様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成　25　年　6　月　24　日現在

機関番号：１７１０４
研究種目：若手研究（B）
研究期間：２０１１　　～　２０１２
課題番号：２３７００３３８
研究課題名（和文）　大量のタンパク質リガンドデータより相互作用の構造的特徴をマイニングする方法の開発
研究課題名（英文）　Development of a method to extract structural features from large-scale protein-ligand data
研究代表者
西郷 浩人　　(Hiroto Saigo)
九州工業大学・大学院情報工学研究院・准教授
研究者番号：90586124

研究成果の概要（和文）：

本研究が行ったのはタンパク質リガンド相互作用の解明において鍵となる、結合中の化合物の部分構造とタンパク質のドメインの対を探索する方法の開発であり、以下の特徴をもつ。1)大量のタンパク質と大量のリガンドを同時に解析できる。2)立体構造が得られない場合にも利用出来る。3)既存のモチーフや商用の記述子のライブラリーを必要とせず、構造データのマイニング法により新規の相互作用モチーフも発見できる。

研究成果の概要（英文）：

We developed a method that searches for protein-ligand pairs which play key roles in protein-ligand interactions. Our method has the following desirable properties; 1) it can analyze large-scale protein-ligand database, 2) it is available in situations where protein tertiary structures are unavailable, 3) it does not depend on commercial descriptors for representing chemical compounds, but can generate them on demand.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,400,000	720,000	3,120,000

研究分野：情報学
科研費の分科・細目：情報学基礎・統計科学
キーワード：医薬生物・ゲノム統計解析

１．研究開始当初の背景

近年、次世代シークエンサーなどの測量機器の発達にともない大量のタンパク質配列データ、そして低分子化合物のデータが蓄積されている。これらのデータはタンパク質リガンド相互作用予測に関わるため、創薬においてその重要性は古くから認識されており、DrugBank[1]や Glida[2]などのデータベースも公開されている。しかしながら、タンパク質配列とリガンドのような種類の異なる大量のデータを同時に扱うことのできる手法の開発は十分とはいえない。

タンパク質と低分子化合物の結合予測シミュレーションには、従来ドッキングが使われてきたが、以下のような欠点が知られている:(1)基本的に１つのタンパク質（レセプター）に対して低分子化合物（リガンド）のライブラリーを１つ１つテストする方法であり、ハイスループットな方法ではない、(2)ターゲットのタンパク質と個々の低分子化合物との結合の強さはスコアであらわされるが、このスコアは必ずしも大域的最適解である保障はない、(3)タンパク質の詳細な立体構造が必要。
これらの弱点を補う方法の開発も行われて

いる。例えば Fukunishi らは、手元にあるデータベース中の全てのタンパク質と低分子化合物のドッキングを網羅的に行ってタンパク質・化合物相互作用行列を作成してから主成分分析を行うことによって、タンパク質の立体構造が得られない場合でも低分子化合物の探索が出来るとしているが、この方法は通常のドッキングよりもはるかに時間のかかる方法であり、大量の CPU を必要とする[3]。

一方、Yamanishi らや Jakob らは、カーネル法を用いた方法を提案している[5]。これらの方法はタンパク質の立体構造は必要としないが、タンパク質のどの部分と化合物のどの部分が結合するのかといった詳細を知ることができないという欠点がある[2]。Nagamine らも、サポートベクターマシンを用いたバーチャルスクリーニング法を提案しているが、同様の問題がある[4]。

----------参考文献----------

[1]DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–906, (2006).
[2]Okuno Y, et al., GLIDA: GPCR-ligand database for chemical genomics drug discovery and tools update. Nucleic Acids Res 36: D907–912, (2008).
[3]Fukunishi Y and Nakamura H Improvement of protein-compound docking scores by using amino-acid sequence similarities of proteins. Journal of chemical information and modeling: 48(1) 148-156, (2008)
[4]Nagamine N et al. Integrating Statistical Predictions and Experimental Verifications for Enhancing Protein-Chemical Interaction Predictions in Virtual Screening. PLoS Comput Biol. 5(6), (2009).

２．研究の目的

【代表者のこれまでの取り組みとの関係】

代表者は、本研究に繋がる取り組みとして、「グラフで表された化合物からの教師有り、教師無し特徴抽出」を研究してきた（KDD2008、ICDM2008、Machine Learning 2009 においてそれぞれ筆頭著者として発表）。この方法は近年開発が進む頻出グラフマイニングに拠っている[6,7]。本提案研究は、グラフで表された化合物に加えて、文字列で表されたタンパク質配列も入力とし、両者を同時にマイニングするというものである。異なる種類の構造データの同時マイニングは、代表者らの知る限り未開の領域であり、世界でもパイオニア的研究になる可能性がある。尚、異なる種類の構造データを組み合わせることによりタンパク質リガンド相互作用の予測精度を高めることが出来ることは、[5]において既に報告されているが、この手法では特徴選択が出来ないため、創薬に向けた相互作用部位の更なる解析は困難である。

