2008－BIO－14 (2)
2008／9／18

構造キーの分割によるTanimoto係数を用いた化合物検索の計算範囲の絞り込み手法

清水隆史*，木戸善之†，瀬尾茂人*，竹中要一*，松田秀雄*

概要 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たどっている．化合物データベースの類似性検索では，構造キーとTanimoto係数を用いた検索がよく利用されるが，化合物数の増加に伴い検索にかかる時間が増加する．そこで本研究では，構造キーを分割することで類似性検索の計算範囲を絞り込む手法を提案し，実際の化合物データベースでの検索の際の計算回数を測定することによりその有効性を示した．

A method for refinement of the calculation range by divided structure key of the compound search using Tanimoto coefficient

Takashi Shimizu*, Yoshiyuki Kido†, Shigeto Seno*, Yoichi Takenaka* and Hideo Matsuda*

Abstract The amount of compounds in public databases goes on increasing. A structure key and Tanimoto coefficient are often used for similarity searches in compound databases. As the number of compounds increases, the search time increases. In this research, we propose a method for refinement of the calculation range by divided structure key, and show the effectiveness by measuring the number of calculation with the data of database.

1 はじめに

化合物のデータは日々増加し続けており，公共データベースであるPubChem[1]には現在1900万を超える膨大な化合物データが登録されている．PubChemの化合物データが2007年7月から12月までの5ヶ月で800万個増加していることや，理論的に合成可能な化合物数は10^{60}個を超えると言われていることから，今後もその数は増加し続けると考えられる．化合物データベースでは構造類似性検索がよく利用されるが，化合物数の増加に伴い検索にかかる時間が増加する．

2 化合物の構造比較

2.1 記述子と類似性評価尺度

化合物の構造を計算機上で扱うため，その構造の数値化が行われている．構造を数値化したものとして，原子数や分子量などの値や，構造キーなどの記述子が利用されている．中でも化合物の構造を表現するものとしては，よく構造キーが用いられる．構造キーは，対象化合物における部分構造の有無を0，1のビットで表現したものの集合（ビット列）である．部分構造としてどのような特徴を定義するのかによって，構造キーのビット数や，構造キーから得られる情報が変化する．MACCS Key[2]はMDL Information Systemsによって運営されているデータベースに採用されている構造キーである．MACCS Keyは166種類の部分構造の有無をビット列として表現している．

構造キーのビット列を比較することにより化合物の類似性を評価できる．類似性の評価尺度としてTanimoto係数[3]やユークリッド距離などがある．Tanimoto係数は，平均的なパフォーマンスがよく[3]，構造キーの類似性を評価する方法としてよく用いられる．

*Department of Bioinformatic Engineering,Graduate School of Information Science and Technology, Osaka University
†The Center for Advanced Medical Engineering and Informatics, Osaka University

2.2 類似度計算時間の短縮法

構造キーを用いた類似度計算時間の短縮法として，ビット数を圧縮する方法と類似度の計算回数を削減する2つの方向でそれぞれ研究が行われてきた．

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にすることにより，ビット比較の計算時間の短縮を図っている．構造キーを圧縮することにより情報が失われてしまうが，圧縮前のビット数を推測することで情報の損失を最小限に抑えようという研究 [4] もある．

2つ目は，類似性尺度の性質から，計算の必要のないものは計算しないようにすることで，類似度計算のための比較回数の削減を行う研究である．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．通常，データベース内のすべての化合物との類似度を計算するが，閾値を満たす可能性のある化合物とだけ類似度計算を行うようにすれば，無駄な計算をしなくてよい．具体例として，構造キーの1のビット数を利用することで Tanimoto 係数の計算範囲を限定する研究 [5] がある．Tanimoto 係数は次のように定義されている．

$$Tc = \frac{c}{a+b-c} \tag{1}$$

Tc は化合物 A と B の Tanimoto 係数，a，b は化合物 A，B の構造キーの1のビット数，そして c は化合物 A，B の構造キーで共通して1になるビット数を表している．化合物 A 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 B 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．式 (1) より，Tc が最大となるのは $c = min(a,b)$ のときである．min(a,b) は a と b のうち，小さい方の数を表している．

$a \geq b$ のとき，Tc が最大となるのは $c = min(a,b) = b$ のときで，$Tc = b/a$ となる．$T \leq Tc$ より，$T \leq b/a$ を満たす必要がある．これを式変形することで b の範囲は $aT \leq b$ となる．$a \leq b$ のとき，Tc が最大となるのは $c = min(a,b) = a$ のときで，$Tc = a/b$ となる．$T \leq Tc$ より，$T \leq a/b$ を満たす必要がある．これを式変形することで b の範囲は $b \leq a/T$ となる．

