SMILES 記法を利用した薬物設計ソフトウェアの開発

藤　秀義，辰巳 絢子，岩本 光司，星野 忠次
千葉大学大学院医学薬学府

Development of Drug Designing Software using SMILES

Hideyoshi Fuji, Junko Tatsumi, Kouji Iwamoto, Tyuji Hoshino
Graduate School of Pharmaceutical Sciences, Chiba University

Abstract

SMILES (Simplified Molecular Input Line Entry System) is a chemical notation system which is based on principles of molecular graph theory. Because chemical structures can be written in a single string using SMILES, it is widely used for chemical databases. In general, 3D chemical structures are required for computer assisted drug design. In this study, we have developed a computer software which can construct a 3D structure of compounds from SMILES and dock it into a target protein.

概要

SMILES 記法とは、グラフ理論に基づいて考案された分子の化学構造を表記する方法である。これを用いることで、化学構造を１行の文字列で表記することができるため、広く化合物データベースに用いられている。通常、コンピュータ上で薬物を設計する場合、化合物の３次元立体構造が必要である。そこで本研究では、SMILES 文字列から化合物の３次元立体構造を簡便に構築し、さらに化合物を標的タンパク質へとドッキングさせるプログラムの開発を行った。

1. 序論

2003 年の 4 月 14 日にヒトゲノム解読完了の宣言がなされ、現在ポストゲノムと呼ばれる時代を迎えている。医薬品開発においては、ゲノム情報から病気の原因を特定し、それを治療する薬物を理論的にデザインする「ゲノム創薬」というプロセスを組むようになった。従来の創薬方法では、数万種類の化合物の活性を調べる必要が有り、コストと時間が膨大に必要であった。また、研究者の経験や勘に頼るところも多く、偶然に活性薬物を発見することも少なくなかった。一方ゲノム創薬では、理論的に標的 DNA や標

的タンパク質を特定することにより、効果が高く副作用の少ない薬物を低コストで効率的に開発することができる。

ゲノム創薬におけるキーポイントは、Structure-Based Drug Design（SBDD）である。SBDD とは、薬物標的である生体高分子の３次元立体構造をもとにして、その形に合った薬物分子をコンピュータ上で設計する方法である。SBDD に必要なものは、生体高分子と化合物の３次元立体構造情報であるが、構造既知のタンパク質や核酸については Protein Data Bank（http://www.rcsb.org）[1]より３次元立体構造情報を得ることができる。一方、化合物の３次元立体構造については、Protein Data Bank に３次元立体構造が登録されているものもあるが、その数には限りがある。その代わり、化合物の２次元平面構造を集めたデータベースは多く存在している。これら化合物の２次元構造を３次元構造に起こすことが SBDD には必要不可欠である。

そこで我々は、化合物の２次元構造情報から３次元構造に変換し、化合物の構造最適化を行うことのできるソフトウェアの開発を思い立つに至った。本研究においては、SMILES 記法から化合物の３次元構造を構築するプログラムの開発を行った。SMILES 記法とは、グラフ理論に基づいて考案された分子の化学構造を表記する方法であり、化学構造を１行の文字列で表記することが可能である[2]。SMILES の表記例を図 1 に示す。SMILES は文字列であることから、広く化合物データベースに用いられている。例えば、National Cancer Institute の化合物データベース（http://cactus.nci.nih.gov）においては、25 万近くの化合物データが SMILES 文字列として登録されている。よって、本研究にて開発したプログラムを用いることによって、化合物データベースに登録された多くの化合物の立体構造構築が容易になり、創薬研究における SBDD の発展に大きく貢献できると期待される。

