0.349	0.337	0.316	0.099	0.304	0.602	0.88	0.345	0.253	0.0554
P4	A		0.477				0.461	0.332		0.443	0.163		0.612	0.767	0.476	0.864	
P5	A		0.476				0.346				0.177		0.496	0.649	0.425		
P6	A		0.487				0.437						0.532	0.426	0.433		
P7							0.504						0.695				
P8													0.495				
P9													0.496				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 3 Minimum RMSD for Rubicon

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.45	0.31	1.53	0.38	D	0.49	1.67	0.57	0.24	0.91	1.17	0.1	2.02	2.46	1.26	0.17	0.39
P1[b]	0.38	0.31	1.28	0.11	D	0.61	1.73	0.89	0.23	0.94	0.99	0.09	1.72	2.12	1.46	0.19	0.46
P2	0.45	0.31	1.49	0.47	D	0.46	2.07	0.45	0.24	0.8	0.66	0.11	2.09	2.54	1.76	0.18	0.39
P3	0.47	0.31	1.35		D	0.49	1.84	0.76	0.25	0.85	1.12	0.09	2.24	2.57	1.3	0.18	0.47
P4	0.35		1.29				2.37	0.61		0.81	1.12		2.53	2.50	1.25	0.17	
P5	0.37		1.42				1.93				1.12		2.02	2.31	1.53		
P6	0.43		1.32				2.22						2.29	2.43	1.68		
P7							2.05						2.2				
P8													2.29				
P9													2.02				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 4 RMSD standard deviation for Rubicon

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.447	0.067	0.348	0.217	D	0.303	0.422	0.455	0.369	0.337	0.288	0.149	0.642	0.385	0.264	0.834	0.300
P1[b]	0.337	0.047	0.296	0.215	D	0.267	0.357	0.493	0.327	0.333	0.330	0.182	0.607	0.482	0.249	0.754	0.252
P2	0.401	0.063	0.405	0.193	D	0.340	0.325	0.45	0.411	0.351	0.345	0.154	0.548	0.405	0.206	0.781	0.261
P3	0.424	0.046	0.294		D	0.367	0.35	0.519	0.350	0.307	0.321	0.181	0.533	0.348	0.290	0.853	0.253
P4	0.428		0.33				0.251	0.471		0.363	0.321		0.503	0.329	0.282	0.761	
P5	0.382		0.327				0.322				0.352		0.628	0.411	0.235		
P6	0.404		0.338				0.316						0.565	0.34	0.215		
P7							0.338						0.477				
P8													0.544				
P9													0.628				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 5 Minimum RMSD for Omega

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	1.82	0.3	1.52	0.61	D	0.42	1.7	0.53	0.68	0.8	1.3	0.24	2.41	3.41	1.06	0.39	0.3
P1[b]	1.26	0.52	1.73	0.68	D	0.68	0.76	0.74	0.45	0.95	1.3	0.24	1.72	2.1	1.18	0.39	0.3
P2	1.77	0.32	1.44	0.37	D	0.49	1.98	1.19	1	0.71	1.64	0.67	2.38	1.98	1.26	0.18	0.3
P3	0.86	0.29	1.72		D	0.68	1.96	0.61	0.45	1.06	1.26	0.24	2.28	1.97	1.04	0.46	0.3
P4	1.27		1.17				1.17	0.66		0.83	1.61		1.99	1.73	0.78	0.19	
P5	1.51		1.27				1.98				1.25		2.57	2.49	0.95		
P6	1.43		1.66				1.15						2.41	3.44	1.1		
P7							1.09						1.86				
P8													2.45				
P9													2.57				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

dramatic difference is observed when using Catalyst or Corina.

Table 9, A and B, show the RMSD values obtained comparing the crystal structure of 1DFR and different sets of conformers.

In detail, Table 8A shows that operating with default parameters and allowing OMEGA to produce as many conformers as it can, the output from the treatment of permuted input strings delivers conformers closer to the bioactive conformation when compared to those generated from the canonical string (280 final unique conformers were produced both form the canonical and the permuted string of 1DFR).

