Vol.2015-BIO-41 No.12
2015/3/20

# 方向性を持たせたグラフ構造変換による仮想化合物ライブラリ構築の研究

吉川 舜亮[†1,a)]　安尾 信明[†1]　吉野 龍ノ介[†2]　関嶋 政和[†1,†2,b)]

**概要**：医薬品研究開発に必要な膨大な期間とコストの効率化を目的として，薬剤の候補となる化合物を探索するという見地から，仮想化合物ライブラリの活用が期待されている．本研究では，入力した化合物データに対して仮想的な化学反応を適用することで新規化合物を生成してライブラリを構築するシステムを開発した．その際に，化合物構造をグラフとして扱い，化合物グラフに対して部分グラフの変換を行うことで仮想的な化学反応を表現している．また，ここで適用する仮想的な化学反応に方向性を持たせることで，目的に応じた特性を持つ化合物を含むライブラリの構築を成功した．

## 1. はじめに

### 1.1 研究背景

医薬品研究開発のプロセスは図 1 に示すように，標的分子の同定から臨床試験まで，多くの過程における絞り込みを経て薬剤は完成する [1]．新薬を完成させるためには十数年もの期間と数百億円もの費用が必要であると言われており [2]，この膨大な期間と費用の削減が医薬品研究開発における課題となっている．

創薬研究において，研究対象となる化合物の理論的な総数は 10 の 60 乗にも昇るとされている [3] のに対して，製薬会社が保有するライブラリに含まれる化合物数は数百万程度であり [4]，探索範囲が不十分であることが問題となっている．このような問題の解決のために，データとして仮想的に化合物を生成する研究が有用である．

今日では仮想的に化合物を生成するシステムを開発する様々な研究が行われている [5][6]．しかし，既存のシステムでは仮想的に生成する化合物の特性や個数を柔軟に変更できないため，開発する薬剤によって異なる研究者の要求に合わない可能性がある．そのため，目的に応じて柔軟にライブラリを構築できるシステムが重要である．

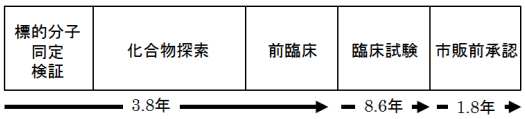


**図 1**　薬剤開発に必要な手順．[1] より改変

### 1.2 研究目的

本研究では，目的に応じた特性を持つ化合物を生成してライブラリを構築するシステムの開発を行った．本システムでは化合物のデータを入力し，入力した化合物データに対して仮想的に化学反応を行うことで新しく化合物のデータを生成し，仮想化合物ライブラリを構築する．その際に，化合物の構造をグラフ構造として扱い，部分グラフの変換によって化学反応による化合物の構造の変換を表現する．この仮想的な化学反応に方向性を持たせることで，目的に応じた特性を持つ化合物を含むライブラリを構築する．また，本システムでは入力した化合物から反応ルールを適用することで新しく生成された化合物グラフを再度入力として適用し，多段階に構造の変換を行う．構造の変換の段階数を経るごとに入力化合物から離れた構造を持つ化合物を生成でき，より多様な化合物が含まれるライブラリの構築が可能である．このシステムで行う部分グラフの変換は実際に合成の現場で使用される化学反応のみを表現しているため，ライブラリに含まれるすべての化合物は合成が可能であり，多段階に部分グラフの変換を行った場合でもその合成経路を辿ることが可能となっている．

[†1] 現在，東京工業大学大学院情報理工学研究科計算工学専攻
Presently with Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[†2] 現在，東京工業大学学術国際情報センター
Presently with Global Scientific Information and Computing Center, Tokyo Institute of Technology
a) yoshikawa.s.ac@m.titech.ac.jp
b) sekijima@gsic.titech.ac.jp