以上より，比較対象化合物 B の範囲を式 (2) のように絞り込むことができる [5]．

$$aT \leq b \leq \frac{a}{T} \tag{2}$$

データベース内の化合物の構造キーの1のビット数が式 (2) 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．

3 類似度検索の高速化

従来研究では，構造キー全体の1のビット数により計算範囲の絞り込みが行われている．そこで本研究では，構造キーを2つに分割することで，計算範囲をより絞り込む手法を提案する．

a_1, a_2, b_1, b_2 を化合物 A，B の構造キーを2つに分割したときのそれぞれの1のビット数，c_1, c_2 を化合物 A，B の構造キーを2つに分割したときにそれぞれ共通して1になるビット数とする．式 (1) は，$a_1, a_2, b_1, b_2, c_1, c_2$ を用いて表すと，次のように表せる．

$$Tc = \frac{c_1 + c_2}{a_1 + a_2 + b_1 + b_2 - (c_1 + c_2)} \tag{3}$$

式 (3) より，c_1, c_2 が大きければ大きいほど Tc は大きくなる．つまり，Tc が最大になるのは，$c_1 = min(a_1, b_1)$ かつ $c_2 = min(a_2, b_2)$ のときである．よって，Tc は以下の条件を満たす．

$$\begin{aligned} Tc &\leq \frac{min(a_1,b_1) + min(a_2,b_2)}{a_1 + a_2 + b_1 + b_2 - (min(a_1,b_1) + min(a_2,b_2))} \\ &= \frac{min(a_1,b_1) + min(a_2,b_2)}{(a_1 + b_1 - min(a_1,b_1)) + (a_2 + b_2 - min(a_2,b_2))} \end{aligned}$$

a と b のうち大きい方の数を max(a,b) と表すと，$a_1 + b_1 - min(a_1,b_1) = max(a_1,b_1)$，$a_2 + b_2 - min(a_2,b_2) = max(a_2,b_2)$ となるので，Tanimoto 係数の上限は式 (4) のようになる．

$$Tc \leq \frac{min(a_1,b_1) + min(a_2,b_2)}{max(a_1,b_1) + max(a_2,b_2)} \tag{4}$$

構造キーを2つに分割することでさらに計算範囲を絞り込むことができる．さらなる計算範囲の絞り込みのために，比較対象化合物Bの分割後の構造キーの前の部分の1のビット数b_1の範囲を求める．

1. $a_1 \geq b_1$かつ$a_2 \geq b_2$のとき

   式(4)より，$Tc \leq \dfrac{b_1+b_2}{a_1+a_2} = \dfrac{b}{a}$となる．

2. $a_1 \geq b_1$かつ$a_2 \leq b_2$のとき

   式(4)より，$Tc \leq \dfrac{b_1+a_2}{a_1+b_2}$となる．

   $a_2 \leq b_2$より，$Tc \leq \dfrac{b_1+a_2}{a_1+b_2} \leq \dfrac{b_1+a_2}{a_1+a_2}$

   b_1の条件式は，

$$\begin{aligned} T &\leq \frac{b_1+a_2}{a_1+a_2} = \frac{b_1+a_2}{a} \\ aT &\leq b_1+a_2 \\ aT-a_2 &\leq b_1 \end{aligned} \tag{5}$$

3. $a_1 \leq b_1$かつ$a_2 \geq b_2$のとき

   式(4)より，$Tc \leq \dfrac{a_1+b_2}{b_1+a_2}$となる．

   $a_2 \geq b_2$より，$Tc \leq \dfrac{a_1+b_2}{b_1+a_2} \leq \dfrac{a_1+a_2}{b_1+a_2}$

   b_1の条件式は，

$$\begin{aligned} T &\leq \frac{a_1+a_2}{b_1+a_2} = \frac{a}{b_1+a_2} \\ b_1+a_2 &\leq \frac{a}{T} \\ b_1 &\leq \frac{a}{T}-a_2 \end{aligned} \tag{6}$$

4. $a_1 \leq b_1$かつ$a_2 \leq b_2$のとき

   式(4)より，$Tc \leq \dfrac{a_1+a_2}{b_1+b_2} = \dfrac{a}{b}$となる．

式(5)，(6)より，b_1の条件は式(7)のようになる．

$$aT-a_2 \leq b_1 \leq \frac{a}{T}-a_2 \tag{7}$$

同様に，b_2の条件式は式(8)のようになる．

$$aT-a_1 \leq b_2 \leq \frac{a}{T}-a_1 \tag{8}$$

以上の導出過程から，$a_1 \geq b_1$，$a_2 \leq b_2$かつ$a \geq b$の場合と，$a_1 \leq b_1$，$a_2 \geq b_2$かつ$a \leq b$の場合に式(2)による従来手法よりも計算範囲を絞り込むことができる．したがって，従来手法と式(7)，(8)を組み合わせることにより，従来手法だけを用いる場合よりも計算範囲を絞り込むことができる．