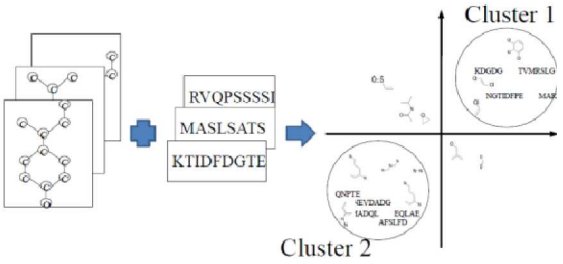

図 1:本研究計画が提案する手法による解析例。
相互作用するタンパク質と低分子化合物の特徴同士がクラスターを形成する。

【研究の独創性】

本計画書で提案する方法は、大量のタンパク質と大量の低分子化合物を同時に扱うことができ、タンパク質のどの部分と化合物のどの部分が結合するのかといった情報は部分グラフや部分文字列として得られる（図 1）。また、タンパク質の立体構造が得られない状況でも精度の高いタンパク質・化合物結合予測を目指しているが、立体構造情報が得られる場合について特徴ベクトルに追加することは容易である。

【研究期間内の達成目標】

(1)タンパク質配列の部分配列とリガンドの部分構造を用いたタンパク質リガンド相互作用予測法の開発とその精度の評価
(2)(1)により得られたタンパク質部分配列とリガンド部分構造の妥当性の検証
(3)成果の国内、国際学会における発表とプログラムソースコードの公開

【予想される結果と意義】

提案研究は、異なる種類のデータを同時にマイニングでき、かつ新しい特徴を発見できるという意味で、独創的である。また、これまでに開発された apriori ベースのマイニング手法はどれでも組み合わせることが出来るので、汎用性や他の分野への応用性も高い。

----------参考文献----------

[5] Yamanishi Y et al. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Biofinformatics 24(13), i232-i240, (2008)

３．研究の方法

【基本的アイデア】
本計画書が提案する研究では、タンパク質配列の特徴量であるモチーフをその有無、低分子化合物の特徴量である部分構造の有無をビットベクトルで表す(図 3)。

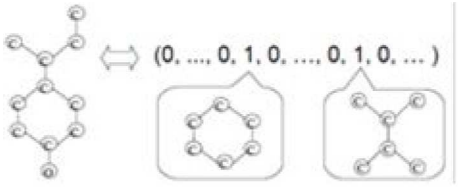

図 2:低分子化合物の特徴量のビットベクトル表現

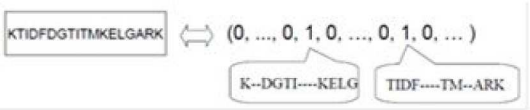

図 3:タンパク質配列中の特徴量のビットベクトル表現

ビットベクトルの生成には頻出パターンマイニングを用い、含まれる特徴を列挙することにより行う。頻出パターンマイニングは、グラフや配列といった構造をもつデータに含まれる全ての部分構造を列挙する方法である[6,7]。図 4 にグラフデータ中の共通部分構造を探索する様子を示す。グラフデータ中の共通部分構造の数は、比較的小さなグラフデータベースにおいても膨大な数にのぼるが、学習に必要な特徴量の数はそれよりもはるかに少ないことが分かっている。また、頻出パターンマイニングにおける逆単調性(anti-monotonicity)を利用した枝狩りにより、計算量、及び使用する記憶領域を更に節約することが可能である。

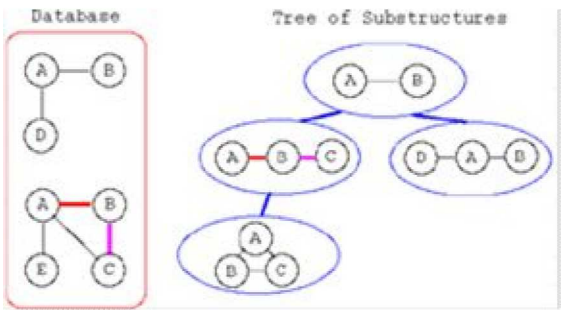

図 4:低分子化合物中の特徴量の木構造を用いた探索

研究計画・方法（つづき）
【平成 23 年度の計画】
代表者は先の研究(KDD2008 で発表)にて、各グラフデータに与えられたラベルとグラフの部分構造の間の共分散を最大化するようなグラフマイニング法を提案した。一方、本提案研究ではグラフで表された化合物と文字列で表されたタンパク質の相関(correlation)を最大化するような部分グラフと部分文字列の対をマイニングする必要があるが、これは容易ではない。なぜなら、相関(correlation)は共分散(covariance)を正規化した値であり、正規化のためには膨大な数の部分構造を予め列挙しておく必要があり、困難であるからである。よって、まずは与えられたグラフと文字列の間の共分散を最大化するようなマイニングを行う。正規化する方法については、専門家であるフランス、キュリー研究所の Yamanishi 博士と協力して開発する予定である。Yamanishi 博士はカーネル正準相関解析の論文を国際学会及び国際誌上において発表しており、適切な助言をいただけると考えている。もし正規化が難しい時は、計算機実験にて正規化しない場合との経験的な差異を調べることとする。平成 23 年度後半までにはこの問題を解決する予定である。

