ChemT, an open-source software for building template-based chemical libraries

Rui M. V. Abreu[1,2,*], Hugo J. C. Froufe[1], Pedro O.M. Daniel[2], Maria-João R. P. Queiroz[3] and Isabel C. F. R. Ferreira[1]

Address:

[1]*CIMO-ESA, Instituto Politécnico de Bragança, Campus de Sta Apolónia, Apartado 1172, 5301-855 Bragança, Portugal.*

[2]*Instituto de Biotecnologia e Bioengenharia, Centro de Genómica e Biotecnologia, Universidade de Trás-os-Montes e Alto Douro (CGB-UTAD/IBB), Apartado 1013, 5001-801, Vila Real, Portugal.*

[3]*Centro de Química, Universidade do Minho, Campus de Gualtar 4710-057 Braga, Portugal.*

*corresponding author

Rui MV Abreu (e-mail: ruiabreu@ipb.pt; tel.: +351 273303219; fax: +351 273325405);

ABSTRACT

In computational chemistry vast quantities of compounds are generated, and there is a need for cheminformatic tools to efficiently build chemical compound libraries. Several software tools for drawing and editing compounds structures are available, but they lack options for automatic generation of chemical libraries. We have implemented ChemT, an easy-to-use open-source software tool that automates the process of preparing custom-made template-based chemical libraries. ChemT automatically generates three-dimensional chemical libraries by inputting a chemical template and the functional groups of interest. The graphical user interface of ChemT is self-explanatory, and a complete tutorial is provided. Several file formats are accepted by ChemT, and it is possible to filter the generated compounds according to different physicochemical properties. The compounds can be subject to force field minimization, and the resulting three-dimensional structures recorded on commonly used file formats. ChemT may be a valuable tool for investigators interested in using *in silico* virtual screening tools, like QSAR modelling or molecular docking, in order to prioritize compounds for further chemical synthesis. To demonstrate the usefulness of ChemT, we describe an example based on a thieno[3,2-*b*]pyridine template. ChemT is available free of charge from our website at http://www.esa.ipb.pt/~ruiabreu/chemt.

Keywords: ChemT; QSAR; chemical library; virtual screening; drug discovery.

1. Introduction

Virtual Screening (VS) of chemical compound libraries, using tools like QSAR (Quantitative Structure Activity Relationships) modelling or molecular docking, has become an important step in the initial stages of the drug discovery process [1].

For an efficient use of these VS tools, the generation of vast quantities of small chemical compounds is many times needed. Cheminformatic tools that efficiently create, manage and examine huge chemical compound libraries are therefore of primordial importance [2].

Several virtual databases are available for VS. Some are freely available including: the DTP depository that provides structural information of over 250 thousand structures [3], ZINC database with over 13 million purchasable compounds [4] and Pubchem with over 31 million of pure and chemically characterized compounds [5], among others. The "iResearch Library" is a more comprehensive database, although only commercially available, with over 95 millions compounds that may be purchased from international chemistry suppliers [6]. These databases of millions of compounds have inspired the development of a different type of databases containing not yet synthesized and characterized compounds. Included on this category are GDB-11, GDB-13 and GDB-15 databases that, with certain atom type and atom number constrains, try to enumerate the ensemble of all possible compounds in the chemical space [7]. GDB-15 in currently the largest database with 28.8 billions compounds.

These databases are good for initial large-scale VS campaigns, particularly when there is still a lack of information on the inhibition mechanism for a given protein target. Lead candidates that perform better on VS are usually viewed as starting points (templates) for synthesis of more interesting compounds. On this stage, designing focused specific chemical compound libraries, using the lead candidates as templates, is frequently required. These libraries can then be used for further virtual testing, before committing resources to chemical synthesis and experimental validation [8]. A number of free or commercial tools are available for generating and editing small

chemical compounds, but they usually work in a one structure at a time approach, were each structure is manually drawn and individually recorded on a file. These tools include ChemDraw [9], Symyx Draw [10], ChemSketch [11] or Corina [12] among others. Automated structure generation in these software tools is usually confined to copy-paste tasks. There are a number of automated generators of combinatorial libraries included in different commercial software products related to drug design, although with a high price tag. Some use a *de novo* approach, were there is a prior knowledge of the structure of a given target, and the software forms the compound structures directly on the receptor binding site [13-16]. Other generators build libraries around a central fragment or scaffold with several substitutions sites by combining a number of fragments [17-19].