The graph in Fig. 3 shows a comparison between the potential energy of the conformers generated from the canonical and the permuted strings. It is quite clear

Table 6 RMSD standard deviation for Catalyst

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.445	0.179	0.596	0.285	0.346	0.413	0.519	0.476	0.382	0.419	0.209	A	0.670	0.806	0.377	0.768	0.278
P1[b]	0.487	0.177	0.586	0.264	0.346	0.413	0.519	0.476	0.382	0.470	0.209	A	0.598	0.806	0.377	0.768	0.278
P2	0.340	0.177	0.586	0.198	0.346	0.302	0.519	0.476	0.384	0.470	0.209	A	0.598	0.764	0.319	0.768	0.278
P3	0.483	0.177	0.626		0.346	0.413	0.519	0.476	0.382	0.439	0.209	A	0.598	0.764	0.319	0.768	0.278
P4	0.483		0.626				0.519	0.476		0.477	0.209		0.670	0.764	0.319	0.768	
P5	0.483		0.622				0.519				0.209		0.670	0.806	0.319		
P6	0.483		0.622				0.426						0.598	0.806	0.377		
P7							0.519						0.598				
P8													0.670				
P9													0.670				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 7 Minimum RMSD for Catalyst

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can[a]	0.53	0.42	0.99	0.06	0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	1.91	2.02	1.27	0.17	0.43
P1[b]	0.26	0.42	0.99	0.06	0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	2.18	2.02	1.27	0.17	0.43
P2	0.26	0.42	0.99	0.38	0.32	0.58	1.16	1.22	0.29	0.82	1.53	0.25	2.18	2.08	1.48	0.17	0.43
P3	0.26	0.42	1.35		0.32	0.57	1.16	1.22	0.72	0.82	1.53	0.25	2.18	2.08	1.48	0.17	1.31
P4	0.26		1.35				1.16	1.22		0.8	1.53		1.91	2.08	1.48	0.17	
P5	0.26		1.35				1.16				1.53		1.91	2.02	1.48		
P6	0.26		1.35				1.3						2.18	2.02	1.27		
P7							1.16						2.18				
P8													1.91				
P9													1.91				

[a]Can = canonical SMILES string

[b]'P1, P2, P3...' = different permutations of the canonical SMILES string

Table 8 Minimum RMSD for Corina

	1aq1	1d1q	1glq	1abe	1azm	1cbx	4dfr	1ebp	1hyt	1mrk	1p93	1phf	4phv	1poc	1i7i	1srj	1tpp
Can	1.14	0.76	2.38	0.63	1.31	0.56	2.57	1.35	0.91	1.09	1.47	0.46	3.78	1.82	4.15	0.16	0.42
P1	1.85	0.76	2.38	0.78	1.35	0.56	2.57	1.34	0.91	1.71	1.44	0.46	2.74	1.86	4.15	0.16	0.42
P2	1.82	0.76	2.41	0.73	1.35	1.52	2.58	1.46	1.49	1.71	1.46	0.46	2.75	2.24	4.15	0.16	0.42
P3	1.83	0.76	2.38		1.35	0.56	2.57	1.29	0.91	1.94	1.81	0.46	2.58	2.22	4.9	0.16	0.42
P4	1.83		2.38				1.98	1.02		1.96	1.22		2.05	2.24	4.9	0.16	
P5	1.82		2.41				1.97				0.56		2.59	2.19	4.15		
P6	1.81		2.41				1.75						2.75	1.82	4.15		
P7							1.97						2.48				
P8													2.05				
P9													2.05				

that the potential energy span produced by the conformers from the permuted input is wider then that originated from the canonical input.

To greater examine this phenomenon we undertook to represent the conformational space explored by the two output sets using PCA analysis. These data are shown in Fig. 4.

In red is depicted the conformational space location of the co-crystal ligand (the bioactive conformer) from 1DFR. Interesting, we note that the immediate conformational neighbourhood around the bioactive conformer predominately originates from the permuted string output. Moreover, the conformational space explored by the conformers from the canonic appears to be smaller and less diverse then that from permuted strings.

In Table 9 the unique conformers produced from the permuted strings are once again closer to the bio-

Table 9 RMSD minimum, maximum, average and standard deviation calculation for the ligand contained in PDB entry 1DFR: the top panel A shows the set where 280 conformers were produced; the panel below B shows the set where 2,239 and 2,326 conformers were produced for respectively the canonic (Can) and the permuted (Perm) SMILES strings

	Can	Perm
A		
Min	1.29	1
Max	3.44	3.44
Average	2.52921429	2.492357143
Dev st.	0.39585861	0.478173962
B		
Min	1.23	0.76
Max	3.47	3.44
Average	2.53093345	2.398963887
Dev st.	0.38276039	0.476240746

active conformer (RMSD 0.76 calculated for permuted strings against 1.23 from canonical), when comparing ensembles of similar size. From the canonical input, OMEGA produced 2,239 conformers. The set originating from the permuted input comprised 2,326 conformers.

