2008－BIO－12 (7)
2008／3／4

Tanimoto係数の性質に基づく化合物の類似度検索の高速化手法

清水隆史，瀬尾茂人，竹中要一，松田秀雄
大阪大学大学院情報科学研究科バイオ情報工学専攻

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．化合物のタンパク質への作用は，化合物の構造に因るところが大きいことから，化合物のデータを単純に蓄積するだけではなく，クラスタリング等により構造情報に基づいたデータの整理が必要である．その際，化合物間で類似度を計算する必要があるが，化合物数の増加に伴い，類似度計算にかかる時間が増加する．そこで本研究では，Tanimoto係数の性質を利用し，類似度計算回数を削減する手法を提案し，化合物データを用いた実験によりその有効性を示した．

A method for speedup of similarity search for compound based on the property of the Tanimoto coefficient

Takashi Shimizu, Shigeto Seno, Yoichi Takenaka and Hideo Matsuda
Department of Bioinformatic Engineering,
Graduate School of Information Science and Technology, Osaka University

Abstract

The amount of compounds in public databases goes on increasing. It is necessary not only to accumulate compound data but also to classify them such as clustering based on their structures, because their activities on proteins depend on the structures. It is necessary to calculate the similarity between compounds, but as the number of compounds increases, the calculation time increases. In this research, we propose a method for reducing the number of calculating similarity based on the property of the Tanimoto coefficient, and show the effectiveness by the experiment with compound data.

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1800万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[2]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．化合物のタンパク質への作用は化合物の構造に因るところが大きいことから，化合物のデータをデータベースに単純に蓄積するだけではなく，構造の類似性によって整理することが非常に重要である．化合物の構造を表現するのに広く使われる記述子として，特定の分子構造が分子内に存在するかどうかをビット列で表現したフィンガープリントがある．このフィンガープリントを用いてデータの整理をする際に，化合物間の類似度を計算する必要が生じる．化合物数をnとすると，化合物間の組合せの数は${}_nC_2$となり，比較回数は$O(n^2)$回である．したがって，データの増加に伴って化合物間の類似度計算回数が増加し，それに伴って類似度の計算時間が増加していく．現在登録されている化合物

数1800万個に対して，類似度計算をすべての化合物間に対して行うのは非常に困難であり，今後データ数が増えるとさらに困難になると考えられる．

また，化合物データは化合物データベースだけではなく，タンパク質データベースに対して相互参照が必要な場合がある．この問題を解決するために，オントロジーの構築によるデータの整理が行われている [3]．オントロジーとは，概念を抽出し，概念の持つ属性と概念間の関係について人間と計算機が理解できる形で記述したものであり，用途によって様々なものがある．例えば，タンパク質をコードする遺伝子の機能分類である Gene Ontology[4] や化合物の分類である Ligand Ontology[5] がある．Ligand Ontology では，タンパク質の階層構造を基にして，それらに相互作用する化合物が分類されている．

複数のデータベースにまたがった検索の例としては，ゲノム情報を用いた創薬が挙げられる．ゲノム創薬の分野では，新薬を開発にするにあたり，遺伝情報を用いて疾患の原因となる遺伝子を同定し，その遺伝子がコードするタンパク質と相互作用する化合物を探索する．このような異分野のデータベースにまたがる解析を助けるために，現在，我々は図 1 のように，Gene Ontology と Ligand Ontology との対応付けを行っている．また，商用データベースである MDDR（MDL Drug Data Report）の化合物データを使用しており，MDDR の化合物データには化合物 ID，化合物の構造情報，分子式，一般名，慣用名，薬理活性情報，文献・特許情報が付随している．この MDDR の化合物データを使用し，働きが未知の化合物データを検索 (図 1 の点線部) できるようなシステムの開発を行っている．このシステムでは，検索の際に類似度計算を行う必要があり，システムを高速化するためにも短時間で類似度を計算する必要がある．