We introduce ChemT (Chemical Templates), a free open-source software tool that can automatically generate three-dimensional (3D) VS-ready chemical compound libraries, by inputting a chemical template structure with the functionalization positions and the functional groups of interest (Fig. 1). In this paper we describe a ChemT application example, using a thieno[3,2-*b*]pyridine template with 4 functionalization positions and 10 possible functional groups (Fig. 2a). In this example a total of 10000 combinations were generated by ChemT and, after filtering the compounds based on physicochemical properties, 18 compounds remained. A 3D thieno[3,2-*b*]pyridine library was finally generated (Fig. 2b). A research project involving similar compounds, and posterior virtual screening with the generated library, was our motivation to develop ChemT.

2. Methodology

ChemT was developed using C-sharp language and compiled for Windows using SharpDevelop version 3.5 [20]. For file format conversions, properties calculation and compound energy minimization, ChemT uses the OpenBabel OBDotNet library (version 0.4) [21]. OBDotNet is a compiled assembly that allows OpenBabel to be used from the various .NET languages, including the C-sharp language, used to develop ChemT.

For compound energy minimization, ChemT uses the Universal Force Field (UFF) available with OpenBabel [22]. The Force fields (including UFF) supported by OpenBabel are not yet compiled as separate loadable libraries so ChemT requires separate installation of OpenBabel. As this feature is on the road-map for the next OpenBabel version, it is expected that next versions of ChemT will not require prior OpenBabel installation. Also, for image rendering and 2D representation of the template molecule, the Indigo chemistry toolkit .NET library was used [23].

The Microsoft .NET Framework (version 4+) is required for using ChemT and, if not already installed, the user will be automatically prompted to download and install it. The Microsoft .NET Framework provides pre-coded solutions to common software development requirements, and manages the execution of applications written for the framework. The deployment size of ChemT is small, since it's executed in the Microsoft .NET Framework runtime environment. ChemT is compatible and was tested with different versions of Windows (XP, Vista and 7).

3. Results and Discussion

Particular attention was paid to ChemT easiness of use, with a GUI (Graphical User Interface) that has a logical working flow and fairly self-explanatory sections and buttons (Fig. 1). A user with no prior knowledge on the intricacies of file formats and force fields can generate a chemical compound library, according to their needs, in a relatively short amount of time. Here we present a description of the main features of ChemT, as well as an example of use.

In the example presented, we set out to generate a chemical compound library of potential new antitumor agents, using a thieno[3,2-*b*]pyridine scaffold as a chemical template. This library can then be subject to Virtual Screening in order to assess the interest in synthesizing the compounds with best results.

3.1. Input section

The first step for using ChemT is to draw, on a suitable program like ChemSketch or ChemDraw, the library template structure by placing an R group on the functionalization positions of interest (Fig. 2a). This template is drawn in the standard 2D nomenclature and can be saved on commonly used file formats (MOL, MOL2, PDB and SDF). Then, the file is selected on the 'template' section and a 2D representation of the template is shown. Each R position is given a different number (Fig. 1). For each numbered R position, specific functional groups (R groups) of interest can then be manually written on the 'Functional groups (R)' section in SMILES format [24]. For convenience a list with the chemical name, SMILES string and 2D representation of the most common functional groups is available on ChemT website. This feature relieves the user from extensive knowledge in SMILES format specifics. ChemT users, especially chemists, will sometimes work with a specific set of functional groups. With this in mind, an option is available for saving a text file with a list of the functional groups of interest. This file can be recalled automatically at any time, to generate more compound libraries.