As for set A, the potential energy has been calculated as previously described. In this case, the potential energy span originating from the permuted strings input is larger. Members of the ensemble are energetically much closer to the energy calculated for bioactive conformation then those members of the ensemble originating from the canonical input (Fig. 5).

Figure 6 shows the plotted result of the PCA. It depicts clearly how the conformational space explored by the conformers originating from permuted

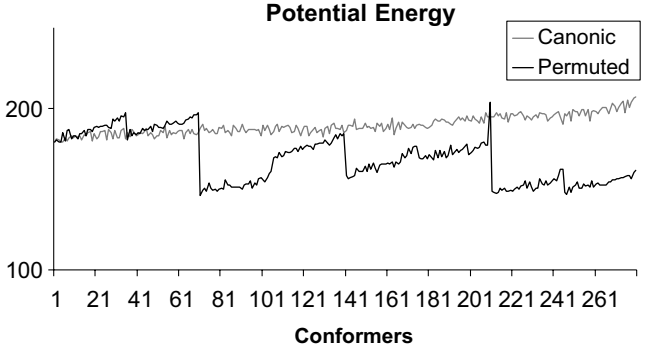

Fig. 3 Potential energy graph for conformer set A, generated using omega (280 members). Black line illustrates result for permuted input string, grey illustrates result for canonical

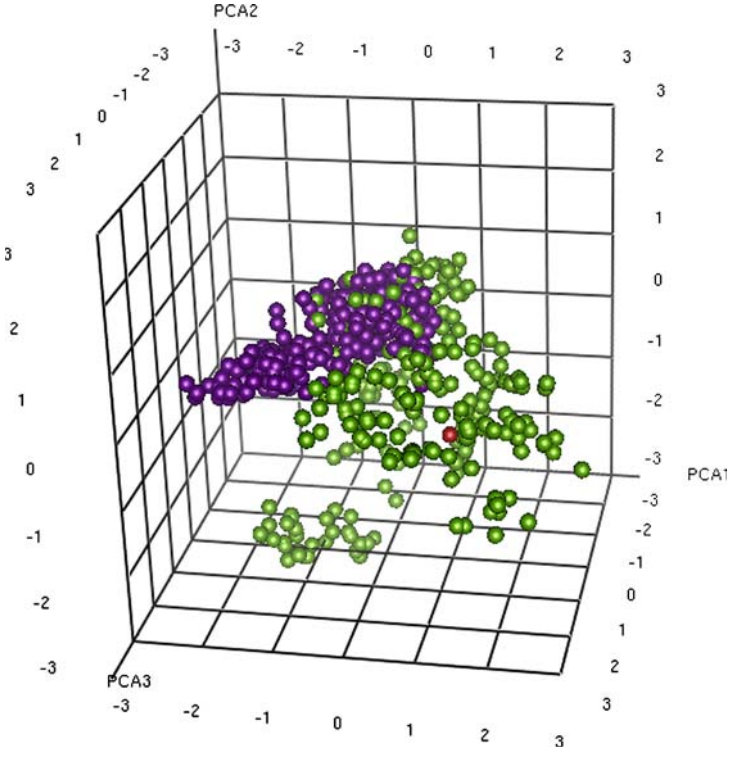

Fig. 4 PCA analysis of conformers generated by Omega: in violet are showed the 280 conformers produced form the canonical string of 1DFR. In green the conformers generated from 8 permutation of the same string. In red the crystal structure of 1DFR

input is diverse and closely surrounds the bioactive space.

Fingerprint analysis

A critical method for incorporating ligand flexibility in a vHTS (Virtual High Throughput Screening) process is to generate multiple conformers of each molecule. One of the key things to take into account at this stage in the process is the balance between the number of conformers generated and the diversity of each with regard to conformational space sampling. Generally, less conformers with a wider degree of RMSD separation between each is beneficial.

To investigate the uniqueness of the conformers generated from each permuted SMILES string and the software used, all the conformers were read into MOE and fingerprint analysis carried out using Esshape3D. By exploring the percentage of conformers that were identical in each set for a molecule it was possible to determine whether this was due to the orientation of each conformer. For example, it was possible that the ensemble of conformers derived from different SMILES permutations of 1AQ1, contained duplicates.