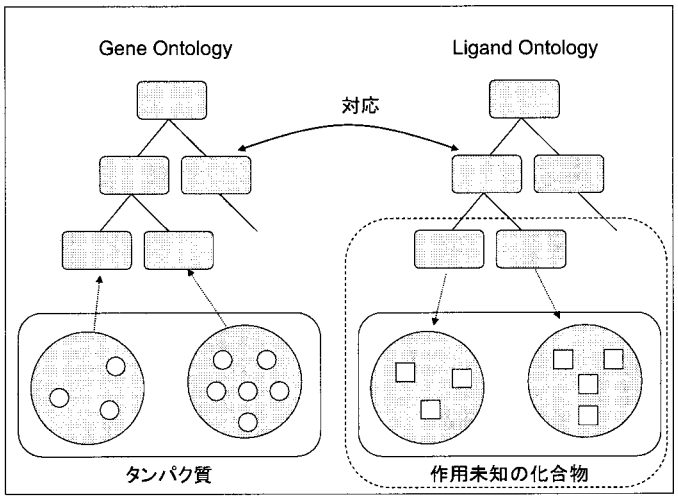

図 1: Gene Ontology と Ligand Ontology の対応付け

2 化合物の構造比較

2.1 記述子

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，部分構造の有無をビット表現したものなどの記述子 [6] 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，よくフィンガープリントが用いられる．フィンガープリントは，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，フィンガープリントのビット数や，フィンガープリントから得られる情報が変化する．代表的なフィンガープリントとして Unity 2D fingerprint[7]，Similog key[8]，MACCS Key[9] がある．

ここでは，例として MACCS Key を用いてフィンガープリントについて説明する．MACCS Key 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フィンガープリントである．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おり，各ビットを見ていくことでどのような部分構造を持っているのかがわかる．例えば，46 ビット目は臭素の有無，164 ビット目は酸素原子の有無を表している．MACCS Key の例を図 2 に示す．図 2 の例では，化合物が臭素と酸素原子を持っているので，46 番目と 164 番目のビットが 1 になっている．

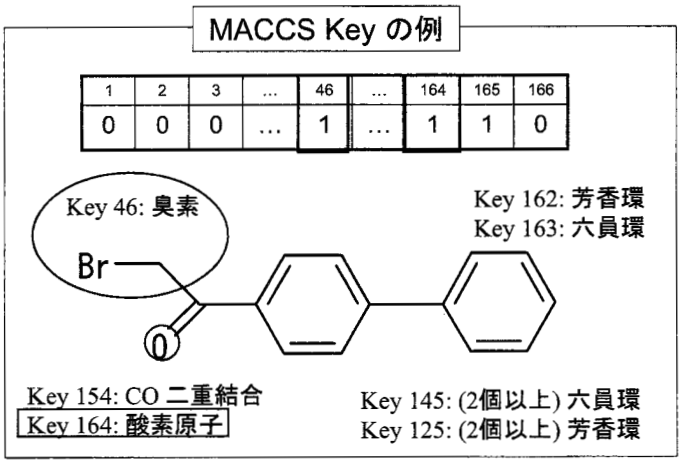

図 2: MACCS Key による化合物の表現の例

2.2 類似性評価尺度

フィンガープリント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 [10] やユークリッド距離 [6] などがある．Tanimoto 係数は，平均的なパフォーマンスがよく [11]，フィンガープリントの類似性を評価する方法としてよく用いられる．

ビット列 X，Y における Tanimoto 係数の定義式は式 (1) のようになる．

$$Tanimoto\text{ 係数 }(X,Y) = \frac{C}{A+B-C} \tag{1}$$

式 (1) の A，B，C はそれぞれ，X 中で 1 になっているビット数，Y 中で 1 になっているビット数，X と Y で共通して 1 になっているビット数を表している．

2.3 類似度計算時間

フィンガープリントを利用し，クラスタリングにより化合物の整理を行うためにはすべての化合物間で類似度を計算する必要が生じる．化合物数を n とすると，化合物間の組合せの数は ${}_nC_2$ となり，比較回数は $O(n^2)$ 回である．したがって，化合物数が増加するにつれて，化合物間の類似度比較回数が 2 乗の割合で増加する．同様に，新たに化合物データが登録された場合にも，登録された化合物データを整理するために類似度計算を行う必要があり，データの急激な増加に伴い，類似度比較の組合せが急激に増加することになる．