【平成 24 年度の計画】
代表者は先の研究(ICDM2008 で発表)にて、各グラフデータの部分構造を全て列挙することなく、主成分分析を行う手法を提案した。この方法は Lanczos 法が計画行列(design matrix)に直接アクセスすることなく固有値分解ができるという性質に基づいているが、２つの計画行列が共に直接アクセスできない時は Lanczos 法が利用できないことが知られている[8]。即ち、部分グラフ特徴量と部分文字列特徴量を同時に生成しようとする本提案研究では Lanczos 法のエラー解析結果が使えないことになる[9]。そこで、平成 24 年の前半には以下の３点を行い、比較する:
1)Witten らの発見的手法[10]　2)１つの計画行列を予め計算しておいた後、もう１つの計画行列を Lanczos 法で求める　3)２つの計画行列を予め計算しておく。
この比較は平成 24 年度初めまでには行い、国際学会に発表する論文を用意するものとする。

【提案研究の評価方法について】
タンパク質リガンド相互作用予測の精度については、DrugBank に登録されている既知の相互作用で確かめることとする。結合箇所の詳細については、同様の研究をしている Yamanishi 博士と議論して検証する。また、共同研究者である Perret 博士はスイス、ノヴァルティス社で実際のバーチャルスクリーニングの手法を使用し改良する立場にあ

るので、本提案研究の実用性について現場からの意見を頂く予定である。

【その他の技術的に困難な可能性がある点】
本計画におけるもう１つの疑問点は、タンパク質配列の特徴量としてギャップなしの文字列を使うかギャップありの Hidden Markov Model (HMM)などのモチーフを使うかである。多数の文字列データからギャップ無しの部分共通文字列を求めることは suffix tree アルゴリズムなどで線形時間で求められる一方、HMM モチーフを求めるためには EM アルゴリズムの繰り返し計算を行わなければならない。後者は精度が高く、前者は精度が低いので、時間と精度はトレードオフである。現在利用できる文字列マイニングアルゴリズムは前者であるが、後者の効率的な実装についても検討したい。

---------- 参 考 文 献 ----------
[6]Inokuchi et al. Applying the apriori-based mining method to mutagenesis data analysis, Journal of Computer Aided Chemistry, 2, 87-92, (2001).
[7]Yan X. and Han J. gSpan:graph-based substructure pattern mining, ICDM, 721-724, (2002).
[8]Parlett B. N. The symmetric eigenvalue problem, In Applied Mathematics, Society for Industrial Mathematics, (1980).
[9] Saad Y. On the rates of convergence of the Lanczos and the block Lanczos methods. SIAM J. Num. Anal., 5, 687-706, (1980).
[10]Witten et al. A penalized matrix decomposition, with applications to sparse principal components and can onical correlation analysis. Biostatistics, 10(3), 515-534, (2009).

４．研究成果
１）正準相関解析に疎性制約を導入することにより、大量の化合物の特徴量とリガンドの特徴量から特徴選択を行えることを示した。
２）ターゲットの QSAR モデルの双対問題に疎性を導入することによって化合物の部分構造マイニングを高速化できることを示した。

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２件）
①
Saigo, H., Kashima, H., Tsuda, K., Fast iterative mining using sparsity-inducing loss functions, IEICE Transaction on Information and Systems, 査読有, E96-D(8), 2013
②
Yoshihiro Yamanishi, Edouard Pauwels, Hiroto Saigo, Veronique Stoven, Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions, Journal of Chemical Information and Modeling, 査読有,51(5), 2011, 1183-1194

〔学会発表〕（計４件）
①
Saigo, H. Learning from treatment history to predict response to anti-HIV therapy, BMIRC International Symposium on Frontiers in Computational Systems Biology and Bioengineering, March 1st, 2013, Iizuka, Japan
②
Suryanto, C.H., Saigo, H., Fukui, K. Protein Clustering on Grassmans Manifold Pattern Recognition in Bioinformatics, November 10th, 2012, Tokyo, Japan
③
Ikeda, N., Saigo, H. An algorithm for searching multiple SNP interactions Joint Conference on Informatics in Biology, Medicine and Pharmacology, October 14th, 2012, Funakoshi, Japan
④
Ikeda, N., Saigo, H. An algorithm for searching multiple SNP interactions Special Interesting Group on Bioinformatics, 2012年08月09日, Iizuka, Japan

〔図書〕（計１件）
①
Hiroto Saigo, Koji Tsuda, IGI Global "Matrix Decomposition-based Dimensionality Reduction on Graph Data" in Sakr, S. and Pardede, E. editors "Graph Data Management: Techniques and Applications", 2011, 260-284

〔その他〕
ホームページ等
http://www.bio.kyutech.ac.jp/~saigo/

６．研究組織
(1)研究代表者
西郷 浩人 (Hiroto Saigo)
九州工業大学・大学院情報工学研究院・准教授
研究者番号：90586124