Figure 7 shows the result from the fingerprint analysis for Omega, Catalyst and Rubicon. The numbers of conformers produced by each software package were normalized to the one with the highest value. Thus, Rubicon (3,845) was 100%, Catalyst (3,623) 94.22% and Omega (2,406) 62.57%. The percentage of unique conformers generated from the permuted SMILES strings by Omega was 99.38% of the total. If the conformers were generated from only the canonical SMILES string, the percentage of the total would be just 11.33%. It is apparent at this stage that generating conformers using Omega with default settings is highly efficacious when combined with permuted SMILES strings. In the case of Rubicon, 87.73% of the total numbers of conformer were unique in conformational space.

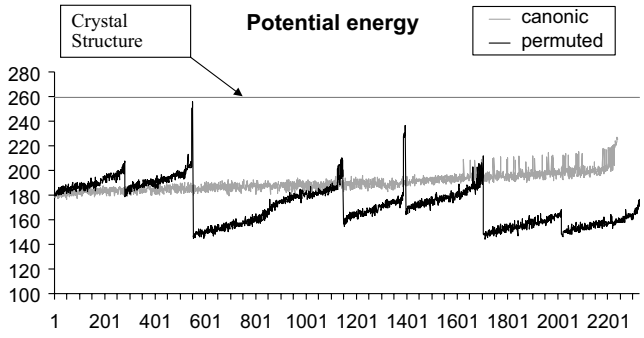

Fig. 5 Potential energy graph for conformer set B, generated using omega (ca. 2200 members). Black line illustrates result for permuted input string, grey illustrates result for canonical

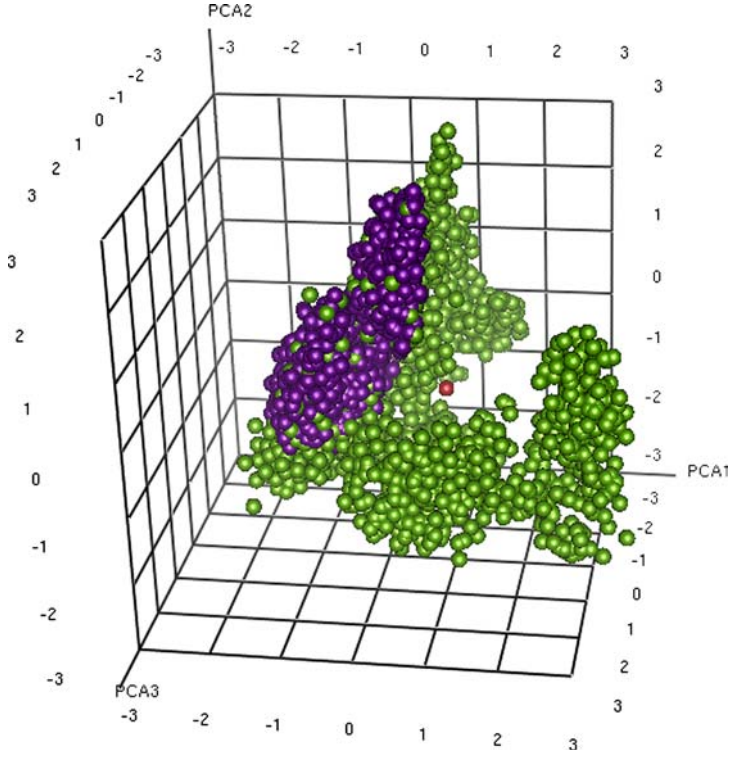

Fig. 6 Fingerprint analysis of conformers generated by Omega. In violet the conformers (about 1200) generated setting an energy window equal of 60 and a RMSD cut off of 0.5 from the single canonical string of the co-crystal ligand from 1DFR. In green the conformers generated from the 8 permutations of the 1DFR ligand keeping all the parameters in OMEGA as default (with the exception of –max_conf, that was set to 310 to generate a comparable number of final conformers with the canonic set. In red is showed the PCA coordinate calculated for the crystal structure