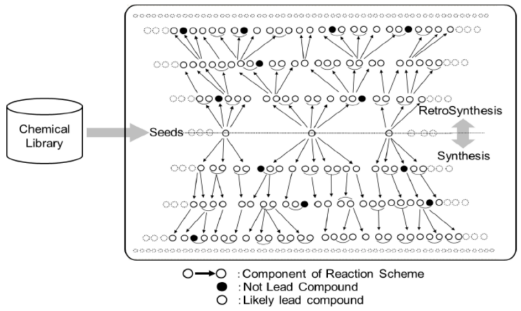


**図 2** 大規模仮想ライブラリ構築の様子．[5] より

### 1.3 本論文の構成

本論文では，まず第二章で仮想的に化合物を生成する既存研究の紹介を行う．第三章では本研究で開発したライブラリ構築システムを説明する．第四章では本研究で開発したシステムを使用してライブラリを構築する実験について記述する．第五章では実験結果を記述し，それ元にした考察を行う．第六章で本論文の結論と今後の課題をまとめる．

## 2. 既存研究

### 2.1 反応スキームによる合成可能な大規模仮想ライブラリの研究

既存研究 [5] では，薬剤の候補となる化合物を発見するために，仮想的に生成する化合物数の増加によって網羅的な探索を行うことを目的としている．この研究のシステムでは，化合物のデータを入力し，その化合物の構造を変換させることにより新しい構造を持つ化合物を生成してライブラリを構築する．その際に，42 万化合物を含む公開化学構造ライブラリ ZINC[7] を入力として用い，化学反応のデータベースから自動的に抽出した様々な化合物の構造変換を適用することで大規模な仮想ライブラリの構築を行う．また，入力化合物の構造から一度変換させた化合物を再度入力化合物として利用することで，更に化合物数を増加させ，1 億以上ものユニークな化合物を含むライブラリを構築している．図 2 にこの大規模仮想ライブラリ構築の様子を示す．

しかし，化合物数の増加によって網羅的な探索を行う試みは非効率的である．この研究では 1 億を超える化合物データを生成しているが，研究対象となる化合物の理論的な総数である 10 の 60 乗には届いておらず，網羅的な探索には不十分な化合物数である．また，1 億を超える化合物データから薬剤の候補となる化合物を選び出す作業には膨大なコストと期間が必要となる．以上の問題から，化合物数を増加させることによって網羅的な探索を行う試みは薬剤開発の効率化には不適切である．

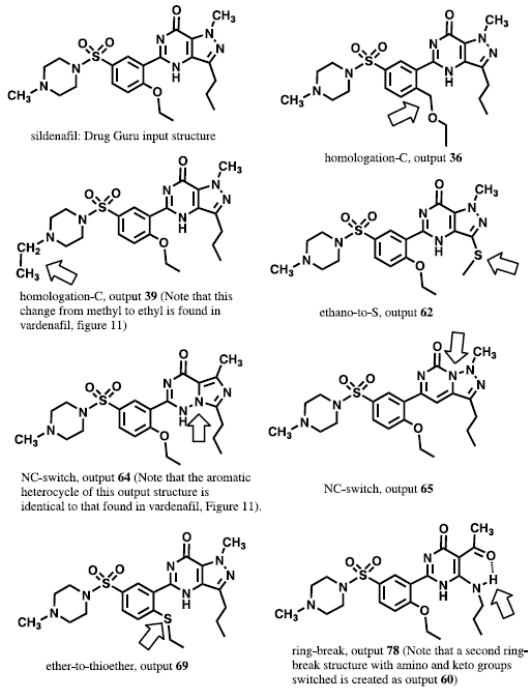


**図 3** Drug Guru による構造最適化の例．左上の入力化合物から得られる出力構造の例．[6] より

### 2.2 Drug guru

入力された化合物構造に対して部分構造の変換を自動的に行う Drug guru[6] は，薬剤の候補として選定された化合物の構造を薬剤として適切な構造に変換する化合物構造の最適化などを目的として用いられる．Drug Guru を使用した構造最適化の例を図 3 に示す．図 3 は，入力化合物の構造と入力化合物の構造から得られる変換後の化合物構造の一部を示している．

Drug guru は化合物構造の変換は一段階しか行われないという問題がある．入力化合物に対して一段階の構造変換では入力化合物と比べて近い構造の化合物しか取得することができず，多様な化合物構造の取得が必要なケースでは不適切なシステムである．

## 3. ライブラリ構築方法

### 3.1 化合物の表記方法

本システムではグラフ理論に基づいて化合物構造を文字列で表す SMILES[8] を用いて化合物を扱う．SMILES は，化合物の原子を頂点，結合を辺として割り当てることで化合物構造をグラフとして表現し，グラフの各頂点と各辺を文字や記号に割り当てることで化合物を文字列として表す記法である．また，本システムでは SMILES の検索用拡張である SMARTS[9] を用いて化合物に対する化学反応の適用を化合物グラフに対する部分グラフの検索と変換によってルール化して取り扱う．本論文では，この仮想的な化学

