Vol.2010-BIO-20 No.7
2010/3/4

Tanimoto係数を用いた類似化合物検索のクラスタリングによる高速化手法

グエン　カム　リー[†1]　瀬尾茂人[†2]
竹中要一[†2]　松田秀雄[†2]

概要　医薬品開発の大きな課題の一つとして，新たな薬の候補となる化合物を効果的に発見するため，化合物集合の中から特定のタンパク質に作用する可能性のある候補化合物を計算機によって探索する過程がある．候補化合物の探索には化合物の構造類似性が用いられていることが多いがデータベースに登録されている化合物量が日々増加しており，類似化合物検索の高速化が必要とされている．本研究では，化合物の部分構造情報を数値化した構造キーと，その類似尺度の一つとして Tanimoto 係数を用いた高速な類似化合物検索方法を提案する．提案手法では，化合物集合をクラスタリングするより類似化合物検索を高速化する．また，提案手法を従来手法と比較し，提案手法を評価する．

A method to speedup compound searching by grouped similar compounds using the Tanimoto coefficient

Nguyen Cam Ly,[†1]　Shigeto Seno,[†2]
Yoichi Takenaka[†2] and Hideo Matsuda[†2]

Abstract: In order to find compounds as candidates for drugs, one of the major problems of drug development, it is necessary to find compounds that might affect the proteins by using computer. Analysis of the compound structure similarity is a widely used method to discover candidate compounds. Nowadays, as the volume of compounds database is becoming rapidly increased, it is important to accelerate the compounds searching. In this paper, we propose a method to speed up compound searching by using structure key and Tanimoto coefficient. We use some methods for grouping similar compounds and evaluate our method by comparing it with the algorithm normally used.

1. はじめに

薬とは異常な働きをするタンパク質に作用するが他のタンパク質に作用しない化合物であり，そのような特定のタンパク質に作用する薬の候補となる化合物を発見することが，医薬品開発の大きな課題である．新薬候補化合物の探索は最終的に実験に頼ることになるが全てのタンパクと化合物の組み合わせを実験するのは時間的にも費用的にも困難である．そのため，計算機によって，大量の化合物の中から薬の候補になる可能性がある化合物を探索してから実験を行う．化合物の働きは，化合物を構成する原子の種類とその結合の仕方という構造によることが大きいため，薬の候補化合物を探索には化合物の構造類似性が用いられることが多い．

合成された化合物の情報を管理するデータベースは PubChem[7]，DrugBank[6] をはじめ，様々なデータベースが存在している．しかし，データベースに登録されている化合物数は増加の一途を辿っている．例を挙げると，2007 年 7 月において PubChem のデータベースに登録された化合物数は約 1000 万件であるが，2009 年 11 月において登録された化合物数は約 2600 万件すなわち約 2.6 倍も増加している．そのため，類似化合物検索を高速化することが重要となっている．

計算機を用いて類似化合物検索を行うために，化合物の構造を数値で表現した記述子及び類似性評価尺度が必要である．その記述子と類似性評価尺度を用いた類似化合物探索の高速化には様々な手法[1),2),8)] があるがいずれも問題点がある．特に，データベースが増大しており，効果の高い類似化合物検索手法が必要である．本研究では，化合物の部分構造情報を数値化した構造キーと，その類似尺度の一つとして Tanimoto 係数を用いた高速な類似化合物検索方法を提案する．

2. 記述子と類似性評価尺度

類似化合物検索を計算機で行うためには，化合物の情報の数値化が必要である．代表的

†1 大阪大学基礎工学部情報科学科
Department of Information and Computer Science School of Engineering Science, Osaka University
†2 大阪大学大学院情報科学研究科バイオ情報工学専攻
Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University

な構造キーには Unity 2D fingerprint[3], Similog key[4], MACCS Key[5] などさまざまなものが提案されているが，いずれも化合物の構造をビット列で表したものである．本研究では MACCS Key を利用する．MACCS Key は化合物構造として 166 種類の部分構造を設定し，166 ビットのビット列として表現することで，化合物が各ビットに指定された部分構造を持っているかどうかがわかる．

