Journal of Molecular Graphics and Modelling 27 (2009) 578–583

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

CLEVER: Pipeline for designing *in silico* chemical libraries

Chun Meng Song [a], Paul H. Bernardo [b], Christina L.L. Chai [b], Joo Chuan Tong [a,*]

[a] Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632, Singapore
[b] Institute of Chemical and Engineering Sciences, 1 Pesak Road, Jurong Island, Singapore 627833, Singapore

ARTICLE INFO

Article history:
Received 30 July 2008
Received in revised form 15 September 2008
Accepted 16 September 2008
Available online 26 September 2008

Keywords:
Virtual combinatorial library
Markush technique
Compound analysis
Chemoinformatics
Chemistry

ABSTRACT

Advances in virtual screening have created new channels for expediting the process of discovering novel drugs. Of particular relevance and interest are *in silico* techniques that enable the enumeration of combinatorial chemical libraries, generation of 3D coordinates and assessment of their propensity for drug-likeness. In a bid to provide an integrated pipeline that encompasses the common components functional for designing, managing and analyzing combinatorial chemical libraries, we describe a platform-independent, standalone Java application entitled CLEVER (Chemical Library Editing, Visualizing and Enumerating Resource). CLEVER supports chemical library creation and manipulation, combinatorial chemical library enumeration using user-specified chemical components, chemical format conversion and visualization, as well as chemical compounds analysis and filtration with respect to drug-likeness, lead-likeness and fragment-likeness based on the physicochemical properties computed from the derived molecules. Also provided is an integrated property-based graphing component that visually depicts the diversity, coverage and distribution of selected compound collections. When deployed in conjunction with large-scale virtual screening campaigns, CLEVER can offer insights into what chemical compounds to synthesize, and more importantly, what not to synthesize. The software is available at http://datam.i2r.a-star.edu.sg/clever/.

1. Introduction

Combinatorial chemistry has become an integral component of the modern drug discovery pipeline [1,2]. Through the introduction of new chemical reactions and commercially available reagents, the size of these libraries has been increasing rapidly over the past few years [3]. Often, such libraries are far too large to be synthesized and screened in their entirety. Additionally, the derived compounds are often redundant in that they exhibit similar physicochemical characteristics. Therefore, a rational approach for combinatorial library design is desirable in order to maximize the outcome of an expensive synthesis and screening campaign [4].

To this end, computational tools have been used for designing combinatorial chemical libraries [3,5]. Virtual library design typically begins with the explicit enumeration of all molecular variants derived from an initial set of chemical "building blocks", followed by subsetting to allow good sampling of all products in the library [6]. Two approaches [3] are available for enumerating molecular variants. Product-based methods for library design, also known as Markush techniques, enumerate libraries by attaching a list of alternative functional groups to variable sites defined on a common scaffold [7]. Reaction-based methods, on the other hand, specify parts of the reacting molecules that undergo chemical transformations and the nature of these transformations. While product-based methods are capable of introducing diversity into the derived libraries simply by varying the functional groups to be attached, they may quickly become prohibitive with full enumeration of compounds, even for moderately sized building block collections. Reaction-based methods tend to provide a systematic roadmap by which the chemical products may be obtained through the use of various reagents. However, one drawback of such techniques lies in the fact that the derived libraries are often smaller, thereby providing less diversity within the available chemical space.

Once the chemical libraries are generated, they may be further optimized for molecular diversity or similarity using descriptors such as chemical composition, chemical topology, three-dimensional structures, functionality or drug-likeness through heuristic rules that detect ADME/Tox (Absorption, Distribution, Metabolism, Excretion and Toxicity) deficiencies [8]. In a quest for advancements in the design of virtual combinatorial libraries, several techniques have been developed [5,9–11]. Conceptualized with specific considerations, each of these tools can be applied to different aspects of

* Corresponding author at: Molecular Design Group, Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632, Singapore. Tel.: +65 6408 2156.
E-mail address: jctong@i2r.a-star.edu.sg (J.C. Tong).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.09.009

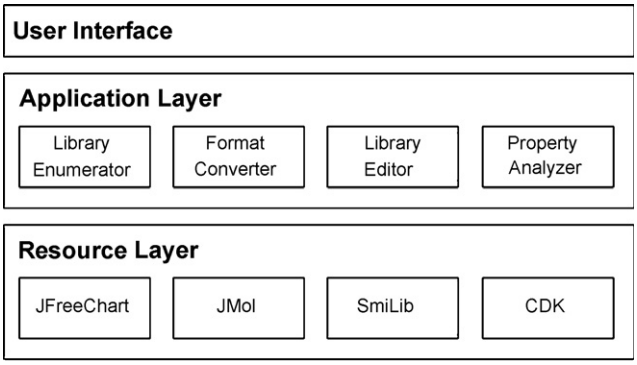


Fig. 1. CLEVER architecture diagram.

combinatorial library design such as library enumeration, format conversion and identification of library subsets.