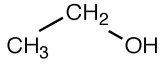
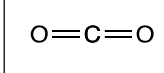
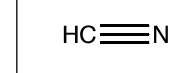
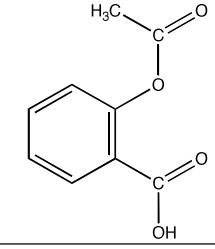
構造式				
分子名	エタノール	二酸化炭素	シアン化水素	アセチルサリチル酸
SMILES	CCO	O=C=O	C#N	CC(=O)Oc1ccccc1C(=O)O

図 1. SMILES による化合物表記例

2. SMILES

2.1. 原子

原子は元素記号にて表記する。塩素のような２文字で表す場合は、'Cl'というように２文字目を小文字で表記する。有機化合物中の元素、すなわち B, C, N, O, P, S, F, Cl, Br, I は、それぞれの原子価に合った数の水素が結合した分子を表記することができる。芳香環中の原子は、小文字で表記することができる。例えば、脂肪族の炭素は'C'、芳香族の炭素は'c'と表記する。

C	methane (CH_4)
N	ammonia (NH_3)
O	water (H_2O)
P	phosphine (PH_3)
Cl	hydrogen chloride (HCl)

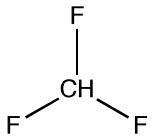
2.2. 結合

単結合、二重結合、三重結合、芳香性の結合は、それぞれ'-', '=', '#', ':'で表記する。通常、単結合'-'と芳香性の結合':'は省略される。

CC	ethane (CH_3CH_3)
C=C	ethane ($CH_2=CH_2$)
CCO	ethanol (CH_3CH_2OH)
C=O	formaldehyde (CH_2O)
O=C=O	carbon dioxide (CO_2)
C#N	hydrogen cyanide (HCN)

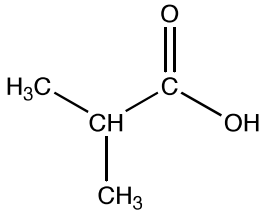
2.3. 分岐

分岐鎖は、()で囲んで表記する。



C(F)(F)F

Trifluoromethane

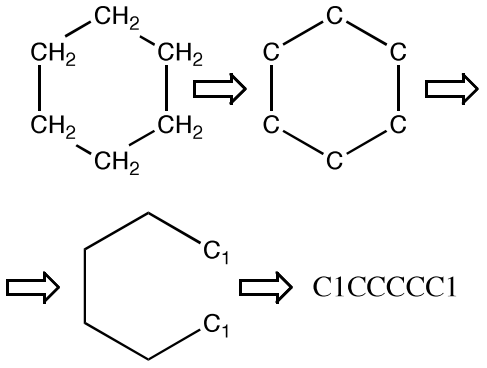


CC(C)C(=O)O

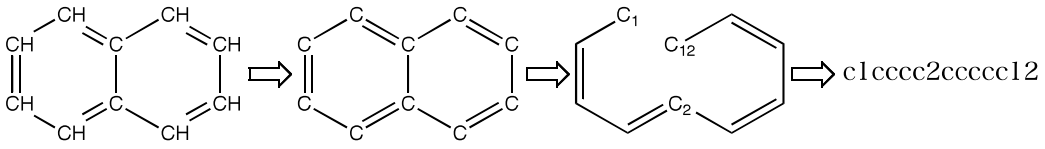
Isobutyric acid

2.4. 環

環状化合物は、まず１つの単結合もしくは芳香性の結合を切断し、鎖状構造とする。切断した結合の両端の原子に結合番号を付け、2.1 から 2.3 までのルールに従い SMILES を表記する。Cyclohexane と Naphthalene における例を下記に示す。



Cyclohexane



Naphthalene

3. 設計

3.1. 原子種と結合情報の抽出

SMILES 文字列から化合物を構成する原子の情報と、それらの原子の結合情報を抽出する。この作業は Open Babel（http://openbabel.sourceforge.net/）を用いて行うことにする。