類似化合物検索では，構造がどの程度似ているのか表わす類似性評価尺度を定義する必要がある．類似度の一つとしては Tanimono 係数がよく用いられる．Tanimoto 係数は，ビット列 X, Y に対して次式で定義される．

$$TC(X,Y) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \quad (1)$$

ただし，

$|X|$: X 中で 1 になっているビット数

$|Y|$: Y 中で 1 になっているビット数

$|X \cap Y|$: X と Y で共通して 1 になっているビット数

Tanimoto 係数は PubChem のデータベースをはじめ様々なデータベースの類似化合物検索で広く用いられる．

3. 従来手法

化合物データベースに登録されている化合物数が日々増加しているため，類似化合物検索の高速化が必要になっている．構造キーを用いた類似化合物検索を高速化するための方法としては，主に構造キーのビット数の圧縮を行う方法[1] と，類似度の性質により計算範囲を絞り込む方法[2),8)] の 2 つの方向がある．

構造キーのビット数の圧縮を行う方法は，ビットを重ねることで数百もしくは数千のビットを数分の 1 にする方法である．ビット数を圧縮することにより，ビット比較の計算時間を短縮することができる．しかし，ビットを圧縮すると情報が失われてしまうという問題がある．

2 つ目の方法は，類似性尺度の性質より導出される条件を用いて類似度計算を行う化合物数を削減する方法である．類似化合物のデータベース検索では，クエリとして与えた 1 つの化合物との Tanimoto 係数の値が指定された閾値より高い化合物を出力するということが行われる．閾値を満たす可能性のある化合物だけ類似度計算を行えば，無駄な計算をしなくてよい．具体例として，Tanimoto 係数の数学的な性質を用い計算範囲を限定する方法を挙げる[2]．式 (1) より，$|X|$,$|Y|$ の値を固定した場合，Tanimoto 係数 TC が最大となるのは $|X \cap Y| = min(|X|, |Y|)$ の時である．また，分母は $|X| + |Y| - min(|X|, |Y|) = max(|X|, |Y|)$ となる．つまり TC の上限は式 (2) のようになる．

$$TC \leq \frac{min(|X|, |Y|)}{max(|X|, |Y|)} \quad (2)$$

ビット列 X で表わす化合物がクエリとして与えられた時，Tanimoto 係数が T 以上となるような化合物の構造キー Y の必要条件を構造キーの 1 のビット数を用いて明らかにする．$|X| \geq |Y|$ の場合と $|X| < |Y|$ の場合に分けると，式 (2) より，以下の式 (3) が成り立つ．

$$|X| \times T \leq |Y| \leq \frac{|X|}{T} \quad (3)$$

すなわち，データベース内の化合物の構造キーの 1 のビット数さらに式 (3) を満たす場合のみ計算を行うようにすることで，計算回数の削減を行うことができる．さらに、構造キーを 2 つに分割することにより，計算範囲をより絞り込む研究[8] がある．

これまで述べたように，1 のビット数の総数を比較することで類似度とる計算の範囲を絞り込むことができる．しかし，クエリ化合物が与えられた時，式 (3) を満たすような化合物がクエリ化合物との類似度が高いとは言えない．そのため，従来手法による絞り込み効果が弱い．実際では，DrugBank のデータベースに登録されている化合物をクエリ化合物，PubChem のデータベースに登録されている化合物を検索対象すると，クエリ化合物との Tanimoto 係数が 0.9 以上になる化合物数は式 (3) を満たす化合物数の 0.1% に満たない．つまり，99.9% 以上の類似計算が無駄になる．また，従来手法はクエリ化合物と高い類似度を持つ化合物を検索する際には有効であるが，比較的類似度の低い化合物を検索する際には，ほとんど効果が期待できないという問題点もある．化合物データベースに登録されている化合物数は増加の一途を辿っており，従来手法よりも効果的な絞り込みの方法が必要である．

4. 提案手法

Tanimoto 係数を利用した類似化合物検索では，クエリ化合物と閾値に指定する Tanimoto 係数の値を入力し，類似度が指定された閾値より高いデータベース内の化合物の情報を出力することが行われる．

本研究では，3 章で述べた検索範囲を絞り込む手法を応用し，検索をさらに高速化するための新しい手法を提案する．

4.1 提案手法の概要

3 章で述べたように 1 のビット数だけで検索範囲を絞り込む条件は弱い．そのため，1 のビット数の総数によるだけでなく，構造キーによる 1 のビットの位置に従って検索範囲を絞り込んだ方がいいと考える．前処理として，共通して 1 になるビット数が多い化合物をグループ化し，グループの代表化合物およびグループ内の化合物全体との類似度の最小値を取り出す．クエリ化合物を与えた時，その化合物と各グループの代表化合物との類似度及び算出されたグループ内の化合物の類似度の最小値により，クエリ化合物とグループ内の化合物の類似度の範囲を算出することができる．その範囲に従って，化合物のグループは以下の 3 つのパターンに分類される．