In an endeavor to consolidate and enhance the functionalities required for designing combinatorial chemical libraries, we have developed CLEVER (Chemical Library Editing, Visualizing and Enumerating Resource) as a user-friendly platform for the design, management and analysis of compound libraries. By systematically evolving only the configurations of essential functional groups, CLEVER is particularly effective for generating libraries based on scaffolds that are already known. Among others, examples of such application areas include lead-optimization, where functional groups are varied in search of better receptor–ligand fits, and the design of prodrugs, where functional groups are modified to improve ADME/Tox or physicochemical properties.

Even though CLEVER adopts a product-based methodology towards library enumeration, it is designed to reduce the typical shortcomings associated with this approach. While full enumeration is an available option, CLEVER also allows for flexible enumeration configurations for each individual functional group of the specified scaffolds. CLEVER can further reduce the size of resulting libraries by computing the physicochemical properties of derived chemical compounds and filtering off unlikely candidates according to customized or predefined schemes such as Lipinski's rule-of-five [12] for drug-likeness. The ability to remove unlikely candidates early in the virtual screening pipeline can lead to significant savings as the number of late-stage failures is reduced [13]. To aid assessment of the diversity, coverage and distribution of chemical libraries, CLEVER is capable of generating histograms and 2D scatter plots using the physicochemical properties of the chemical collections. Finally, in preparation for subsequent docking or screening studies, CLEVER can also convert each molecule found in chemical collections into their respective 3D structure data files (SDF). In essence, CLEVER provides a comprehensive suite of tools that are effective for designing and refining chemical libraries *in silico*.

2. Methodology

CLEVER, which is implemented using the Java 3D Application Programming Interface (API), consists of five key modules. Fig. 1 shows the system architecture of CLEVER, consisting of five modules for chemical library editing, enumeration, conversion, visualization and analysis.

