様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25 年　5 月　30 日現在

機関番号：82626
研究種目：若手研究(A)
研究期間：2009～2012
課題番号：21680025
研究課題名（和文）　In-silico 創薬のための機械学習を用いた生理活性配座予測

研究課題名（英文）　In-silico prediction of active comformations by machine learning

研究代表者
津田 宏治　（Tsuda Koji）
独立行政法人産業技術総合研究所・生命情報工学研究センター・主任研究員
研究者番号：90357517

研究成果の概要（和文）：
プロテインデータバンク(PDB)に登録されているタンパク質三次元構造の数が７万を超えるなど、利用できる三次元情報は、増加を続けている。これらのデータから、機能情報を抽出するため、SketchSort という全ペア類似度検索アルゴリズムを開発し、120 万個におよぶリガンド結合サイトに適用した。また、その結果を Possum というデータベースにまとめ、一般に公開した。

研究成果の概要（英文）：
The amount of protein 3D structure information is ever increasing as the number of registered proteins in Protein Data Bank (PDB) exceeded 70,000. To extract functional information, we developed a fast all pairs similarity search algorithm called sketchsort and applied it to 1.2 million ligand binding sites. The found pairs are summarized in a database called Possum and provided for public use.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	6,500,000	1,950,000	8,450,000
2010 年度	3,500,000	1,050,000	4,550,000
2011 年度	3,500,000	1,050,000	4,550,000
2012 年度	1,800,000	540,000	2,340,000
年度			
総　計	15,300,000	4,590,000	19,890,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学Ａ
キーワード：機械学習、生体分子、生理活性

１．研究開始当初の背景

プロテインデータバンク(PDB)に登録されているタンパク質三次元構造の数が７万を超え、３万を超えるヒトのタンパク質に関して三次元構造が全て予測されるなど、利用できる三次元情報は、増加を続けている。これらの大域的な類似性を発見するのは、BLAST などのソフトで簡単にできるが、局所的な類似性を発見するのは、計算量的に容易でない。しかし、タンパク質の三次元構造データベースから、機能的、進化的な知識を獲得するには、やはり局所的な類似性を発見する必要がある。

タンパク質三次元構造の中で、最も注目すべきなのは、他の分子（リガンド）と結合する部位（リガンド結合サイト）である（図１）。リガンド結合サイトの全体全比較の研究はこれまでも行われてきたが、計算量に難があ

った。例えば、2008 年の Minai et al.による研究では、48347 個のリガンド結合サイトの全体全比較を行うのに、グリッドコンピューティングを用いて 29 日要している。これは、CPU 1 個を用いると約 2 年かかる計算になる。従って、研究開始当初には、大規模に、全体全比較を行った研究は無かった。

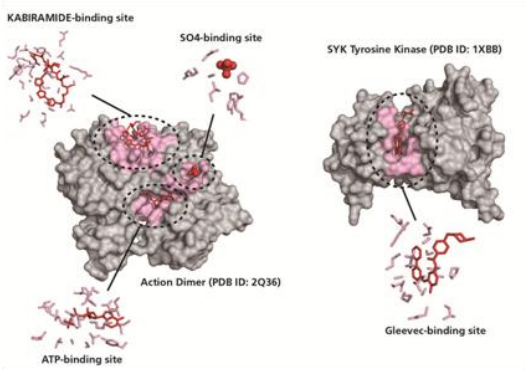

図1：リガンド結合サイトの例

２．研究の目的

本研究では、機械学習の知見を生かして、超高速な全ペア類似度検索アルゴリズムを開発する。また、現在知られているすべてのリガンド結合サイトに加え、ghecom というプログラムによって予測された結合サイト候補を加えて、できるだけ大きなデータベースを作成し、それに当該アルゴリズムを適用することによって、これまでに指摘されてこなかった類似サイトを発見し、In-silico 創薬に役立てる。また、計算結果は、ウエブサイトを通して公開し、興味を持った研究者がいつでもアクセスできるようにする。