- IN グループ：全ての化合物がクエリ化合物と似ているグループ
- OUT グループ：全ての化合物がクエリ化合物と似ていないグループ
- BOUNDARY グループ：上記以外のグループ

そして実際にクエリ化合物とグループ内の化合物で個別に類似度の計算を行う必要があるのは 3 つ目に属するグループのみである．

4.2 アルゴリズム

提案アルゴリズムは，化合物グループを作成する「前処理」と，クエリ化合物が与えられた時に実行する「検索の処理」の 2 段階で構成される．以下順に説明する．

前処理

検索を高速化するための前処理について説明する．前処理の対象は，データベース内の化合物集合である．

- Step P1:
  1 のビット数の総数によって化合物をソートする．
- Step P2:
  同じ 1 のビット数を持つ化合物を，構造キーの類似度に基づいてグループ化する．1 のビット数が i の j 番目のグループを G_{ij} と呼ぶ．各グループは，代表として一つの化合物を持ち，それを O_{ij} とする．また，代表化合物とそのグループ内の任意の化合物との Tanimoto 係数の最小値を記憶しておく．その最小値を $TCmin_{ij}$ とする．

Step P1 と Step P2 終了後のデータベース内の化合物の概念図を図 1 の上部分に示す．図 1 の左上部分は，Step P1 終了後の 1 のビット数について昇順になっていることを表す．図 1 の右上部分は，1 のビット数が i の化合物集合がグループ化されたことを表す．

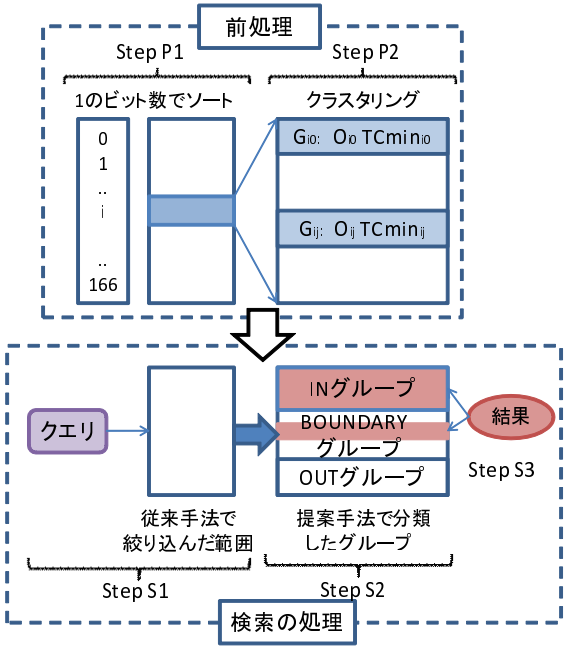
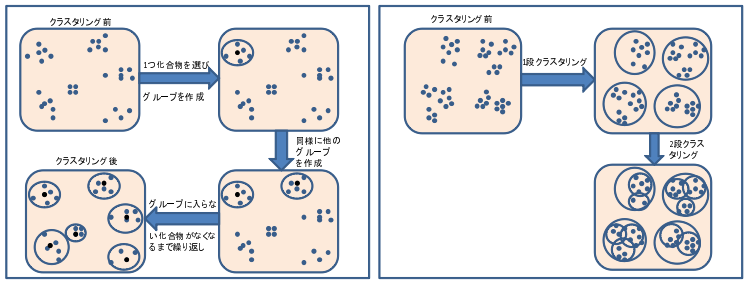


図 1 提案手法の概念

Step P2 において化合物のグループ化，すなわち化合物のクラスタリングを行っている．提案アルゴリズムは特定のクラスタリングアルゴリズムを前提としていない．そこで以下に本研究の性能評価に用いた 3 種類のクラスタリング手法を記す．

- メソッド 1
  Tanimoto 係数の閾値の最小値 $TCmin$ を与え，各グループ G_{ij} に対し，Tanimoto 係数 $TCmin_{ij}$ の値が TCmin 以上になるようにクラスタリングを行う．イメージ図を図 2(a) に示す．まず，クラスタリング対象データベースから 1 つの化合物を化合物グループの代表化合物を選択し，データベースにある同じ 1 のビット数を持つ化合物集合の中に Tanimoto 係数が $TCmin$ より高い化合物を取り出し，選択した化合物にあるグループに入れる．グループに配属されない化合物に対し，また任意の代表化合物を選び，述べた動作と同様に全ての化合物がグループに入るまで繰り返す．