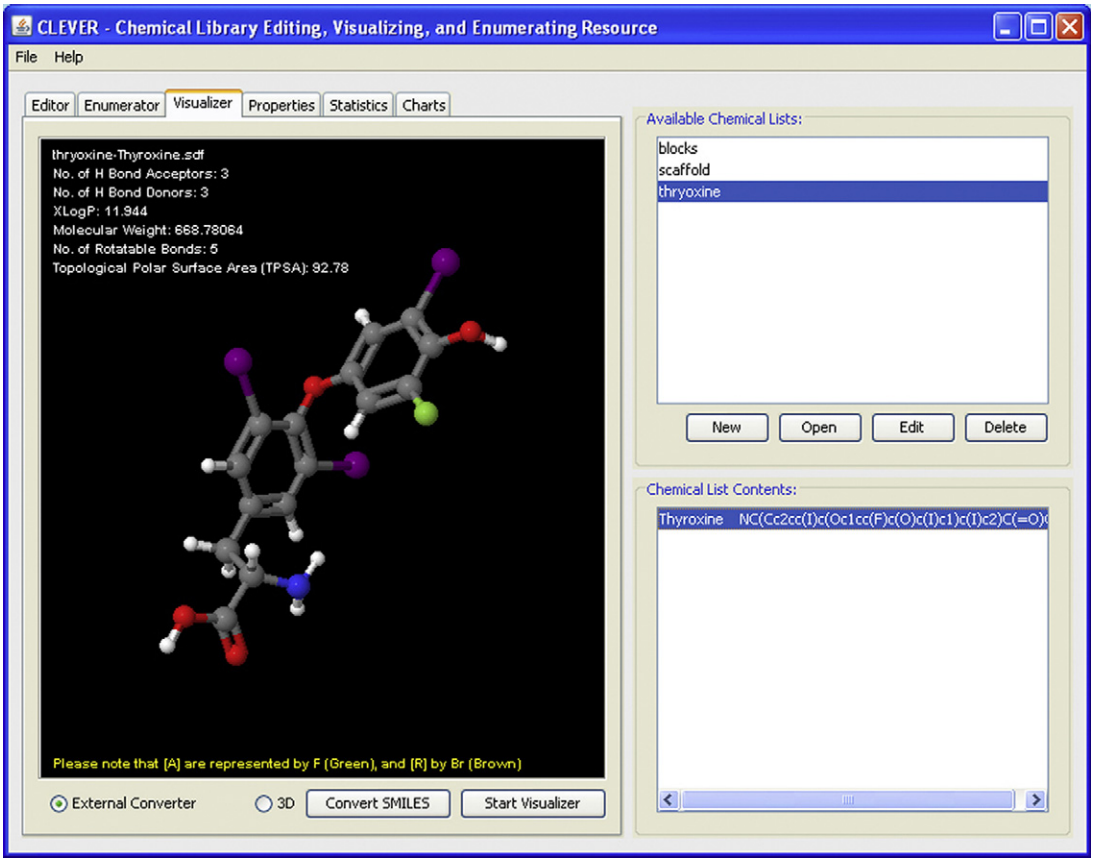


Fig. 2. CLEVER as a chemical visualizer. Display of the Thyroxin molecule in 3D space, together with various physicochemical properties computed.

Table 1
SMILES string configuration for scaffold and building blocks.

Type	Identifier	SMILES
Scaffold	S1	NC(Cc2cc([R1])c(Oc1ccc(O)c([R2])c1)c([R3])c2)C(═O)O
Block	B1	C[A]
Block	B2	C(C)(C)([A])
Block	B3	F[A]
Block	B4	CC[A]
Block	B5	C═C[A]
Block	B6	C1CCCCCC1([A])
Block	B7	C2ncc1[nH]cnc1n2([A])

Table 2
Possible reaction schemes and the respective sizes of derived libraries.

Reaction scheme definition							Size of derived library
1	1	1–7	1	1–7	1	1–7	343
1	1	1–3	1	1–2; 4	1	1–2; 5	27
1	1	1–2	1	3–6	1	7	8

Reaction schemes are defined for each scaffold (specified in the first column), followed by pairs of entries that stipulate the linker and blocks to be used for each of the attachment sites. Therefore, columns 2 and 3 define the linker and blocks to be used for the first attachment point, columns 4 and 5 for second attachment point, etc.

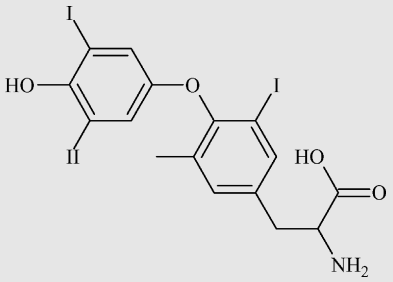
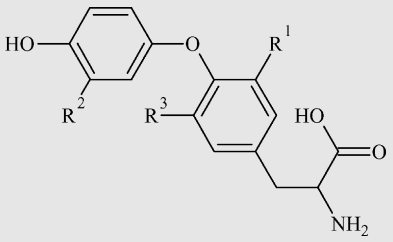
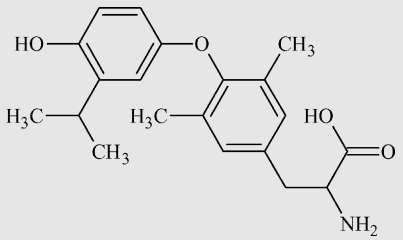
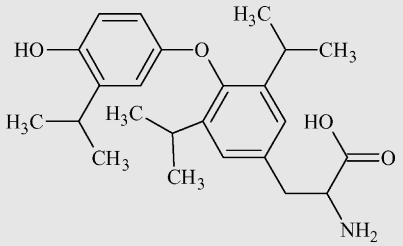
2.1. Chemical library editor

The chemical library editor is a tool for creating and manipulating chemical libraries, which include scaffolds and attachment blocks. The selected notation for chemical compounds is the Simplified Molecular Input Line Entry Specification (SMILES) [14], which represents chemical compounds in linear textual forms. Library files are essentially plain text files that contain a record on each line, with an entry identifier and the actual SMILES string delimited by a tab character.