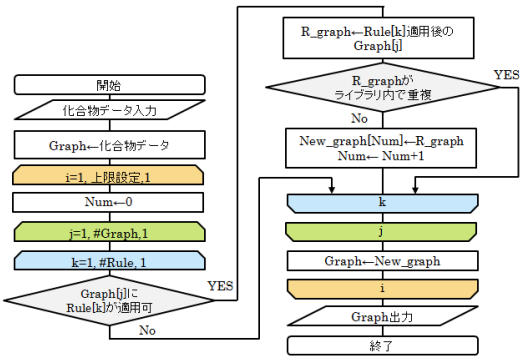


**図 4** ライブラリ構築のフローチャート

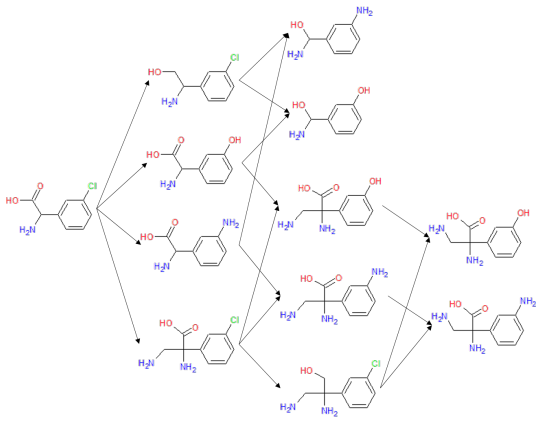

**図 5** 多段階の化合物グラフ変換の様子

反応を反応ルールと呼ぶ.

### 3.2 ライブラリ構築の概要

ライブラリ構築のフローチャートを図 4 に示す．まず，変換させたい化合物のデータセットをシステムに入力する．入力された化合物データはシステム内でグラフ構造の集合として扱い，ライブラリの構築を開始する．このフローチャートでは i,j,k という 3 種類の文字を使用して 3 種類の繰り返しを表現しており，i,j,k はそれぞれ段階数，各化合物グラフ，各反応ルールを表している．まず，一番目の化合物グラフに対して各反応ルールが適用可能かを判断し，適用可能な場合はルールを適用して部分グラフの変換を行い，新規化合物グラフの生成を行う．ここで新しく生成された化合物グラフはライブラリ内のすべての化合物グラフに対して重複確認を行い，重複していなければ新しい化合物グラフの集合に加える．各化合物グラフに対して同様にすべての反応ルールを適用することで，一段階の新規化合物グラフの生成が完了する．ここで生成された新規化合物グラフを再度システムの入力化合物群として適用することで二段階目の新規化合物グラフの生成を行う．このように上記の作業を繰り返すことで，多段階の反応ルールの適用による化合物グラフの生成を行う．段階数の上限を構築開始時に設定することで，構築するライブラリの規模を目的に応じて調節することができる．段階ごとに化合物数が増加する様子を図 5 に示す.

### 3.3 構造変換の方向性

単純な化合物数の増加による化合物空間の網羅的な探索目指すことは非効率的である．そこで本研究では，入力化合物に適用する反応ルールに方向性を持たせることで，目的に応じた特性を持つ化合物を含むライブラリの構築を目指す.

本研究では，以下の 3 種類の反応ルールの方向性を作成した.

(1) 水素結合の供与体（水素を供与する側, hydrogen bond donor）の個数を増加させる方向

**表 1** 反応ルールの個数

| 反応ルールの方向性 | 反応ルールの個数 |
|---|---|
| (1) 供与体数増加方向の反応ルール | 56 |
| (2) 受容体数増加方向の反応ルール | 66 |
| (3) 環構造数増加方向の反応ルール | 22 |

(2) 水素結合の受容体（水素を受け取る側, hydrogen bond acceptor）の個数を増加させる方向

(3) 環構造の数を増加させる方向

これらの方向性は複数を選択して使用することも可能である.