Vol.2010-BIO-20 No.7
2010/3/4



(a) メソッド 1　　(b) メソッド 2

図 2　クラスタリングメソッド

– メソッド 2
多段クラスタリングを行った．イメージ図を図 2(a) に示す．詳細は以下のようになる．
(1)　メソッド 1 と同様に，Tanimoto 係数の閾値の最小値 $TCmin_1$ を与え，各グループ G_{ij} に対し，Tanimoto 係数 $TCmin_{ij}$ の値が $TCmin_1$ 以上になるようなクラスタリングを行う．
(2)　$TCmin_2$ ($TCmin_2 > TCmin_1$) を与え，Tanimoto 係数の閾値の最小値として使用する．Step1 でクラスタリングした各グループをクラスタリング対象データとして扱い，さらにクラスタリングする．
(3)　同様に，$TCmin_k$ ($TCmin_k > TCmin_{k-1}$) を与え，Tanimoto 係数の閾値の最小値として使用する．Step(k-1) でクラスタリングした各グループをクラスタリング対象データとして扱い，さらにクラスタリングする．

– メソッド 3
k-means 法[9] を用いて，クラスタリングを行う．クラスタリング対象，すなわち同じ 1 のビット数を持つ化合物集合から k 個の化合物を選択する．クラスタリング対象の各化合物に対し，選択されえた k 個の化合物との比較を行い，類似度が最も高い化合物と同じグループに入れる．分類された各グループに対し新たな代表ビット列を作成する．つまり，グループ内の全ての化合物の構造キーにおいて 1 になっている率が高いビットを 1，残りのビットを 0 にするビット列を代表ビット列として作りだす．ただし，代表ビット列の 1 のビット数は化合物の構造キーの 1 のビット数と同じようにする．新たな代表ビット列を選択した後，新しいグループを作成する．その過程を i 回繰り返す．

検索の処理

化合物 Q がクエリとして与えられた時，Tanimoto 係数が T 以上となるような化合物を検索する処理のアルゴリズムを記す．

– Step S1
クエリとして与えられた化合物の構造キーの 1 のビット数を計算する．その 1 のビット数と指定された閾値から式 (3) に基づいて，比較対象の化合物の構造キーの 1 のビット数の範囲を計算する．

– Step S2
絞り込まれた範囲に対し，クエリ化合物を各グループの代表化合物と比較し，Tanimoto 係数を計算する．その Tanimoto 係数及びグループ内の化合物との Tanimoto 係数の閾値，与えられた類似度の閾値 T により，グループを以下の 3 種類に分類する．
* IN グループ：次の条件を満たしたグループである．

$$|Q \cap O_{ij}| \geq (|O_{ij}| + |Q|) \times \frac{T}{1+T} + |O_{ij}| \times \frac{1 - TCmin_{ij}}{1 + TCmin_{ij}} \qquad (4)$$

グループ内の全ての化合物がクエリ化合物と閾値以上の類似度を持つ．
* OUT グループ：次の条件を満たしたグループである．

$$|Q \cap O_{ij}| < (|O_{ij}| + |Q|) \times \frac{T}{1+T} - |O_{ij}| \times \frac{1 - TCmin_{ij}}{1 + TCmin_{ij}} \qquad (5)$$

グループ内の全ての化合物とクエリ化合物の類似度が閾値未満である．
* BOUNDARY グループ．上記の 2 つの条件を満たさないグループである．クエリ化合物に似ている化合物も似ていない化合物も含む可能性がある．

– Step S3
分類ごとに以下の処理を実施する．
* IN グループは，グループ内の化合物とクエリ化合物との類似度の計算を省略して，グループ内の全ての化合物を結果として出力する．
* OUT グループは，グループ内の化合物との計算を行わず出力もしない．

* BOUNDARY グループについては，グループ内の全ての化合物とクエリ化合物との類似性を計算し，あらかじめ入力として与えられた閾値より大きな類似度を持つ化合物のみを出力する．

検索の処理を図 1 の下部分に示す．図 1 の左下部分はクエリ化合物と閾値が与えられた時，式 (3) により検索範囲を絞り込むことを表す．図 1 の右下部分は Step2 終了後，グループが 3 種類に分類されることを表す．