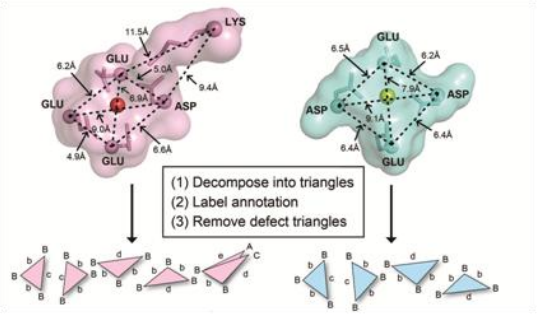

図2：幾何的特徴量の抽出

３．研究の方法

まず、PDB に含まれている 61927 個のタンパク質三次元構造から、226438 個の既知リガンド結合サイト、及び、991900 個のリガンド結合サイト候補を抽出した。双方を合わせると約 120 万個となる。次に、各々のリガンド結合サイトに対して、幾何的な特徴抽出を行った。この方法では、リガンド結合サイト周辺にあるアミノ酸を一点とする三角形を全て抽出し、その三角形を、辺の長さなど、幾何的な特徴と、アミノ酸の種類によって集計する（図２）。

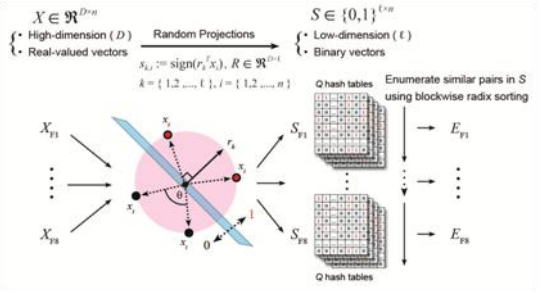

図3：SketchSortアルゴリズムの概要

本研究で開発した高速全ペア類似度検索アルゴリズムは、SketchSort と呼ばれる。このアルゴリズムでは、まず特徴ベクトルを Locality Sensitive Hashing と呼ばれるランダム射影によって、0/1 のビットベクトルに射影する。このビットベクトルのハミング距離（ミスマッチの数）は、元の特徴ベクトルのコサイン距離に関する単調増加関数になっているため、ハミング距離が近いビットベクトルを列挙できれば、コサイン距離が近い特徴ベクトルのペアを発見できることになる（図３）。

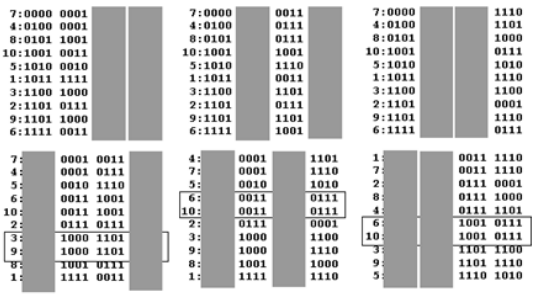

図4：複合ソート法

ビットベクトルの中から類似したものを発見するために、複合ソート法を用いる（図4）。この方法では、ハミング距離 d 以内のペアを全列挙したい場合には、ビットベクトルを d +k 個のブロックに分割し、その内 d 個をマスクして、ソーティングを行う。マスクの仕方は、全ての場合を試すので、複数回ソーティングは行われる。これにより、ハミング距離が d 以内のペアは、必ず一回はソーティング結果の中に隣接して現れるため、簡単に発見することができるようになる。しかし、ハミング距離が d ＋１以上のものも発見される場合があるので、隣接して現れたペアに関しては、全体のハミング距離を計算して、誤りを除いておく。

SketchSort では、ランダム射影を用いるため、コサイン距離が閾値以内のペアを 100%完全に列挙することはできず、必ず見逃しが生じる可能性がある。しかし、この確率(Missing edge ratio)は理論的に計算でき、非常に低い確率(10 の-6 乗など)に抑えることができるので、応用上は問題にならない。

４．研究成果

SketchSort を 120 万個のリガンド結合サイトに適用し、コサイン距離が 0.15 以下のペアを全列挙したところ、約 8800 万個の類似ペアを得た。その計算は、通常のＰＣを用いて、僅か４時間ほどで終了した。発見されたペアの中で、CATH code によって双方がアノテーションを持つペアを選ぶと約 1600 万ペアあった。さらに、構造アラインメントツール TM-align を用いてフィルタリングしたところ、約 347 万個の確信度の高い類似サイトペアを得ることができた。