水素結合の供与体や受容体を持つ化合物は水との間に水素結合を作ることで水に融解する性質を持つ．このことから水素結合の供与体と受容体は親水性を計る指標となっており，薬剤開発において親水性は頻繁に用いられる指標である [10]．また，環構造を多数含む化合物は安定性が高く，安定性の高い化合物は薬剤として望ましいと考えられる [11].

上記の 3 種類の方向性にそれぞれ適用する反応ルールの個数を表 1 に示す．本システムには全部で 308 種類の反応ルールが導入されており，方向性を選択することで各方向において考慮する特性に該当する部分グラフが増加する反応ルールを選択してライブラリの構築を行った.

## 4. 実験

### 4.1 実験概要

システムの評価を行うために，ナミキ商事の building block 統合データベース（2013 年 6 月版）から 500 種類の化合物を抽出してシステムの入力化合物として用いた．反応ルールの適用の段階数の上限は 3 とし，以下の 4 種類のライブラリの構築を行った.

(1) all_library: すべての反応ルールを適用して構築したライブラリ

(2) donor_library: 供与体数増加方向の反応ルールを適用して構築したライブラリ

**表 2** 反応ルール数，ライブラリの化合物数の比較

| ライブラリ名 | 反応ルール数 | 化合物数 |
|---|---|---|
| all_library | 308 | 2946223 |
| donor_library | 56 | 42257 |
| acceptor_library | 66 | 135751 |
| ring_library | 22 | 26977 |

**表 3** 実験環境

| | |
|---|---|
| CPU | Intel core i7-3770(3.4GHz) |
| メモリ | 16GB |
| 言語 | Python 2.7.4 |
| OS | Ubuntu 13.04 |

(3) acceptor_library: 受容体数増加方向の反応ルールを適用して構築したライブラリ
(4) ring_library: 環構造数増加方向の反応ルールを適用して構築したライブラリ

まず，本システムに導入されている 308 種類すべての反応ルールを適用した all_library と呼ぶライブラリの構築を行った．また，供与体数増加方向の反応ルール，受容体数増加方向の反応ルール，環構造数増加方向の反応ルールを適用した donor_library, acceptor_library, ring_library と呼ぶライブラリの構築を行った．4 種類のライブラリに含まれる化合物数，適用した反応ルール数を表 2 に示す．構築された 4 種類のライブラリに含まれる化合物をグラフとして扱い，供与体，受容体，環構造に該当する部分グラフの個数を比較することで，本システムで構築されたライブラリの評価を行う．供与体数，受容体数，環構造数はいずれも RDKit[12] によって計算を行った．

### 4.2 実験環境

本実験に使用した CPU，メモリ，プログラミング言語，OS を表 3 に示す．

## 5. 結果・考察

### 5.1 供与体数増加方向

入力化合物と供与体数増加方向反応ルールを適用して構築したライブラリである donor_library に含まれる化合物の供与体数の平均と標準偏差を表 4 に示す．表 4 から，donor_library は入力化合物と比べて供与体数の平均が 1.72 程度大きいことがわかる．また，入力化合物と donor_library に含まれる化合物の供与体数の割合の分布を図 6 に示す．図 6 から，入力化合物は供与体数が 1 の化合物の割合が最大である一方で，donor_library は供与体数が 3 の化合物の割合が最大であり，供与体数を多く含む化合物の割合が大きいライブラリを構築できたことがわかる．

**表 4** 入力化合物と donor_library 内の化合物の供与体数の平均と標準偏差

| ライブラリ名 | 平均 | 標準偏差 |
|---|---|---|
| 入力化合物 | 0.96 | 0.75 |
| donor_library | 2.68 | 0.70 |

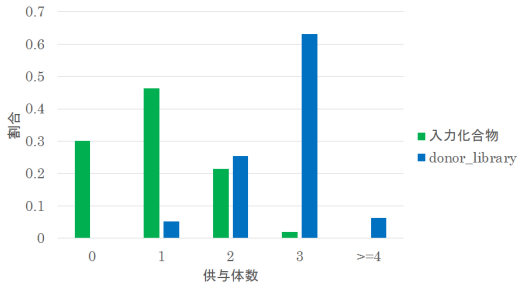


**図 6** 入力化合物と donor_library 内の化合物の供与体数の割合の分布

**表 5** 入力化合物と acceptor_library 内の化合物の受容体数の平均と標準偏差

| ライブラリ名 | 平均 | 標準偏差 |
|---|---|---|
| 入力化合物 | 3.37 | 1.39 |
| acceptor_library | 4.25 | 1.32 |