2.2. Chemical library enumerator

The combinatorial library enumerator allows for rapid combinatorial library generation in SMILES format. Using SmiLib2 [9] as the enumeration engine, CLEVER enables users to design new combinatorial libraries by specifying the chemical scaffolds, attachment blocks, linkers and reaction schemes. In accordance with the SMILES format required by SmiLib2, attachment points on blocks are represented by "[A]" and functional groups to be permutated on the scaffolds are depicted by "[R*n*]", where *n* is a

Table 3
Enumerated molecules and resemblance to existing molecules.

Molecule	Structure	Description
Thyroxine		Original molecule, PubChem CID: 853
S1		Scaffold, with attachment points R^1, R^2 and R^3
S1.1_B1.1_B2.1_B1		Derived, corresponds to PubChem CID: 108112
S1.1_B2.1_B2.1_B2		Derived, corresponds to PubChem CID: 21295179

numerical value unique to each functional group to be varied. Once the required parameters are supplied, CLEVER automatically generates the settings required and invokes SmiLib2 execution accordingly. The resultant library is then added back into the available chemical list.

2.3. Chemical format converter

The chemical format converter is responsible for transforming linear SMILES strings into SDFs bearing 3D coordinates. Although the conversion can be performed using the Chemistry Development Kit (CDK) APIs [15], CLEVER is also configurable to harness the prowess of commercially available and open-source converters by interfacing with external programs such as CORINA [16] and OpenBabel [10]. The generated structure data files are saved to disk for future analysis and molecular docking studies.

2.4. Chemical visualizer

The chemical visualizer (Fig. 2), a Jmol [17] derivative, facilitates the interactive display of molecular structures in 3D space together with their associated physicochemical properties. CLEVER features an automatic browsing functionality, which is useful for systematic viewing of huge compound collections. The animation settings that apply as each molecule is displayed can also be configured.

2.5. Chemical analyzer

The chemical analyzer enables quality control of compound collections through the computation and analysis of important physicochemical properties including the number of hydrogen bond acceptors and donors, XLogP (partition coefficient) values, molecular weights, number of rotatable bond and the Topological Polar Surface Area (TPSA) of compounds. Using these properties, compounds can be assessed and filtered as a new collection on the basis of drug-likeness [12,18], lead-likeness [19], fragment-like [20], Vernalis-leads [21] and also Vernalis-fragments [21].

The integrated graphing component creates two types of charts. Histograms that depict the distribution of chemicals within selected libraries can be generated based on a particular physicochemical property and configurable bin sizes. The diversity and coverage of chemical libraries can also be visualized via plots that scatter chemical compounds in 2D space based on the selected physicochemical properties for the *X* and *Y* axes. The software is available at http://datam.i2r.a-star.edu.sg/clever/.

3. Results and discussion

Among others, CLEVER can be deployed appropriately for generating novel chemical libraries, assessing the characteristics of chemical collections, as well as filtering chemical collections based on their physicochemical properties. Here we present three examples that illustrate the utility of CLEVER for each of these tasks.

3.1. Enumerating novel chemical libraries

To generate novel chemical compound libraries using CLEVER, we first begin with a scaffold molecule. For the purpose of this illustration, the Thyroxine (PubChem CID: 853) molecule was enumerated by replacing the iodine molecule with "blocks" from a