In our example of a thieno[3,2-*b*]pyridine library, we assigned 4 possible functionalization positions, as indicated by the R groups (Fig. 2a), and a total of 10 functional groups were introduced (Fig. 1). The user should be careful with blank lines as they will be considered a functional group and should use an explicit H letter for R=H. The example files can be introduced by simply selecting the ‘File>Example’ menu or can be downloaded at the ChemT website.

3.2. ChemT algorithm

Clicking the 'Generate Combinations' button initiates the ChemT algorithm that, using a dynamic matrix, finds all the possible chemical permutations, each representing a different chemical compound. All the permutations are then presented in SMILES format on the 'Template and Functional Groups' section (Fig. 1), and the number of combinations is presented. This step is very fast and a SMILES list library is built with an average speed of about 1000 compounds per second,

depending on the computer used. In our thieno[3,2-*b*]pyridine library example, the use of 4 possible functionalization positions and 10 functional groups generated a total of 10000 combinations, each represented by a SMILES string.

When generating all combinations, it's important to note that ChemT uses the first atom of the functional group SMILES string, as the one that will be covalently linked to the functionalization positions (R positions) of the template structure. This is illustrated by the first three functional groups used in the example (Fig. 1), that represent three possible pyridine rings linked by a C-C bond in the: 2 (c1Ncccc1; SMILES notation), 3 (c1cNccc1) and 4 (c1ccNcc1) positions, respectively.

3.3. Filter Application

Many VS tools are computer intensive, and processing thousands or hundreds of thousands of compounds can be a time consuming task. To help in making an intelligent selection of the compounds that will be generated in our final chemical library, several filters can be applied at this stage. The user can apply directly a Lipinski's Rule of Five (LRF) filter that sets limits for different physicochemical properties: molecular weight (MW), partition coefficient (LogP), number of hydrogen bond donors (HBD) or acceptors (HBA) [25]. With the LRF filter, the compounds are required to meet 3 out of 4 properties limits: between 160 and 500 Dalton for MW, between -0.4 and 5.6 for LogP, no more than 5 HBD and no more than 10 HBA. The LRF was proposed after a study where it was observed that drugs reaching Phase II clinical trials usually tended to have values for MW, LogP, HBD and HBA, falling within certain limits [25]. The default values for MW and LogP limits were selected according to Goose *et al.* [26]. Still, because there are several versions of LRF with different property limits, the user is able to change directly the limit values for each property. An option to change the default values of the compound property limits is available on the Settings menu. Also, ChemT is flexible enough to filter the compounds using limits for only

one property, or for a selected set of properties. Filter application is fast, with the LRF filter being the slowest because it calculates values for all 4 properties. Depending on the computer used, our tests showed that the LRF filter achieves average speeds of about 40 compounds per second. Applying a single property filter is much faster. Per example, the MW filter alone is applied at speeds of about 250 compounds per second. On this stage, any unwanted chemical compound can be removed by left-clicking the corresponding line in the SMILES list box. Also, by clicking the 'Save' button a text file with the library of compounds can be saved in SMILES format. ChemT can also be used to convert and filter compound libraries from other sources (provided they are available in SMILES format). Per example the GDB-11, GDB-13 and GDB-15 databases are supplied in SMILES format and can be inputted on ChemT using the 'Open' button. We successfully tested several subsets of the GDB-13 database using ChemT [27].

When applying the default limits of the LRF filter to our thieno[3,2-*b*]pyridine library example, a total of 7827 compounds were selected out of the original 10000 compounds. To refine the filtering process the LRF filter was unselected and each individual filter of the 4 properties was selected. This dropped the number of compounds from 7827 to 1035. This reduction occurs because compounds must respect the 4 property filter limits, and not just 3 out of 4 property filters limits as with the LRF filter. Then, by restricting the MW, LogP and HBD limits to 450-500, 0-5 and 5-5, respectively, we ended up with a more manageable 18 chemical compounds library (Fig. 2b). It's important to note that the filters are applied to the compounds shown on the 'Template and Functional Groups Combinations' field. This means that a filter is applied on top of the last filter results.