3.2. 化合物の３次元立体構造構築

SMILES 文字列から得られた原子種と結合情報から分子の３次元立体構造を構築するために、距離幾何学法[3]を用いた。距離幾何学法は、分子の構造を導くのに４段階の手順を踏む。N 個の原子からなる分子を扱う場合、1）まず始めに原子間距離の上界下界行列（N×N 行列）を計算する。これは、分子内の各原子対に許される距離の最大値と最小値を要素とする行列である。2）次に、各原子間距離に対して、上界と下界の間の適当な値がランダムに与えられる。3）得られた距離行列から直交座標への変換を行い、4）直交座標のさらなる精密化を行う。

3)における詳しい計算式を以下に述べる。まず最初に、計量行列 G を計算する。G の行列要素(i, j)は、原点から原子 i と j へ向かうベクトルの内積に等しい。

$$G_{ij} = \mathrm{i} \bullet \mathrm{j} \qquad \text{（式 1）}$$

要素G_{ij}は、余弦定理を用いることにより距離行列から計算することができる。

$$G_{ij} = (d_{i0}^{\ 2} + d_{j0}^{\ 2} - d_{ij}^{\ 2})/2 \quad \text{（式 2）}$$

ここで、d_{i0}は原点から原子 i までの距離、d_{ij}は原子 i と j の間の距離を表す。

座標系の原点となるのは、一般に分子の中心である。中心からの各原子の距離は、次式により原子間距離から求めることができる。

$$d_{i0}^{\ 2} = \frac{1}{N}\sum_{j=1}^{N} d_{ij}^{\ 2} - \frac{1}{N^2}\sum_{j=2}^{N}\sum_{k=1}^{j-1} d_{jk}^{\ 2} \quad \text{（式 3）}$$

計量行列 G は正方対称行列である。このような行列は、一般に次のように分解できる。

$$\mathrm{G} = \mathrm{V}\mathrm{L}^2\mathrm{V}^{\mathrm{T}} \qquad \text{（式 4）}$$

L^2 の対角要素は G の固有値、V の列はその固有ベクトルになっている。計量行列から原子座標を求めるため、式１を次のように書き直す。

$$\mathrm{G} = \mathrm{X}\mathrm{X}^{\mathrm{T}} \qquad \text{（式 5）}$$

ここで、X_{ij}は原子 i の j 番目の座標軸での

位置を示す。式4と式5が等しいと置けば、次式が得られる。

$$X = VL \qquad \text{（式 6）}$$

行列 L^2 は対角項しかもたないので、行列 L も対角行列として表すことができる。すなわち、行列 L の転置行列 L^T は、元の行列 L と等しい（$L = L^T$）。式6より、原子座標を得るには、固有値の平方根に固有ベクトルを掛ければよいことがわかる。

4)における座標の精密化には、分子力学計算を行い、分子構造のエネルギー極小化を行う。

3.3. 標的タンパク質の薬物結合部位探索

タンパク質–薬物間の相互作用には疎水相互作用が重要な役割を果たしていると考えられている。そこで、タンパク質周囲の疎水性ポテンシャルを計算することにより薬物結合部位の探索を行う。方法としては、タンパク質周囲に一定間隔で格子点を発生させ、各格子点における疎水性ポテンシャルの計算を行う[4]。疎水性ポテンシャルの計算式を式7に示す。

$$\Delta G_H = -2.0R\exp(-D/10) \qquad \text{（式 7）}$$

ここで、ΔG_H は疎水性エネルギー（kcal/mol）。$R = R_1R_2/(R_1 + R_2)$、R_1はタンパク質中の炭素原子の半径、R_2は格子点上のプローブ炭素原子の半径。$D = R_{12} - (R_1 + R_2)$、R_{12}はタンパク質中の炭素原子と格子点間の距離。計算には、タンパク質中の疎水性残基（Gly, Ala, Val, Leu, Ile, Met, Trp, Phe, Pro）のカルボニル炭素を除く炭素原子を用いた。各格子点における疎水性エネルギーを計算後、エネルギーの低い方から格子点100個を取り出し、格子点の多く集まった領域を薬物結合部位とする。