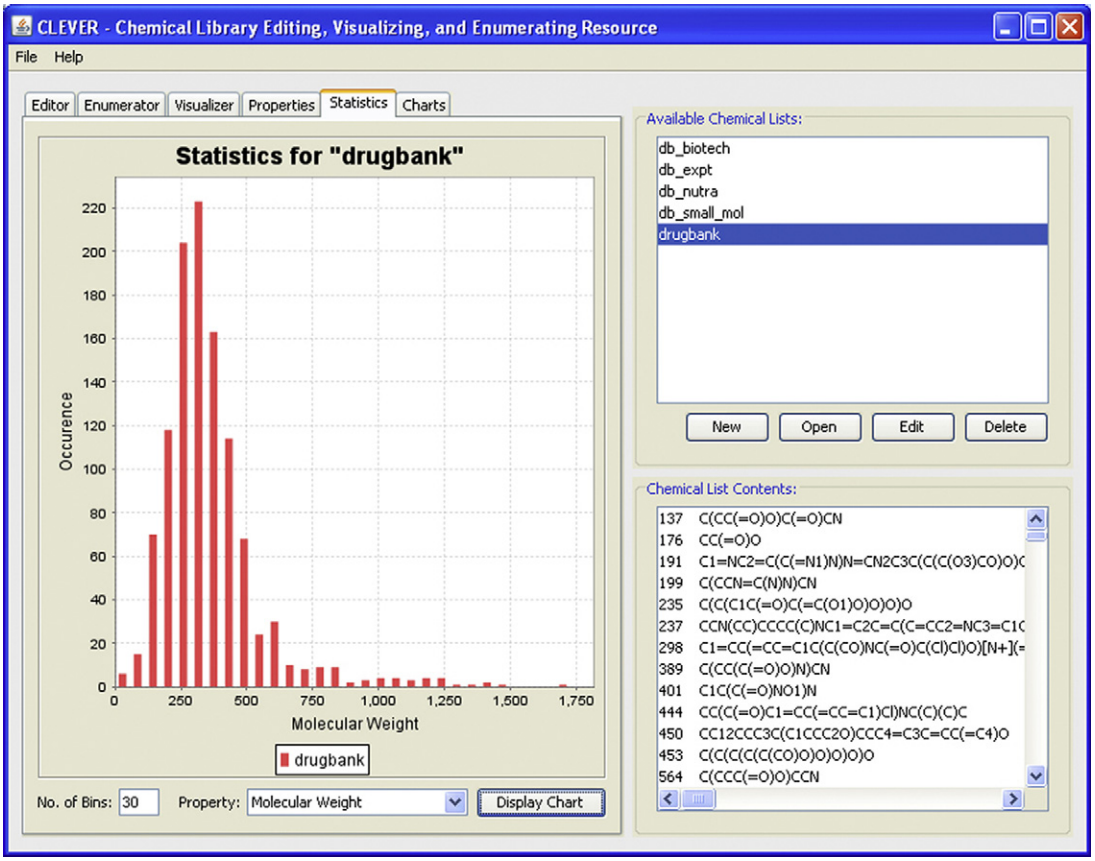


Fig. 3. Distribution of compounds of a selected collection(s). Based on the computed properties of individual chemical compounds, CLEVER dynamically plots histograms that provide insights into the distribution of the selected collection(s).