3.4. Library generation

After filter application, the next step is to select the output file format and generate a 2D or 3D chemical compound library. On the 'output file' field the user can select the folder, name file and

one of the file formats available: MOL, MOL2, PDB and SDF. Finally, the generation of a compound library is started by clicking the 'Generate Library' button. Building a 3D library is considerably longer due to the minimization process when applying the UFF force field. Our tests showed that ChemT generates, on average, 5 compounds per minute for a 3D library and 40 compounds per second for a 2D library. Nevertheless, for a 3D library, speed is largely dependent on size and number of rotatable bonds of the compounds. Also the 64-bit version of ChemT can be used on compatible computers and will drastically increase library generation speed.

By default, one file with all the 3D compounds structures is generated. This is a normal option for VS projects, especially when using the SDF format. Still an alternative is available for saving each compound in individual files by simply selecting the 'Multiple files' tick box. The 3D thieno[3,2-*b*]pyridine library example was generated and is shown as a superimposition of the 18 compounds (Fig. 2b).

ChemT is available for free for academic researchers under an open-source license. This will eventually allow the structure-based drug design community to independently create and add improvements to ChemT. A detailed tutorial for ChemT is available on the Help menu.

3.5. Future plans

Planned features in future versions of ChemT include: (1) an integrated tool for drawing the template structures and the functional groups without the need of third-party software, (2) inclusion of more filters, (3) the ability to predict changeable positions/synthetically-accessible places to add the R groups and (4) a ChemT version for Linux operating systems.

4. Conclusions

ChemT automatically generates 2D or 3D chemical compounds libraries, based on a user-defined chemical template, in an easy-to-use approach and with limited knowledge needed on the

nuances of chemical file formats. Several options are available including: LRF filters, individual property filters and multiple files option for saving compounds. Furthermore, ChemT supports file formats more frequently used in VS projects using tools like molecular docking and 3D-QSAR. ChemT can be a very useful tool in the drug discovery process, by automatically generating compound libraries, which can be virtually evaluated using different VS tools. The results of this virtual evaluation can then be used to guide and prioritize chemical synthesis of new compounds.

Acknowledgements

The authors are grateful to Foundation for Science and Technology (Portugal) and COMPETE/QREN/UE for financial support through research project PTDC/QUI-QUI/111060/2009, and Rui M.V. Abreu thanks Foundation for Science and Technology (Portugal) for the SFRH/PROTEC/49450/2009 grant.

References

[1] G. Pujadas, M. Vaque, A. Ardevol, C. Blade, M.J. Salvado, M. Blay, J. Fernandez-Larrea and L. Arola, *Protein-ligand Docking: A Review of Recent Advances and Future Perspectives*, Curr. Pharmaceut. Anal. 4 (2008), pp. 1-19.

[2] C.M. Song, S.J. Lim and J.C. Tong, *Recent advances in computer-aided drug design,* Brief. Bioinform. 10 (2009), pp. 579-591.

[3] DTP (Developmental Therapeutics Program of the National Cancer Institute) repository: http://dtp.nci.nih.gov/.

[4] J.J. Irwin, and B.K. Shoichet, *ZINC - a free database of commercially available compounds for virtual screening*, J. Chem. Inf. Model. 45 (2005), pp. 177-182.

[5] E.E. Bolton, *PubChem: Integrated Platform of Small Molecules and Biological Activities*, Annu. Rep. Comput. Chem. 4 (2008), pp. 217–241.

[6] iResearch library database: http://www.chemnavigator.com/.