4 実験と考察

4.1 実験条件

類似度比較による計算回数と計算時間の変化を評価するため，本手法を2004年版のMDDRに登録されている全ての化合物データ145,295件に対して適用した．構造キーにはMACCS Keyを用いている．MACCS Keyの生成はMOEというソフトウェアで行った．構造キーは90ビットと76ビットに分割している．閾値TとMDDRの全ての化合物データを1つずつクエリとして与え，類似度計算対象化合物をMDDRの化合物データとして，分割した76ビットに従来手法と式(8)を適用している．今回の実験では，化合物データの構造キーのビット数はあらかじめ計算している．使用した計算機は，インテルXeon3.0GHz 2MB L2キャッシュ 800MHz FSB，メモリ12GBで，C言語によりプログラミングを行っている．

閾値Tを0.8から0.95まで0.05間隔で変化させ，従来手法と本手法を適用した場合の比較回数と計算時間について比較を行った．

4.2 結果

比較回数，計算時間は図1，2のようになった．従来手法だけを使用した場合と，従来手法と式(8)を組み合わせた本手法とを比較すると，本手法を適用することで全ての閾値において比較回数が減少している．しかし，計算時間は類似度0.95の場合のみ本手法の方が4秒計算時間が短くなっている．さらに，閾値Tを0.95から0.99まで0.01間隔での計算時間を調査すると，図3のようになった．

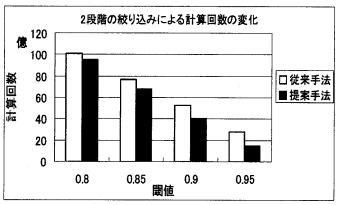

図 1: 2 段階の絞り込みによる計算回数の変化

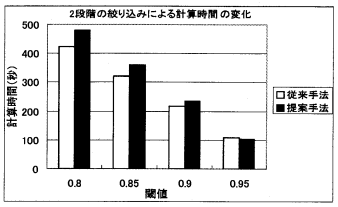

図 2: 2 段階の絞り込みによる計算時間の変化

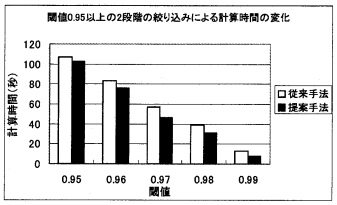

図 3: 閾値 0.91 以上の 2 段階の絞り込みによる計算時間の変化

4.3 考察

図 1 より，本手法を適用した方が式 (2) だけで絞り込みを行うよりもより計算範囲を絞り込むことができたことがわかる．しかし，計算時間は図 2，3 より，閾値が 0.95 以上の場合に計算時間が短くなっている．これは，Tanimoto 係数の計算にかかる計算時間が短いために，式 (8) の計算等にかかる時間の影響を受けているからだと考えられる．閾値 0.95 以上のときに，式 (8) の計算等の影響を上回るほど計算範囲をより絞り込めたために，計算時間がより短くなっているのだと考えられる．構造キーが 166 ビットの場合でも閾値 0.95 以上の場合に本手法を使用することでより計算範囲を絞り込むことができ，計算時間も短縮することができた．PubChem のような，より長い構造キーを利用しているデータベースでは，Tanimoto 係数の計算時間がより長くなる．しかし，本手法は構造キーのビット数に基づいているため，構造キーの長さが変わっても式 (8) の計算等の時間に影響がない．つまり，Tanimoto 係数の計算回数が多いほど，構造キーが長くなった時の影響が大きくなる．そのため，0.95 より低い閾値でも本手法の効果が期待できると考えられる．

5 おわりに

構造キーの 1 のビット数と分割後の 1 のビット数に基づき，比較対象の化合物を絞り込むことによる類似度計算の計算範囲を絞り込む手法を提案した．そして，実際の化合物データに対して適用する実験により，計算範囲の絞り込みを確認した．

謝辞

北海道大学の伊藤公人先生と中央大学の田口善弘先生には，本研究を行うきっかけとなる御助言を頂きまして心より御礼申し上げます．

本研究は，一部，科学研究費特定領域研究「基盤ゲノム」および「情報爆発」によっている．

参考文献

[1] D. L. Wheeler, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research*, Vol. 36, pp. D13–D21, 2008.

[2] MDL Drug Data Report : Internet address. http://www.mdli.com/.

[3] P. Willett. Similarity-based approaches to virtual screening. *Biochemical Society Transactions*, Vol. 31, pp. 603–606, 2003.

[4] 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.

[5] 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.