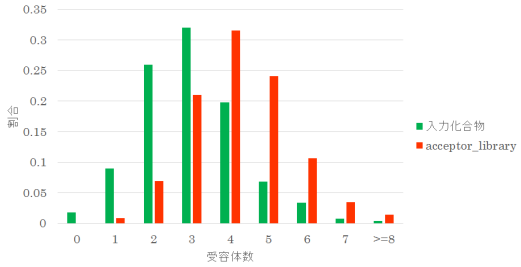


**図 7** 入力化合物と acceptor_library 内の化合物の受容体数の割合の分布

### 5.2 受容体数増加方向

入力化合物と受容体数増加方向反応ルールを適用して構築したライブラリである acceptor_library に含まれる化合物の受容体数の平均と標準偏差を表 5 に示す．表 5 から，acceptor_library は入力化合物と比べて受容体数の平均が 0.88 程度大きいことがわかる．また，入力化合物と acceptor_library に含まれる化合物の受容体数の割合の分布を図 7 に示す．図 7 から，入力化合物は受容体数が 3 の化合物の割合が最大である一方で，acceptor_library は受容体数が 4 の化合物の割合が最大であり，受容体数を多く含む化合物の割合が大きいライブラリを構築できたことがわかる．

### 5.3 環構造数増加方向

入力化合物と環構造数増加方向反応ルールを適用して構築したライブラリである ring_library に含まれる化合物

**表 6** 入力化合物と ring_library 内の化合物の環構造数の平均と標準偏差

| ライブラリ名 | 平均 | 標準偏差 |
|---|---|---|
| 入力化合物 | 1.78 | 0.75 |
| ring_library | 3.36 | 0.87 |

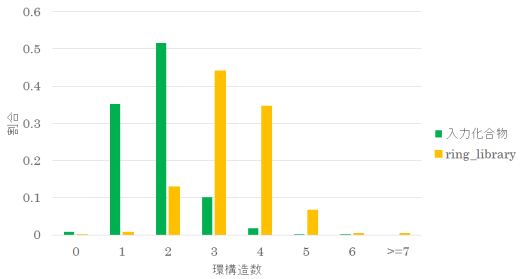


**図 8** 入力化合物と ring_library 内の化合物の環構造数の割合の分布

の環構造数の平均と標準偏差を表 6 に示す．表 6 から，ring_library は入力化合物と比べて環構造数の平均が 1.58 程度大きいことがわかる．また，入力化合物と ring_library に含まれる化合物の環構造数の割合の分布を図 8 に示す．図 8 から，入力化合物は環構造数が 2 の化合物の割合が最大である一方で，ring_library は環構造数が 3 の化合物の割合が最大であり，環構造数を多く含む化合物の割合が大きいライブラリを構築できたことがわかる．

### 5.4 構築時間

4 種類のライブラリの構築時間を表 7 に示す．500 種類の入力化合物に対してすべての反応を考慮するライブラリである all_library を構築したところ，約 294 万種類の化合物を含むライブラリを構築し，約 27.5 時間の実行時間が必要となっていた．一方で，他の 3 種類のライブラリの構築に要する時間は数十秒から数分程度であり，数万から十数万種類の化合物を含むライブラリを構築した．このことから，反応の方向性を選択することで，all_library と比べて短時間でライブラリを構築できたことがわかる．

すべての反応ルールを適用した all_library の構築時間において，全体の約 88.7%が重複確認に必要な時間となっていたことがわかった．m 段階目に生成された化合物数を $N_m$ として，m 段階目までに生成された化合物数の合計 $M$ は

$$M = \sum_{k=1}^{m} N_k$$

となる．ここで m+1 段階目の反応ルールの適用における重複確認の回数は

$$\begin{aligned} &M + (M+1) + ... + (M + N_{m+1}) \\ &\quad = N_{m+1}^2 + 2N_{m+1}M - M^2 \end{aligned}$$

となる．多くの反応ルールが適用され，$N_{m+1}$ が $M$ と比べて充分に大きい場合，重複確認に必要な計算量は $O(N_{m+1}^2)$ となることがわかる．