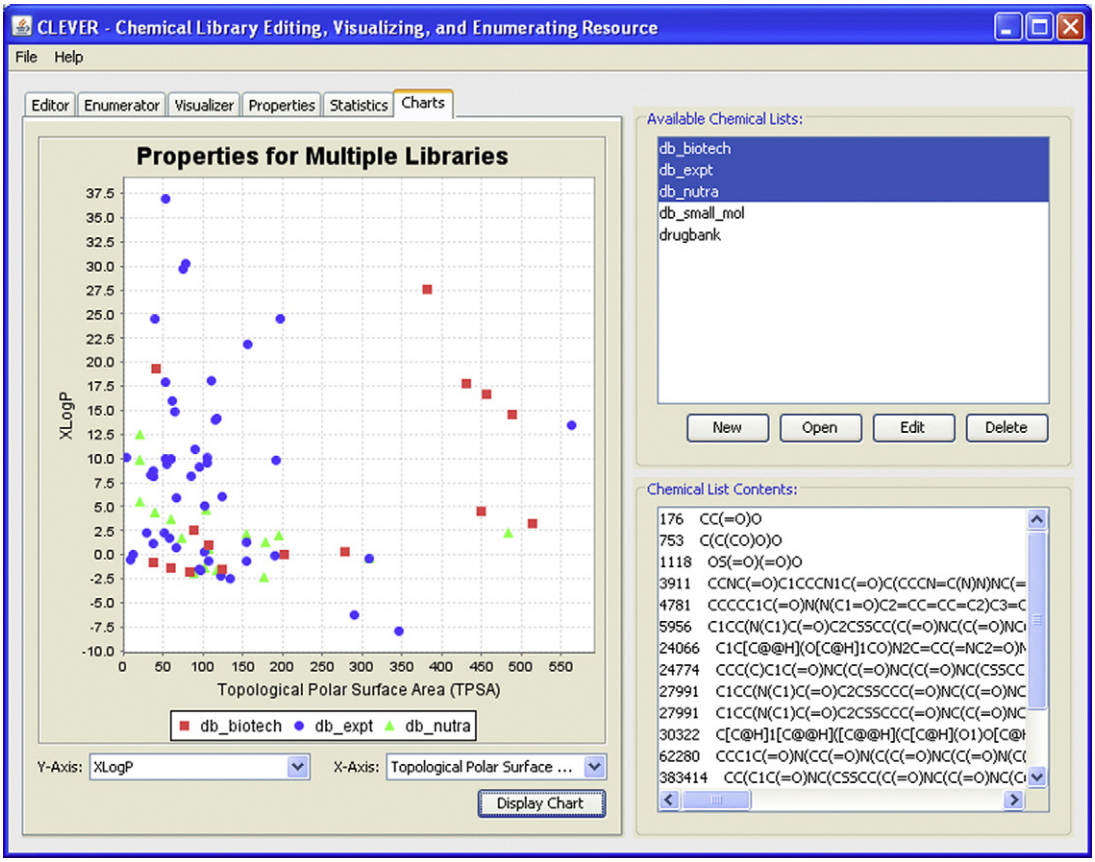


Fig. 4. Scatter plots of one or more libraries. CLEVER can also graphically depict the diversity and coverage of selected libraries by plotting individual chemical compounds on scatter plots with the selected physicochemical properties on the *X* and *Y* axes, respectively.

library of just seven small molecules. The SMILES string for each of the molecules used in this exercise is presented in Table 1.

Despite the modest number of building blocks used, the total number of chemical molecules derived by fully enumerating each of the three attachment points on the scaffold is at a staggering 343, and would continue to increase exponentially as larger libraries of building blocks are used. However, by specifying a reaction scheme that dictates the specific groups of building blocks to be used on each attachment point, the number of resulting molecules can be reduced substantially. Table 2 exemplifies possible reaction schemes and the respective sizes of the derived libraries. By enumerating using only the first two blocks (B1 and B2) for each of the three attachment points, we generated a library comprising of eight chemical compounds, of which two correspond to known derivatives of Thyroxine as shown in Table 3.

3.2. Analyzing properties of chemical libraries

The utility of CLEVER for analyzing properties of chemical libraries is exemplified here using data from DrugBank [22]. With over 4000 drugs in its collection organized under subcategories of "nutraceutical", "biotech", "experimental", "small molecules", etc., this resource is an excellent source for characterizing properties of drugs. Data from DrugBank were downloaded, parsed for the SMILES representation for each compound and organized into collections for analysis using CLEVER.

Using the "Statistics" module of CLEVER (Fig. 3), the distribution of drug molecules obtained from DrugBank can be displayed as histograms according to the selected physicochemical property and chosen bin sizes. An alternative view of the physicochemical properties of selected libraries is available via the "Charts" component, where in this case the TPSA and XLogP values of multiple drug collections were plotted as 2D scatter plots (Fig. 4).

3.3. Filtering chemical libraries

CLEVER incorporates several familiar schemes that describe certain classes of chemical molecules based on their physicochemical properties. These schemes include drug-likeness [12,18], lead-likeness [19] and fragment-likeness [20]. User-specified filtration criteria may also be specified for filtering chemical libraries. Using an initial collection of randomly selected molecules from DrugBank,

Table 4
Chemical filtration schemes and resulting collection sizes.

Filtering condition					Derived library size
Scheme	Hydrogen acceptors	Hydrogen donors	Molecular weight	XLogP	
None	–	–	–	–	1101
Drug-like [12,18]	0–5	0–10	0.0–500.0	−5.0 to 5.0	217
Lead-like [19]	0–6	0–3	150.0–1000.0	−2.0 to 4.0	140
Fragment-like [20]	0–4	0–2	150.0–250.0	−2.0 to 3.0	25

different filters were applied to derive new collections. The filtration criteria and resulting library sizes are outlined in Table 4.