具体的な計算時間を示すため，MDDR の化合物データを用いた予備実験を行った．4.1 節に示す計算機を使用し，MDDR の化合物データ約 14 万 5 千個，フィンガープリント MACCS Key でそれぞれの化合物間の類似度を Tanimoto 係数により測定したところ，組み合わせ数が ${}_{145000}C_2$ 通りで約 3800 秒かかった．これを PubChem のデータ約 1800 万件に対して外挿すると，データ数が約 120 倍のため，計算時間は 120^2 倍，つまり約 600 日かかってしまう．

2.4 類似度計算時間の短縮法

フィンガープリント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．

フィンガープリントのビット数の圧縮を行う方法は，ビットを重ねることにより行う．例として，8 ビットを 4 ビットに圧縮する場合を図 3 に示す．8 ビットを 4 ビットずつに分解し，それぞれの OR を取ることにより 4 ビットに圧縮している．こうして類似度計算の際のビット計算回数を減らそうとしている．しかし，このように単純に圧縮しただけだと情報が失われてしまう．また，圧縮前のビット数を推測することで圧縮による情報の損失を最小限に抑えようという研究 [12] もある．

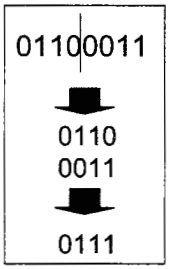

図 3: フィンガープリントの圧縮の例

2 つ目は，類似度の計算回数を減らすことで総計算時間を減らす研究である．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図っている．具体例として，フィンガープリントの 1 のビット数を利用することで計算範囲を限定する研究 [13] がある．

3 類似度検索の高速化

類似化合物の検索では，Tanimoto 係数の値 (以降，類似スコアと呼ぶ) を閾値として指定し，クエリとして与えた 1 つの化合物との類似スコアがその閾値 T 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している．しかし，閾値 T の値によってスコアを満たす化合物数は限られてくる．そこで，すべての化合物間で類似度を計算するのではなく，閾値 T 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．そこで本研究では，無駄な計算を省くことにより，化合物の高速な類似度検索を行う手法を提案する．

Tanimoto 係数の定義式 (1) における C の値は，A，B の共通ビットのため，A 及び B 以下である．つまり，比較する化合物のフィンガープリントの 1 のビット数によって類似度の最大値が変化する．類似度検索の高速化は，閾値 T の値とクエリとした与えた化合物の 1 のビット数 A に基づき，T を満たすためのもう一方の化合物の 1 のビット数 B の範囲を決定することにより行う．

$A > B$ のとき，T の最大値は式 (1) より条件式 $T \leq B/A$ を満たす．これを式変形することで比較対象となる B の範囲は $AT \leq B$ となる．

$A < B$ のとき，類似スコア T の最大値は条件式 $T \leq A/B$ を満たす．これを式変形することで比較対象となる B の範囲は $B \leq A/T$ となる．

以上より，比較対象 B の範囲は式 (2) のように絞り込むことができる．

$$AT \leq B \leq \frac{A}{T} \tag{2}$$

仮に，フィンガープリントのビット数が 10 の場合について説明を行う．クエリとして与えた化合物の 1 のビット数 A が 5 の場合を仮定し，もう一方の化合物の 1 のビット数 B が 3 や 7 の場合の類似スコアの最大値を図 4 に示す．図 4 のように，B が 3 や 7 の場合，類似スコアの最大値はそれぞれ 0.6 や 0.714 となる．図 5 は，A が 5 の場合の B と類似スコアの最大値の関係を表しており，横軸は B の 1 のビット数を，縦軸は Tanimoto 係数の値を表している．図 5 は，1 のビット数によって類似スコアの最大値が決まることを示している．このとき，閾値 T を 0.8 とすると，B が 4，5，6 の場合のみ 0.8 以上になる可能性がある．つまり，それ以外の場合は計算の必要がないことがわかる．