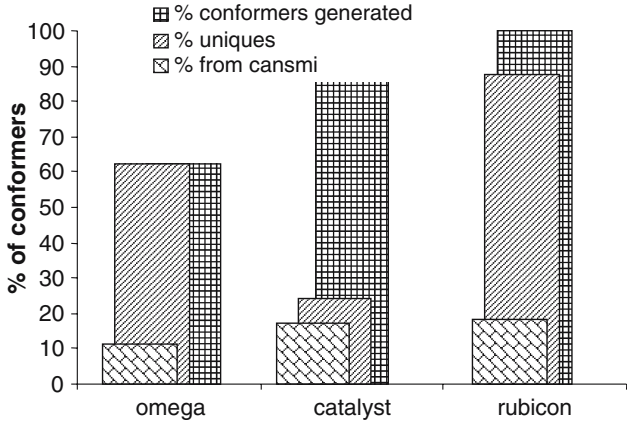

Fig. 7 Fingerprint analysis of conformers generated by Omega, Catalyst and Rubicon. The height of the 3 bar graphs is proportional to the total amount of conformers generated. Cansmi: Canonical SMILES string

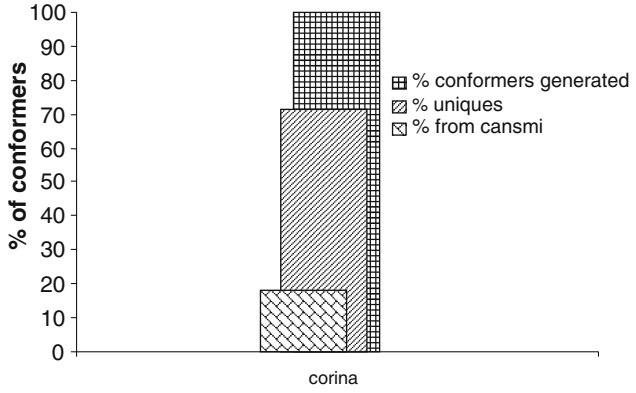

Fig. 8 Fingerprint analysis of conformers generated by Corina

Table 10 RMSD minimum, maximum, average and standard deviation calculation for 1DFR: the table shows the set where 280 conformers were produced using one connection table (one CT) and 10 permutation of a single connection table (Permuted)

1DFR	One CT	Permuted
Min	1.14	0.37736
Max	3.26	3.26
Average:	2.373614865	1.787744
Dev st.	0.377360288	1.280565

Again starting from a single SMILES string, only 18.11% would be unique. To a lesser extent, but still evident, was Catalyst with the percentage of unique conformers being 23.98% of the total. Beginning from a single SMILES string would allow 17.06% unique conformers to be produced.

Since Corina produced just one conformer per string (94 total amount = 100%) it is represented on its own in Fig. 8 so as to keep it to scale. The percentage of unique conformations using Corina was 71.28%. With only the canonical SMILES string it would have been 18.18% of the total.

Connection table analysis

Table 10 depicts the RMSD difference between the co-crystal structure conformer and the conformers produced using permuted connection table input. As the values suggest, the same result obtained above for the permutation of input SMILES strings is reproduced for the permutation of more complex (SDF) connection table input. As explored for the SMILES strings, different conformers were generated for the input sets (different standard deviation). Moreover, beginning from 10 unique input permutations, the 280 output conformers generated are closer to the crystal structure (average: 1.78, min: 0.37) than the 280 conformers originating from a single connection table (average: 2.37, min: 1.14).

Conclusion

In this study we have shown that permutation of input strings leads to the generation of different numbers of conformers and that some software programs are affected more than others by this depending on their innate algorithm. Depending on the software used the conformers generated from permuted SMILES strings are more diverse than if they were generated from a single representation. The differing amounts of conformations generated are not an inherent fault of the software packages used, rather it is an outcome of the input SMILES string and how the software deals with that SMILES string. When a SMILES string is permuted this possibly has an incidental effect on the order of the torsions which are subsequently used to generate the conformations.

Conformers produced from permuted SMILES strings exhibit a smaller RMSD value, when compared to the X-ray structure, than do conformers generated from the canonical SMILES string. Other work carried out by our group has shown the dramatic effect that this phenomenon has on Enrichment and False Positive rates. Thus, many validations of rigid docking/scoring combination platforms where the same dataset was used, but converted to SMILES strings using different software, may show better or worse performance as a direct result of this phenomenon and not due to the platforms utilised. We suggest that all validations of rigid docking/scoring platforms account for this effect prior to the docking step. Increasing the

number of conformers generated, may possibly reduce this effect.