ただし，メソッド 2 においては，クエリ化合物を与えた時，大きいグループからチェックを始めて，さらに内部の小さいグループへと分岐探索と類似した処理を行う．

4.3 アルゴリズムの正当性

本節では，記述したアルゴリズムの正当性を証明する．すなわち，提案アルゴリズムを用い，クエリ化合物と閾値が与えられた時，化合物データベースから類似度が閾値以上になる全ての化合物を出力することを証明する．そのため，IN グループはグループ内の全ての化合物がクエリ化合物の Q と閾値 T 以上の類似度を持つことと，OUT グループはグループ内の全ての化合物がクエリ化合物 Q と閾値未満の類似度しか持たないいことを証明する．

まず，幾つかの補題を導入する．

補題 *1*：任意の化合物の構造キー X, Y に対し，次の式が成り立つ

$$|X| \geq |X \cap Y| \tag{6}$$

$$|X \cap Y| + |X \cap \bar{Y}| = |X| \tag{7}$$

$\bar{X}$ は X の補数すなわち X の各ビットを反転させたビット列である．$X \cap Y$ は X，Y で共通したビット列を表す．

以上で述べた式は自明であるため補題 1 の証明を省略する．

補題 *2*：任意の化合物の構造キー O_{ij}, E, Q に対し，次の不等式が成り立つ

$$|Q \cap E| \geq |Q \cap O_{ij}| - |O_{ij}| + |O_{ij} \cap E| \tag{8}$$

$$|Q \cap E| \leq |Q \cap O_{ij}| + |E| - |O_{ij} \cap E| \tag{9}$$

証明*:*

$|Q \cap E| \geq |(Q \cap E) \cap O_{ij}|$ ((6) より)
$= |(Q \cap O_{ij}) \cap E|$
$= |Q \cap O_{ij}| - |Q \cap O_{ij} \cap \bar{E}|$ ((7) より)
$\geq |Q \cap O_{ij}| - |O_{ij} \cap \bar{E}|$ ((6) より)
$= |Q \cap O_{ij}| - |O_{ij}| + |O_{ij} \cap E|$ ((7) より)

よって，不等式 (8) が成り立つ．

$|Q \cap E| = |(Q \cap E) \cap O_{ij}| + |(Q \cap E) \cap \bar{O}_{ij}|$ ((7) より)
$= |(Q \cap O_{ij}) \cap E| + |(E \cap \bar{O}_{ij}) \cap Q|$
$\leq |Q \cap O_{ij}| + |E \cap \bar{O}_{ij}|$ ((6) より)
$= |Q \cap O_{ij}| + |E| - |E \cap O_{ij}|$ ((7) より)

よって，不等式 (9) が成り立つ．

アルゴリズムの正当性の証明

各グループ内における化合物の類似度の関係は以下のように表わされる．前処理の規則より，グループ G_{ij} におけて代表化合物 O_{ij} を与える時，全ての化合物は Tanimoto 係数が $TCmin_{ij}$ 以上であるため，グループ G_{ij} にある任意の化合物 E が次の条件を満たす．

$$TCmin_{ij} \leq \frac{|E \cap O_{ij}|}{|E| + |O_{ij}| - |E \cap O_{ij}|}$$

式変形により式 (10) が導出される．

$$|E \cap O_{ij}| \geq (|E| + |O_{ij}|) \times \frac{TCmin_{ij}}{1 + TCmin_{ij}} \tag{10}$$

IN グループの条件 (式 (4) で表わす) を満たすグループ G_{ij} の中の任意の化合物 E がクエリ化合物 Q との類似度が閾値 T 以上であることを証明する．それは，次の不等式が成り立つことを証明すればよい．