3.4. 標的タンパク質への薬物ドッキング

探索した薬物結合部位の慣性主軸と化合物の構造の慣性主軸を重ね合わせることによって、薬物ドッキングを行う。

4. 実験

4.1. 化合物の３次元立体構造構築の検証

化合物の３次元構造構築のテストに、hexane（SMILES：CCCCCC）と cyclohexane（SMILES：C1CCCCC1）の計算を行った。計算結果を図 2 に示す。比較対象として、量子化学計算ソフト Gaussian03 を用いて最適化した構造を図 3 に示す。Hexane と cyclohexane の RMSD 値を計算したところ、Gaussian03 による最適化構造とのずれは、それぞれ 1.781Å、0.013Å であった。

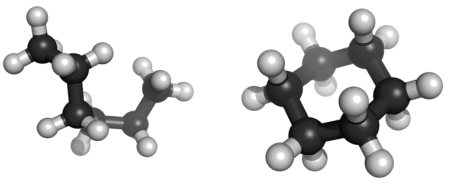


図 2．距離幾何学法による化合物の最

適化構造．左側が hexane、右側が cyclohexane.

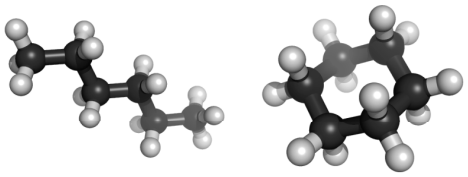


図 3. Gaussian03 による化合物の最適化構造．左側が hexane、右側が cyclohexane.

4.2. 標的タンパク質の薬物結合部位探索，薬物ドッキングの検証

立体構造が既知である25種のタンパク質-リガンド複合体について、薬物結合部位の探索を行った。計算に用いた 25 種類のタンパク質-リガンド複合体を表 1 に示す。格子点を 1Å 間隔で発生させて計算したところ、18 種のタンパク質の薬物結合部位の同定に成功した。成功例のいくつかを図 4 に示す。結晶構造のリガンド分子の位置と計算した疎水性領域が一致していることがわかる。

薬物結合部位の探索に失敗した 7 種について、格子点間隔 0.5Å で計算した。すると、7 種類中 3 種について薬物結合部位を同定することができた（図５）。これら 3 種のタンパク質-リガンド複合体に着目してみると、結合しているリガンド分子が小さいということが共通して言える。3PTB, 1HSL, 2GBP のリガンドの分子量は、それぞれ 121.16, 155.16, 180.15 であり、リガンドの溶媒露出表面積を見ても、312.58Å^2, 334.12Å^2, 341.09Å^2 と小さな分子であることが分かる。よって、リガンド分子の大きさによって格子点間隔を変えて計算する必要があると考えられる。予測の出来なかった 4 種の複合体は、ヘテロ 2 量体で大きなタンパク質や、リガンド分子の親水性が高いものが含まれる。これらの複合体に対しては、計算方法の改良の余地がある。

次に、薬物結合部位の同定に成功した 21 種に対して、疎水性領域と薬物の慣性

Protein-ligand complex	PDB ID	Protein-ligand complex	PDB ID
ACE-lisinopril	1O86	Glucoamylase-acarbose	1AGM
Acetylcholinesterase-aricept	1EVE	HDAC8-aryl hydroxamic acid	1W22
D-Ala D-Ala peptidase-cefotaxime	1CEF	Histidine-binding protein–histidine	1HSL
β-lactamase-clavulanate	1BLC	HIV-1 protease-hydroxyethylene	1AAQ
Carbonic anhydrase-acetazolamide	1JD0	HIV-1 reverse transcriptase-efavirenz	1FK9
COX-1-aspirin	1PTH	Cholesterol oxidase-androstenolone	1COY
COX-2-SC-558	6COX	Myoglobin–imidazole	1MBI
CPA-L-benzylsuccinate	1CBX	Penicillopepsin-pepstatin analogue	1APT
CYP2C9-S-warfarin	1OG5	Thermolysin-Leu-NHOH	4TLN
DHFR–methotrexate	4DFR	Thrombin-NAPAP	1ETS
Enolase–phosphonoacetohydroxamate	1EBG	Trypsin-benzamide	3PTB
FABP–palmitic acid	2IFB	Xanthine oxidase-Mo-pt	1FIQ
GBP-D-glucose	2GBP		