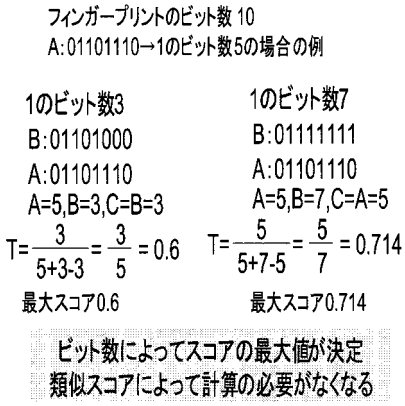

図 4: Tanimoto 係数の最大値の例

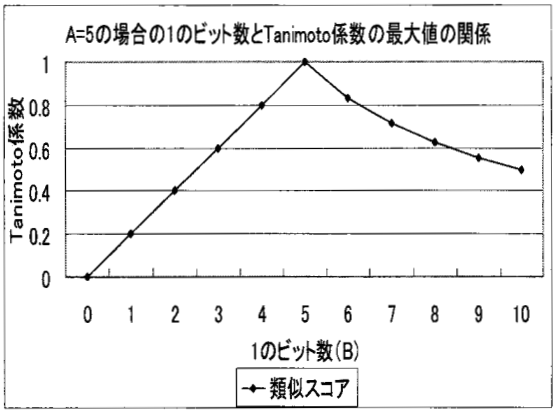

図 5: Tanimoto 係数の最大値の変化

1 つの化合物をクエリとして与え，データベース内の化合物の中から類似スコアが閾値 T 以上の化合物を探索する際の高速化手法の流れについて説明する．まず，クエリとして与えられた化合物のフィンガープリントのビット数 A を計算する．次に，A と指定された閾値 T から式 (2) に基づいて，比較対象の化合物のフィンガープリントのビット数 B の範囲 R を計算する．そして，データベース内のデータを 1 つずつ取り出し，取り出した化合物のビット数が範囲 R を満たすかどうかを判定し，満たすなら類似スコアを計算し，満たさなければ次の化合物の判定を行う．これをデータベース内のすべての化合物に対して行う．

この高速化手法では，範囲 R を満たさない化合物についても範囲を満たすかどうかの判定を行うことになるが，あらかじめデータベース内のデータを，フィンガープリントの 1 のビット数に基

づいてソートしておくことで，範囲内の化合物のみ計算を行うことができ，さらに高速化を行うことができる．

4 実験と考察

4.1 実験条件

類似度比較回数による時間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．フィンガープリントには MACCS Key を用いている．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クエリ以外の MDDR の化合物データとして本手法を適用している．今回の実験では，化合物データのフィンガープリント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 (R)Xeon(R)3.0GHz 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 0.6 から 0.95 まで 0.05 間隔で変化させ，本手法を適用した場合の化合物間の類似度比較回数と計算時間を測定する実験を行った．

そして，本手法を使用しない場合と，閾値 T を変化させ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

実験結果は図 6 のグラフである．横軸が閾値となる類似スコアで，縦軸は左が比較回数，右が計算時間を表している．また，一番左の閾値なしというのは本手法を適用しない場合の結果である．

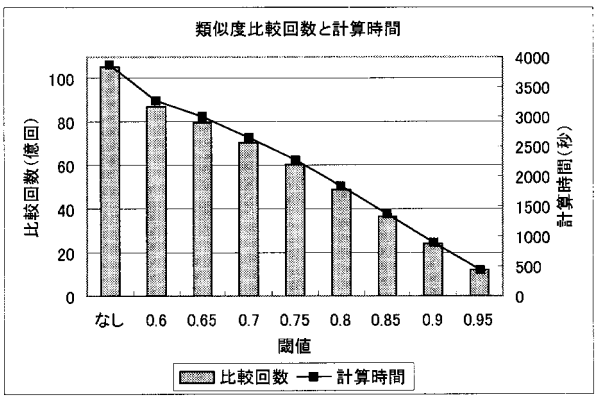

図 6: 類似度比較回数と計算時間

本手法を使用しなかった場合と使用した場合を比較すると，手法を適用することにより比較回数・計算時間共に減少している．また，各閾値間で比較すると，閾値を上げていくにつれて，比較回数・計算時間共に減少している．特に，本手法を使用しなかった場合と類似スコア T を 0.95 にした場合で比べると，比較回数・計算時間共に約 1/10 にまで減少していることがわかった．