$$T \leq \frac{|E \cap Q|}{|E| + |Q| - |E \cap Q|} \tag{11}$$

一方不等式 (11) が次の不等式に同様である．

$$|E \cap Q| \geq (|E| + |Q|) \times \frac{T}{1 + T} \tag{12}$$

そのため，不等式 (12) が成り立つことを証明する．

$|E \cap Q| \geq |Q \cap O_{ij}| - |O_{ij}| + |O_{ij} \cap E|$ ((8) より)
$\geq (|O_{ij}| + |Q|) \times \frac{T}{1+T} + |O_{ij}| \times \frac{1-TCmin_{ij}}{1+TCmin_{ij}} - |O_{ij}| + |O_{ij} \cap E|$ ((4) より)
$\geq (|O_{ij}| + |Q|) \times \frac{T}{1+T} + |O_{ij}| \times \frac{1-TCmin_{ij}}{1+TCmin_{ij}} - |O_{ij}| + (|E| + |O_{ij}|) \times \frac{TCmin_{ij}}{1+TCmin_{ij}}$ ((10) より)
$= (|O_{ij}| + |Q|) \times \frac{T}{1+T} + |O_{ij}| \times \frac{1-TCmin_{ij}}{1+TCmin_{ij}} - |O_{ij}| + (2 \times |O_{ij}|) \times \frac{TCmin_{ij}}{1+TCmin_{ij}}$ ($|E| = |O_{ij}|$ より)
$= (|O_{ij}| + |Q|) \times \frac{T}{1+T}$

$= (|E| + |Q|) \times \frac{T}{1+T}$ $(|E| = |O_{ij}|$ より)

従って，式 (4) を満たす時，グループ G_{ij} に属している全ての化合物がクエリ化合物との類似度が閾値 T 以上である．

OUT グループの条件 (式 (5) で表わす) を満たすグループ G_{ij} の中の任意の化合物 E がクエリ化合物 Q との類似度が閾値 T 未満であることを証明する．それは，次の不等式が成り立つことを証明すればよい．

$$T > \frac{|E \cap Q|}{|E| + |Q| - |E \cap Q|} \tag{13}$$

一方不等式 (13) が次の不等式に同様である

$$|E \cap Q| < (|E| + |Q|) \times \frac{T}{1+T} \tag{14}$$

そのため，不等式 (14) が成り立つことを証明する．

$|E \cap Q| \leq |Q \cap O_{ij}| + |E| - |O_{ij} \cap E|$ ((9) より)

$< (|O_{ij}| + |Q|) \times \frac{T}{1+T} - |O_{ij}| \times \frac{1-TCmin_{ij}}{1+TCmin_{ij}} + |E| - |O_{ij} \cap E|$ ((5) より)

$\leq (|O_{ij}| + |Q|) \times \frac{T}{1+T} - |O_{ij}| \times \frac{1-TCmin_{ij}}{1+TCmin_{ij}} + |E| - (|E| + |O_{ij}|) \times \frac{TCmin_{ij}}{1+TCmin_{ij}}$ ((10) より)

$= (|O_{ij}| + |Q|) \times \frac{T}{1+T} + |O_{ij}| \times \frac{1-TCmin_{ij}}{1+TCmin_{ij}} - |O_{ij}| + (2 \times |O_{ij}|) \times \frac{TCmin_{ij}}{1+TCmin_{ij}}$ $(|E| = |O_{ij}|$ より)

$= (|O_{ij}| + |Q|) \times \frac{T}{1+T}$

$= (|E| + |Q|) \times \frac{T}{1+T}$ $(|E| = |O_{ij}|$ より)

従って，式 (5) を満たす時，グループ G_{ij} に属している全ての化合物がクエリ化合物との類似度が閾値 T 未満である．

以上により，本アルゴリズムの Step S2 で分類された各グループについて，Step S3 の処理を行うことによりデータベースからクエリ化合物との類似度が閾値 T 以上である全ての化合物を出力することを証明した．

5. 実験と結果

5.1 実　　験

提案手法の有用性を検証するための 2 つの実験を行った．実験 1 では，提案手法を検索条件の閾値の変化に伴い，各メソッドの効果を検証する．実験 2 は，将来化合物数が増加しても提案手法の効果があることを検証する．

実験 1 では，1 つの化合物データベースを検索対象，もう 1 つのデータベースに含まれる化合物をクエリ対象として用い，閾値を変化させ，クラスタリング法による検索の計算時間と類似度の計算回数を評価する．実験 2 では，閾値を固定し化合物データベースに含まれる化合物数の変化による計算時間を評価する．

ここで，従来手法とは式 (3) のことを指す．

使用した計算機はインテル Core 2 Duo 2.40GHz，メモリ 2.0GB で，C 言語によりプログラミングを行った．

実験 *1*

2009 年 11 月において PubChem に登録されている化合物 26,026,918 件，DrugBank に登録されている全ての化合物データ 4,886 件に対して従来手法と提案手法の 3 種類を適用した．PubChem に登録される化合物を検索対象，DrugBank に登録される化合物をクエリ化合物として実験を行った．構造キーは MACCS Key を用いている．