4. Conclusions

The CLEVER program provides a suite of tools suitable for editing, enumerating, converting, visualizing and analyzing chemical libraries. The graphical user interface provides an uncomplicated means for rapidly generating novel chemical libraries and streamlining them to exclude unlikely candidates. Recently, it has been estimated that a minimum series of 200 compounds is necessary for high probability conformation of Structure–Activity Relationship (SAR) with at least two hits from the same series, while more substantial early SAR (at least five hits from the same series) may be obtained by using series of approximately 650 compounds each [22]. When used in conjunction with these criteria or other virtual screening initiatives, CLEVER can contribute in the design of sensible chemical collections that may be eventually evolved and refined into novel drugs. Future works will revolve around selection schemes that are capable of discriminating between drugs and non-drugs. While generic schemes such as those proposed by Lipinski have been generally accepted, it is also clear that these policies are inadequate for describing drug-like characteristics. As such, the next step for improving CLEVER is to develop and incorporate novel schemes that can predict drug-like properties of the generated candidates with reasonable accuracies.

References

[1] E.J. Martin, R.E. Critchlow, Beyond mere diversity: tailoring combinatorial libraries for drug discovery, J. Comb. Chem. 1 (1999) 32–45.
[2] M.J. Valler, D. Green, Diversity screening versus focussed screening in drug discovery, Drug Discov. Today 5 (2000) 286–293.
[3] E.A. Jamois, Reagent-based and product-based computational approaches in library design, Curr. Opin. Chem. Biol. 7 (2003) 326–330.
[4] A.R. Leach, M.M. Hann, The in silico world of virtual libraries, Drug Discov. Today 5 (2000) 326–336.
[5] J.F. Truchon, C.I. Bayly, GLARE: a new approach for filtering large reagent lists in combinatorial library design using product properties, J. Chem. Inf. Model. 46 (2006) 1536–1548.
[6] D.K. Agrafiotis, V.S. Lobanov, F.R. Salemme, Combinatorial informatics in the post-genomics ERA, Nat. Rev. Drug Discov. 1 (2002) 337–346.
[7] B.A. Leland, B.D. Christie, J.G. Nourse, D.L. Grier, R.E. Carhart, T. Maffett, S.M. Welford, D.H. Smith, Managing the combinatorial explosion, J. Chem. Inf. Comput. Sci. 37 (1997) 62–70.
[8] R.D. Brown, M. Hassan, M. Waldman, Combinatorial library design for diversity, cost efficiency, and drug-like character, J. Mol. Graph. Model. 18 (2000) 427–437, 537.
[9] A. Schüller, V. Hähnke, G. Schneider, SmiLib v2.0: a Java-based tool for rapid combinatorial library enumeration, QSAR Comb. Sci. 26 (2007) 407–410.
[10] R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, E.L. Willighagen, The Blue Obelisk—interoperability in chemical informatics, J. Chem. Inf. Model. 46 (2006) 991–998.
[11] J. Sadowski, Optimization of chemical libraries by neural networks, Curr. Opin. Chem. Biol. 4 (2000) 280–282.
[12] C.A. Lipinski, Drug-like properties and the causes of poor solubility and poor permeability, J. Pharmacol. Toxicol. Methods 44 (2000) 235–249.
[13] V. Lobanov, Using artificial neural networks to drive virtual screening of combinatorial libraries, Drug Discov. Today: BIOSILICO 2 (2004) 149–156.
[14] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[15] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, E.L. Willighagen, Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bioinformatics, Curr. Pharm. Des. 12 (2006) 2111–2120.
[16] J. Sadowski, A hybrid approach for addressing ring flexibility in 3D database searching, J. Comput. Aided Mol. Des. 11 (1997) 53–60.
[17] H. Angel, Biomolecules in the computer: Jmol to the rescue, Biochem. Educ. 34 (2006) 255–261.
[18] A.K. Ghose, V.N. Viswanadhan, 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) 55–68.
[19] S.J. Teague, A.M. Davis, P.D. Leeson, T. Oprea, The design of leadlike combinatorial libraries, Angew. Chem. Int. Ed. Engl. 38 (1999) 3743–3748.
[20] R.A. Carr, M. Congreve, C.W. Murray, D.C. Rees, Fragment-based lead discovery: leads by design, Drug Discov. Today 10 (2005) 987–992.
[21] N. Baurin, R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parratt, P. Greaney, et al., Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds, J. Chem. Inf. Comput. Sci. 44 (2004) 643–651.
[22] D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res. 36 (2008) D901–D906;
M.J. Lipkin, A.P. Stevens, D.J. Livingstone, C.J. Harris, How large does a compound screening collection need to be? Comb. Chem. High Throughput Screen. 11 (2008) 482–493.