[7] J.-L. Reymond, R. Deursen, L.C. Blum and L. Ruddigkeit, *Chemical space as a source for new drugs*, Med. Chem. Commun. 1 (2010), pp. 30–38.

[8] S. Ghosh, A. Nie, J. An and Z. Huang, *Structure-based virtual screening of chemical libraries for drug discovery,* Curr. Opin. Chem. Biol. 10 (2006), pp. 194-202.

[9] N. Mills, *ChemDraw Ultra 10.0*, J. Am. Chem. Soc. 128 (2006), pp. 13649-13650.

[10] Symyx Draw software: http://www.symyx.com.

[11] ChemSketch software: http://www.acdlabs.com/download/.

[12] Corina software: http://www.molecular-networks.com/.

[13] J.R. Fischer, U. Lessel, and M. Rarey, *LoFT: Similarity-Driven Multiobjective Focused Library Design,* J. Chem. Inf. Model. 50 (2010) pp. 1-21.

[14] P.S. Kutchukian, D. Lou and E.I. Shakhnovich, *FOG: Fragment Optimized Growth algorithm for the de novo generation of molecules occupying drug like chemical space,* J. Chem. Inf. Model. 49 (2009), pp. 1630-42.

[15] B.C. Pearce, D.R. Langley and J. Kang, *E-novo: an automated workflow for efficient structure based lead optimization*, J. Chem. Inf. Model. 49 (2009), pp.1797-1809.

[16] E.W. Lameijer, J.N. Kok, T. Bäck and A.P. Ijzerman, *The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules*. J. Chem. Inf. Model. 46 (2006), pp. 545-52.

[17] SYBYL software package, Tripos Inc.: http://tripos.com/.

[18] Cerius2 Version 4.10, Accelrys, Inc.: http://accelrys.com/.

[19] MOE software, Chemical Computing Group Inc.: http://www.chemcomp.com/.

[20] SharpDevelop software: http://sharpdevelop.com/.

[21] R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner and E.L. Willighagen, *The Blue Obelisk-interoperability in chemical informatics*, J. Chem. Inf. Model. 46 (2006), pp. 991-998.

[22] C.J. Casewit, K.S. Colwell and A.K. Rappe, *Application of a universal force field to main group compounds*, J. Am. Chem. Soc. 114 (1992), pp. 10046-10053.

[23] Indigo chemistry toolkit: http://ggasoftware.com/opensource/indigo.

[24] D. Weininger, *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*, J. Chem. Inf. Comput. Sci. 28 (1988), pp. 31-36.

[25] C.A. Lipinski, F. Lombardo, B.W. Dominy and P.J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Del. Rev. 23 (1997), pp. 3-25.

[26] A.K. Ghose, V.N. Viswanadhan and J.J. Wendoloski, *A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases*, J. Comb. Chem. 1 (1999), pp. 55-68.

[27] L.C. Blum and J-L Reymond, *970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13*, J. Am. Chem. Soc. 131 (2009), pp. 8732–8733.

[28] R.C. Calhelha and M.J.R.P. Queiroz, *Synthesis of new thieno[3,2-b]pyridine derivatives by palladium-catalyzed couplings and intramolecular cyclizations*, Tetrahedron Lett. 51 (2010), pp. 281-283.

[29] The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC: http://pymol.org/.

Figures

Fig. 1 Main panel of ChemT. Top left the template input field section. Bottom left the functional groups input section where the functional groups of interest can be introduced. Top right the Template and Functional groups combinations section, where all compound combinations are shown in SMILES format and where filters can be applied. Bottom right the output section, were library output file format, 2D or 3D option and the multiple file option can be selected.

Fig. 2 (a) Example of a thieno[3,2-*b*]pyridine chemical template with 4 functionalization positions of interest indicated by R groups [28]; (b) superimposition of 18 compounds generated in the thieno[3,2-*b*]pyridine library example. Figure 2a was drawn using ChemSketch [11] and figure 2b was prepared using Pymol software [29].