4.2 節で述べた各クラスタリング法のパラメータを次のように設定した．

メソッド 1: $TCmin$:=0.7

メソッド 2: 3 段クラスタリングを行った．

$TCmin_1 := 0.5, TCmin_2 := 0.7, TCmin_3 := 0.9$

メソッド 3: i:= 10, k:= $n^{0.8}$．ただし，n はグループ化対象の化合物の総数である．

閾値 T を 0.6 から 0.95 まで 0.05 間隔で変化させた従来手法と本手法を適用した場合の計算回数と計算時間について比較を行った．

実験 *2*

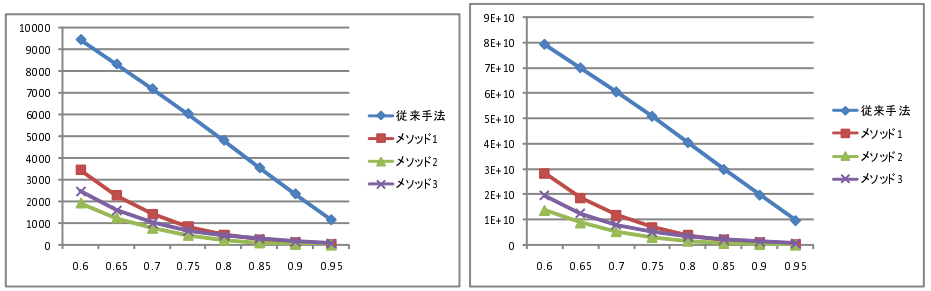
検索対象化合物数を変化させ，従来手法およびメソッド 1 を適用した場合の計算時間の変化の考察を行った．クエリ化合物として DrugBank に登録されている全ての化合物を扱う．検索対象化合物は，2009 年 11 月まで PubChem に登録された全ての化合物とその $1/2, 1/4, 1/8, .., 1/128$ の 8 つのデータベースを用いる．閾値 T に 0.95 を与える．クラスタリング法はメソッド 1 を用いる．各データベースに対し，$TCmin$ を 0.55 から 0.70 まで 0.01 間隔で変化させた提案手法を適用し，計算時間が最小になる $TCmin$ を選ぶ．

5.2 結　　果

実験 *1*

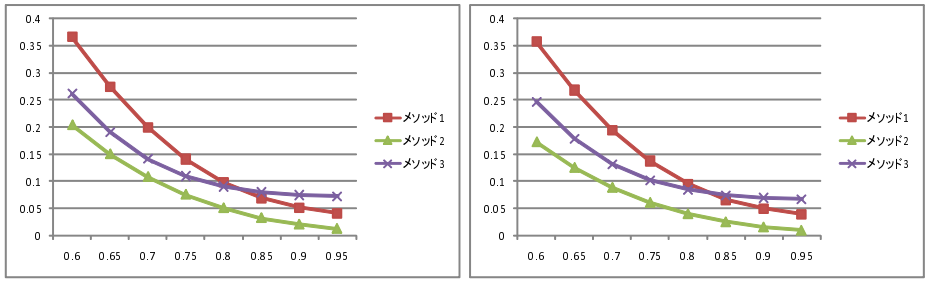
提案手法，従来手法による計算回数，計算時間の結果はそれぞれ図 3(a)，図 3(b) のようになる．横軸は閾値を表しており，縦軸はそれぞれ計算時間（秒），計算回数を表している．

Vol.2010-BIO-20 No.7
2010/3/4



(a) 計算時間　　(b) 計算回数

図 3　各手法における計算時間と計算回数の比較



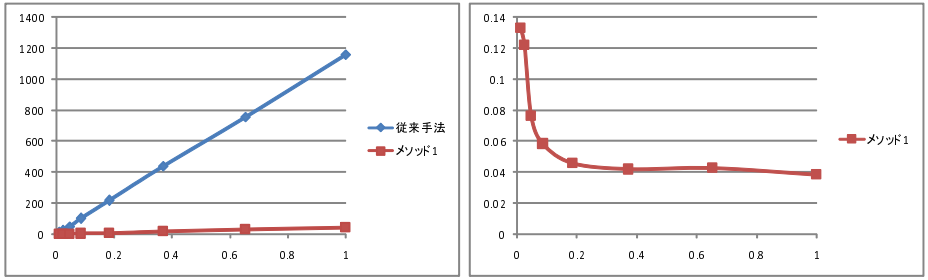
(a) 計算時間の削減率　　(b) 計算回数の削減率

図 4　従来手法に対する提案手法の計算時間と計算回数の削減率

図 3(a)，図 3(b) より，すべての閾値において，3 種類の提案手法は従来手法よりも計算回数と計算時間が削減されている．

「従来手法の計算時間 ÷ 提案手法の計算時間」で従来手法と比較した提案手法の計算回数計，計算時間の割合を算出すると，結果はそれぞれ図 4(a)，図 4(b) のようになる．