表 1．計算に用いたタンパク質—リガンド複合体の結晶構造

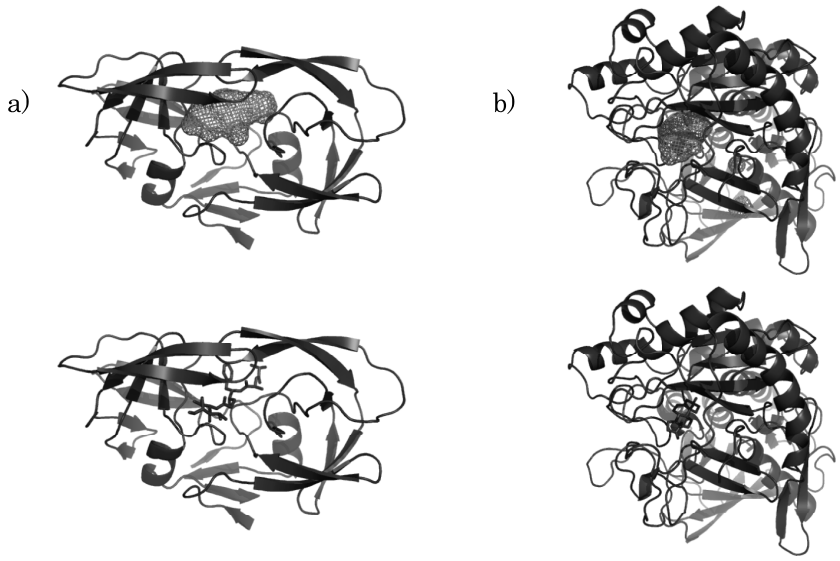


図 4. 計算で同定された薬物結合部位（格子点間隔 1Å で計算）を上側に示し、下側にタンパク質–リガンド複合体の結晶構造を示した。結合部位をメッシュ表示で示し、リガンド分子をスティック表示で示した。a) HIV-1 protease (PDB ID: 1AAQ), b) Cholesterol oxidase (PDB ID: 1COY).