4.3 考察

MDDR の化合物データを用いた実験では 4.2 節のような結果になったが，MDDR は薬理活性が既知のデータベースであり，データに偏りがある可能性がある．そのため，公共データベースである PubChem に対しても本手法が有効であるとは限らない．本手法の高速化の効果はフィンガープリントの 1 のビット数の分布に依存する．そこで，実験で用いた MDDR の化合物データと，PubChem の化合物データ約 1000 万件について，MACCS Key の 1 のビット数の分布を調査している．それぞれの 1 のビット数の分布は図 7，図 8 のようになる．横軸が 1 のビット数で，縦軸が化合物数である．

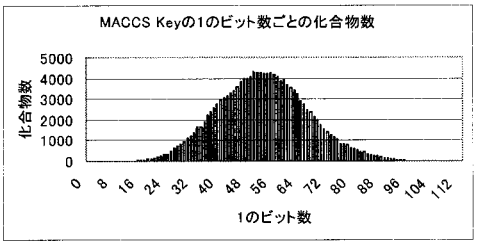

図 7: MDDR の 1 のビット数の分布

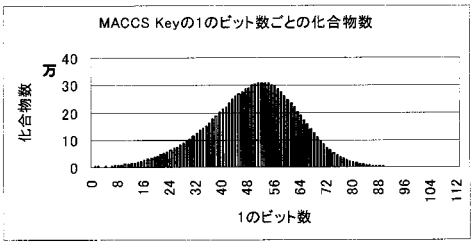

図 8: PubChem の 1 のビット数の分布

2 つの図より，どちらも同様の分布をしていることがわかる．フィンガープリントの 1 のビット数の分布が図 7 のような形になるデータであれば同様の結果が得られると考えらるので，MDDR のデータに対して適用した結果と同様の結果が PubChem のデータに対しても得られると推測できる．この結果を踏まえて，PubChem のデータ約 1800 万件に適用すると，閾値 T が 0.95 の場合 600 日から約 2ヶ月に短縮できると考えられる．

5 おわりに

フィンガープリントの 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検索の高速化を行う手法を提案し，実際に化合物データに対して適用する実験により高速化を確認した．

今回提案した手法を，現在開発を行っている検索システム (図 1) に利用する予定である．4.3 節に，閾値 T を 0.95 にした場合，2ヶ月に短縮できると述べたが，本手法を適用しただけではまだ不十分だと考えられる．今後の改善点としては，情報が失われないフィンガープリントのビットの圧縮法の考案のような，計算時間短縮のための他の手法との組み合わせることでさらに高速化が行える可能性が考えられる．

謝辞

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, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, Vol. 36, pp. D13–D21, 2008.

[2] C. M. Dobson. Chemical space and biology. *Nature*, Vol. 432, pp. 824–828, 2004.

[3] 高井貴子, 高木利久. 生命科学のためのオントロジー. 情報知識学会第 9 回研究報告会講演論文集, pp. 13–18, 2001.

[4] The Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Research*, Vol. 36, pp. D440–D444, 2008.

[5] A. Schuffenhauer, J. Zimmermann, R. Stoop, J. J. van der Vyver, S. Lecchini, and E. Jacoby. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *Journal of Information and Computer Sciences*, Vol. 42, pp. 947–955, 2002.

[6] J. Gasteiger, and T. Engel, 監訳 : 船津公人, 訳 : 船津公人, 佐藤寛子, 増井秀行. ケモインフォマティクス. 丸善株式会社, 2005.

[7] J. W. Raymonda and P. Willettb. 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.

[8] A. Schuffenhauer, P. Floersheim, P. Acklin, and E. Jacoby. Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of Chemical Information and Computer Sciences*, Vol. 43, pp. 391–405, 2003.

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

[10] J. W. Godden, L. Xue, and J. Bajorath. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. *Journal of Chemical Information and Computer Sciences*, Vol. 40, No. 1, pp. 163–166, 2000.

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

[12] S. Swamidass and P. Baldi. Mathematical correction for fingerprint similarity measures to improve chemical retrieval. *Journal of Chemical Information and Modeling*, Vol. 47, pp. 952–964, 2007.

[13] S. Swamidass and P. Baldi. 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.