これらのペアは、構造の類似性のみから選ばれており、アミノ酸配列の類似性は考慮されていないため、配列は似ていないのに、構造は類似しているアナロガスなペアを多く発見することができた。例えば、EF-hand containing protein (PDB ID: 1XO5)と、transferase (PDB ID: 1NUD)との間に発見されたペアを見てみると、二つのサイトは、双方とも、カルシウムイオンの結合サイトであるが、一方は、典型的な EF-hand motif であり、他方は、ベータストランド間のループに存在するなど、全く周辺の条件が違っている。

また、これまで機能未知であったタンパク質と機能既知のタンパク質の間にも類似ペアが発見された。例えば、ThiS-ThiF protein complex (PDB ID: 1ZUD)と、DR_0571 protein (PDB ID: 3C4N)の間に見つかったペアや、ribosomal protein S6 kinase (PDB ID: 2WNT)と Serine/Thereonine-Protein kinase (PDB ID: 2WTK)の間に見つかったペアなどが、それに当てはまる。このような類似性は、タンパク質の機能推定において有力な情報となる。

発見された膨大な類似サイトペアの中で、我々が目を通すことができたのは、ほんの一部に過ぎず、他のタンパク質研究者にとって興味深い例が多数存在すると思われる。そこで、我々は、発見したペアをデータベース化し、Possum というウエブサイトで検索できるようにした。このデータベースは、リガンド結合サイトに関するものでは現在世界最大である。例えば、SiteBase という他のデータベースは、約 33000 個のリガンド結合サイトの情報しか提供していない。

Possum は幾つかの検索機能を用意している。SearchK では、ある既知のリガンド結合サイトを指定すると、それに類似したサイトのリストを返すことができる。また、SearchP では、ある構造既知のタンパク質を指定すると、それに属するリガンド結合サイトの全てに関して、類似サイトのリストを返す。

本研究では、タンパク質構造解析という、あまりアルゴリズムの高速性が重視されていなかった分野において、高速性がもたらす価値を主張するという新しいアプローチを取ったことが、生命医薬情報連合大会で、ベストポスター賞を獲得するなどの成功につながったと思われる。同様のアプローチは、新世代シクエンサーデータの解析にも適用され、SlideSort というソフトウエアの形で発表されている。今後は、同様のアプローチを、メタボロミクス、プロテオミクスなどの分野のデータに適用することで、新たな価値を創造していきたい。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 9 件）

①J. Ito, Y. Tabei, K. Shimizu, K. Tsuda and K. Tomii, “PoSSuM: a Database of Similar Protein-Ligand Binding and Putative Pockets”, Nucleic Acids Research, 査読有, 40, D541-D548, 2012, DOI:10.1093/nar/gkr1130

②J. Ito, Y. Tabei, K. Shimizu, K. Tomii and K. Tsuda, “PDB-scale Analysis of Known and Putative Ligand-binding Sites with Structural Sketches”, Proteins, 査読有, 80, 747-763, 2012, DOI: 10.1002/prot.23232

③K. Shimizu and K. Tsuda, “SlideSort: All Pairs Similarity Search for Short Reads”, Bioinformatics, 査読有, 27, 464-470, 2011, DOI:10.1093/bioinformatics/btq677

④Y. Tabei and K. Tsuda, “SketchSort: Fast All Pairs Similarity Search for Large Databases of Molecular Fingerprints”, Molecular Informatics, 査読有, 30, 801-807, 2011, DOI: 10.1002/minf.201100050

〔学会発表〕（計 1 件）

K. Tomii, J. Ito, Y. Tabei, K. Shimizu, K. Tsuda, “Possum: A Database for Predicting Protein-Ligand Interactions”, 生命医薬情報学連合大会, 2012/10/14, 東京都、タワーホール船堀

〔その他〕

ホームページ等

http://possum.cbrc.jp/PoSSuM/

６．研究組織

(1)研究代表者

津田宏治（Tsuda Koji）
独立行政法人産業技術総合研究所・生命情報工学研究センター・主任研究員
研究者番号：90357517