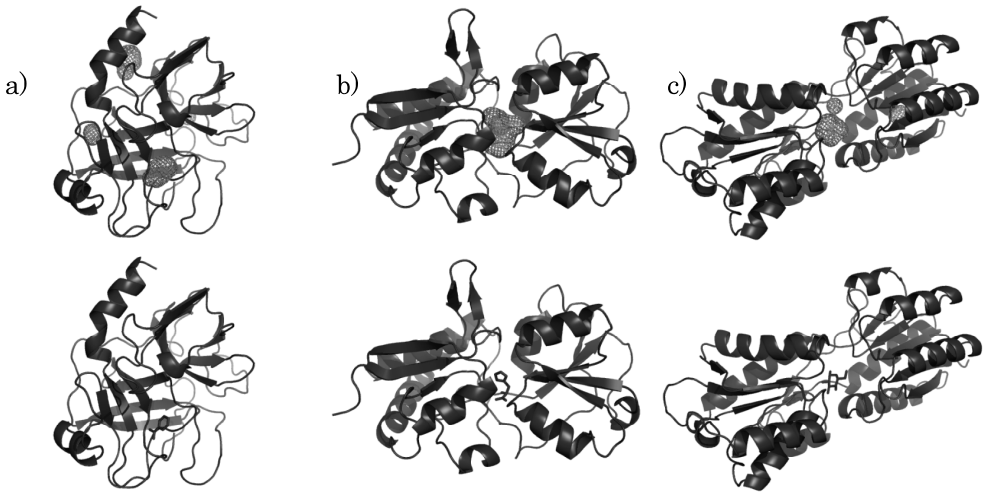


図 5. 計算で同定された薬物結合部位（格子点間隔 0.5Å で計算）を上側に示し、下側にタンパク質–リガンド複合体の結晶構造を示した。表示方法は図 4 と同様。a) β-trypsin (PDB ID: 3PTB), b) Histidine-binding protein (PDB ID: 1HSL), c) D-galactose/D-glucose binding protein (PDB ID: 2GBP).

主軸を合わせる方法を用いて、薬物ドッキングを行った。成功の目安として結晶構造中のリガンド分子の構造と、ドッキング後のリガンド分子の構造のRMSD値を計算した。6 種が RMSD 値 2.0Å 以下に含まれ、10 種が 3.0Å 以下に含まれた。Cholesterol oxidase におけるドッキング結果を図 6 に示す。

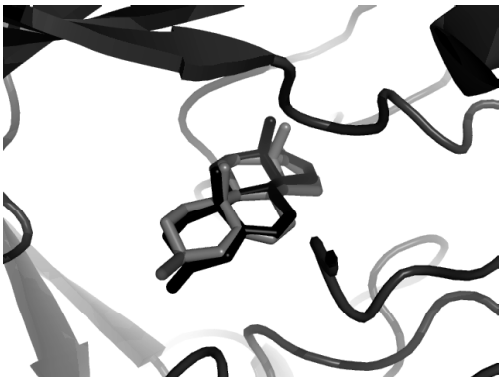


図 6. Cholesterol oxidase（PDB ID: 1COY）に対する androstenolone（スティック表示）のドッキング結果．黒のスティック表示が結晶構造、灰色のスティック表示がドッキング構造．RMSD 値は 0.435Å.

5. 今後の展望

5.1. 化合物の３次元構造構築について

hexane や cyclohexane といった鎖状炭化水素、環状炭化水素では、SMILES から３次元構造を構築することができた。まだヘテロ原子や二重結合、三重結合を含む化合物での検証を行っていないため、様々な化合物でのテストを行い、さらなる精度の向上が必要である。

5.2. 薬物結合部位探索について

リガンド分子の大きさによって、格子点間隔を変えて計算する必要があることが今回の実験によってわかった。リガンド分子の大きさの指標として、溶媒接触表面積を計算し、自動的に計算条件を決定するプログラムを開発する予定である。

5.3. 薬物ドッキングについて

疎水性領域とリガンド分子の慣性主軸を合わせることで、18 種のタンパク質–リガンド複合体のうち 10 種において、結晶構造に近い構造（RMSD 値 3.0 以下）を得ることができた。より結晶構造に近い構造を得るためには、タンパク質–リガンド間の相互作用を加味した計算が必要である。今後は、タンパク質–リガンド間の疎水相互作用と水素結合をもとに、標的タンパク質とリガンド分子の両方をフレキシブルに動かすことのできるドッキングプログラムを作成する予定である。

謝辞

本研究は、独立行政法人情報処理推進機構（IPA）の 2006 年度未踏ソフトウェア創造事業（未踏ユース）の支援を受けて行われた。

参考文献

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, Vol. 28, pp. 235-242, 2000.

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

[3] A.R.リーチ著, 江崎俊之訳. 分子モデリング概説 −量子力学からタンパク質構造予測まで−. 地人書館, pp. 451−458.

[4] N. Yamaotsu and S. Hirono. Determination of the Binding Site of Ligand Molucules in Protein Using Hydrophobic Potential. *The 3rd Annual Meeting of Chem-Bio Informatics Society,* pp. 140-141, 2002.