**表 7** 4 種類のライブラリの構築時間

| ライブラリ名 | 構築時間 | 化合物数 |
|---|---|---|
| all_library | $9.9 \times 10^4$ 秒（約 27.5 時間） | 2946223 |
| donor_library | 70 秒 | 42257 |
| acceptor_library | 394 秒 | 1335751 |
| ring_library | 32 秒 | 26977 |

## 6. 終わりに

### 6.1 本論文の結論

本研究では，化合物構造をグラフ構造として表現し，方向性を持たせた反応ルールの適用を繰り返し行うことで新規化合物を生成し，ライブラリを構築するシステムを開発した．本システムでは，水素結合の供与体増加方向，水素結合の受容体増加方向，環構造増加方向の 3 種類の反応ルールの方向性を作成し，

ナミキ商事の building block 統合データベース（2013 年 6 月版）から抽出した 500 種類の入力化合物に対してライブラリの構築を行ったところ，目的に応じた特性を持つ化合物の割合が大きいライブラリを構築することに成功した．また，方向性を指定することで，すべての反応ルールを適用するよりも短時間でのライブラリ構築が可能となった．以上から，方向性を指定することで短時間で目的に応じた化合物を含むライブラリを構築できるシステム開発した．

### 6.2 今後の課題

今後の課題として，実際の創薬の現場で利用されることを目標に，更なる反応ルールの充実と方向性の追加が挙げられる．今回の実験では分子量が小さい化合物を入力し，化合物を大きくすることを前提にライブラリの構築を行った．しかし，極端に分子量の大きい化合物は薬剤として適切ではないため，分子量を小さくする方向の反応ルールを適用する方向性を作成することで実用性が高まると考えられる．

また，本システムは反応ルールが適用可能な場合はすべての部分構造に対して適用している．しかし，薬剤としての効果を持つ部分構造に対して反応ルールを適用して生成される化合物はライブラリに含まれる化合物として適切ではないため，指定した部分構造に対しては反応ルールを適用しないというオプションの作成が今後の課題として考えられる．

本システムは方向性を選択することにより，ライブラリ構築時間を短縮したが，入力化合物数や反応ルール数が増加し，生成する化合物数が多くなると構築時間が長くなると考えられるため，ライブラリ構築の並列化や重複確認の効率化による高速化が課題として考えられる．

## 参考文献

[1] J.K.Willmann, N.V.Bruggen, L.M.Dinkelborg *et al.*, Molecular imaging in Drug Development,*Nat.Rev.Drug.Discov.*, Vol.7, pp.591, (2008)
[2] T.YAGI and M.Okubo, JPMA News Letter, No.136, pp.33, (2010)
[3] R.S.Bohacek, C.McMartin and W.C.Guida, The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective, *Med.Res.Rev.*, Vol.16, pp.3, (1996)
[4] 日暮那造, 井上篤, 「化合物ライブラリのトレンドとエーザイにおける取り組み」 *CICSJ Bulletin*, Vol.23, No.5, pp.156, (2005)
[5] 西村拓朗, 船津公人, 「大規模バーチャルライブラリ開発の試み」 *CICSJ Bulletin*, Vol.29, No.3, pp.49, (2011)
[6] K.D.Stewart, M.Shiroda, C.A.James; Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorg.Med.Chem.*, Vol.14, No.20, pp.7011, (2006)
[7] J.J.Irwin, B.K.Shoichet, ZINC-A Free Database of Commercially Available Compounds for Virtual Screening. *J.Chem.Inf.Model.*, Vol.45, No.1, pp.177, (2005)
http://zinc.docking.org/index.shtml
[8] D.Weininger, SMILES, a Chemical Language and Information System. 1. Introduction and Encoding Rules. *J.Chem.Inf.Comput.Sci.*, Vol.28, No.1, pp.31, (1988)
http://www.daylight.com/
[9] http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
[10] C.A.Lipinski, F.Lombardo, B.W.Dominy, *et al.*, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv.Drug.Deliv.Rev.*, Vol.23, No.1, pp.3, (1997)
[11] T.J.Ritchie, S.J.F.Macdonald, The The impact of aromatic ring count on compound developability - are too many aromatic rings a liability in drug design? *Drug.Discov.Today.*, Vol.14, No.21, pp.1011, (2009)
[12] http://www.rdkit.org/