実験 *2*



(a) 計算時間　　(b) 計算時間の削減率

図 5　検索対象化合物数の増加に伴う計算時間の推移

従来手法，提案手法を適用した化合物数の変化による計算時間の結果は図 5(a) のようになる．横軸は PubChem のデータベース比を表しており，縦軸は計算時間を表している．図 5(a) より，全てのデータベースにおいて，メソッド 1 による提案手法は従来手法よりも計算時間が削減されている．

「従来手法の計算時間 ÷ 提案手法の計算時間」で従来手法と比較した提案手法の計算時間の割合を算出すると，結果はそれぞれ図 5(b) のようになる．図 5(b) より，化合物数が増加しても時間削減率は増加しないことが分かる．つまり，化合物データベースが大きくなっても，提案手法の効果が低くならない．

6. 考　　察

3 章で述べたように，化合物の構造キーの 1 のビット数だけで検索範囲を絞り込む条件だけでは不十分であり，無駄な類似計算が多い．提案した手法が実験で用いた全ての閾値，全てのデータベースに対し，類似計算時間及び類似検索回数を削減することができる．それは，検索対象化合物集合をグループ化してから検索を行うため，明らかに類似している化合物と明らかに類似していない化合物に対して比較を行わなくて良いためである．

図 4 により，多段クラスタリングの方が 1 段だけのクラスタリングより効果が高いことがわかる．分岐探索のように，明らかに比較の必要がない大きなグループを除いてから小さいグループを調査するので検索回数を削減することができる．本実験では 3 段クラスタリ

ングまでしか行わなかったが，より多段階にクラスタリングを行うと効果が高くなると期待できる．

図 4 により，閾値が 0.8 以下の場合メソッド 3 の方がメソッド 1 より効果が良いが閾値が 0.85 以上の場合にはメソッド 1 の方がメソッド 3 より効果が高い．そのため，閾値によって適切なクラスタリング法を用いると本手法の効果が高くなると考えられる．

図 5 により，化合物が多くなっても提案手法の効果が失わない．そのため，データベースに登録されている化合物数がさらに急激に増加することが見込まれる将来においても提案手法は有効である．

7. おわりに

本研究では，Tanimoto 係数の数学的性質に基づいて化合物をあらかじめクラスタリングし，検索時の計算範囲を絞り込むことで検索を高速化する手法を提案した．また，提案手法を実際の化合物データベースに対し適用しその有効性を示した．提案手法では類似化合物検索時間を削減できたことを示した (閾値 0.95 の時メソッド 3 において従来手法より約 77 倍速くなった) 上で，将来のデータベース増大にも対応していることを示した．

参 考 文 献

1) S. Swamidass,*et al.* Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *Journal of Chemical Information and Modeling*, Vol. 47, pp. 952-964, 2007.
2) S.Swamidass, *et al.* Bounds and algorithms for fast exact searches of chemical fingerprints in linear and sublinear time. *Journal of Chemical Information and Modeling*, Vol. 47, pp. 302-317, 2007.
3) J. W. Raymonda, *et al.* Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. Journal of Computer-Aided Molecular Design, Vol. 16, pp. 59-71, 2002.
4) A. Schuffenhauer, *et al.* Similarity metrics for ligands reflecting the similarity of the target proteins. Journal of Chemical Information and Computer Sciences, Vol. 43, pp. 391-405, 2003.
5) L. Xue, *et al.* Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *Journal of Chemical Information and Computer Sciences*, Vol. 43, pp. 1218-1225, 2003.
6) DrugBank: Internet address. http://www.drugbank.ca/.
7) E. W. Sayers, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research*, Vol. 37, pp. D5-D15, 2009.
8) T.Shimizu, *et al.* A Method for Reducing Bounds of Compound Search by Dividing Structure Key, *Proceedings of the 2008 Annual Conference of the Japanese Society for Bioinformatics (JSBi2008),* P077/T01, Senri-Chuo , 2008.
9) J.B.MacQueen.Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1967.