Importantly, all of the software tested compares the generated conformers and discards those that are judged to be too closely related. The ultimate measure of this is an RMSD assessment. For example, one of the comparison steps in Omega involves applying a conformation RMSD exclusion threshold of 0.8 Å. If this value is reduced an increased amount of conformations will be accumulated but they will be more conformationally similar.

Conversely, when using permutation of the same SMILES string those conformers would not be discarded and thus some of the conformers obtained might have similar RMSD values.

Finally, we suggest a novel way of generating conformers when using Omega, Rubicon or Corina with their default settings. Permuted input strings, in conjunction with the canonical SMILES string, could be utilised to more rapidly and comprehensively sample conformational space of a molecule. This method extends to connection table representations of 2D molecular structure such as SDF. PCA analyses clearly illustrate superior conformational sampling arising from use of permuted input. To ensure that the final result does not include any duplicates a quick fingerprint analysis can be carried out as we have shown. We also propose that future versions of software employed in the generation of conformers from SMILES strings could include a switchable flag allowing the canonicalisation of the SMILES strings before conformer generation. The same flag, when off, would allow the users to generate multiple molecular conformations starting from permutations of the same SMILES string.

The effect is also observed using more complex molecular connection tables rather than SMILES strings. To this end, the same method described above may be used for the efficient generation of a diverse unique set of conformers. Clearly, SMILES notation presents a far more limited set of possible atom/bond orderings unless the use of dot disconnections or closure bonds are employed. Thus, exploiting SD or indeed any format incorporating a connection table may be used to generate (#atoms!)*(#bonds!) possible orderings, resulting in efficient sampling of conformational space. Finally, although not currently implemented in any software, canonicalisation of connection tables would also be useful in the validation process of vHTS tools.

Acknowledgement This work was supported through funding from Science Foundation Ireland and the Irish Health Research Board.

References

1. Hou T, Xu X (2004) Curr Pharm Des 10(9):1011
2. Liao C, Liu B, Shi L, Zhou J, Lu XP (2005) Eur J Med Chem 40(7):632
3. Bringmann BKA (2004) Frequent SMILES. Lernen, Wissensentdeckung und Adaptivität, Workshop GI Fachgruppe Maschinelles Lernen, part of LWA 2004
4. Weininger D (1988) J Chem Inf Comput 28:31
5. Knox AJS, Meegan MJ, Carta G, Lloyd DG (2005) J Chem Inf Model 45(6):1908–19
6. Vigers GP, Rizzi JP (2004) J Med Chem 47(1):80–89
7. Kauppi B, Jakob C, Farnegardh M, Yang J, Ahola H, Alarcon M, Calles K, Engstrom O, Harlan J, Muchmore S, Ramqvist AK, Thorell S, Ohman L, Greer J, Gustafsson JA, Carlstedt-Duke J, Carlquist M (2003) J Biol Chem 278(25):22748
8. Cronet P, Petersen JF, Folmer R, Blomberg N, Sjoblom K, Karlsson U, Lindstedt EL, Bamberg K (2001) Structure (Camb) 9(8):699
9. Daylight Chemical Informations Systems Inc. (URL: http://www.daylight.com)
10. Molecular Operating Environment (MOE), developed and distributed by Chemical Computing Group (http://www-chemcompcom)
11. Babel v2.0A3, distributed by Openeye Scientific Software
12. Chemsketch v8.17, www.acdlabs.com
13. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261(3):470
14. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267(3):727
15. CORINA 3.6, distributed by Molecular Networks GmbH
16. Cambridge Structural Database, http://www.ccdc.cam.ac.uk/
17. OMEGA 1.8.1, distributed by Openeye Scientific Software
18. Catalyst v4.9.1, www.accelrys.com
19. RUBICON, distributed by Daylight Chemical Informations Systems Inc
20. Shanno DF, Phua KH (1980) ACM Trans Math Software 6:618
21. Bostrom J, Greenwood JR, Gottfries J (2003) J Mol Graph Model 21(5):449
22. Bostrom J (2001) J Comput Aided Mol Des 15(12):1137
23. Ivanciuc O (2003) In: Gasteiger J (ed) Handbook of chemoinformatic , vol. 1. p 103
24. Kudo Y, Sasaki S (1974) J Chem Document 14(4):200
25. Babel 1.100.2, Distributed by Openeye Scientific Software
26. oechem, RMSD, Distributed by Openeye Scientific Software