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２４年　４月　１日現在

機関番号：１２３０１
研究種目：若手研究（Ｂ）
研究期間：２０１０～２０１１
課題番号：２２７００３１２
研究課題名（和文）リガンド予測のための記述子自動獲得アルゴリズムの開発
研究課題名（英文）　Development of automatic descriptor extraction algorithms for ligand prediction.
研究代表者
加藤　毅（KATO　TSUYOSHI）
群馬大学・大学院工学研究科・准教授
研究者番号：４０４０１２３６

研究成果の概要（和文）：

本研究は，指定された化合物がターゲットタンパク質のリガンドとなりうるかを予測するリガンド予測問題に関するものである．タンパク質と相互作用するリガンドを予測するには効果的な記述子が重要となる．大域的モデルでは，各々の局所問題に特化した記述子を扱うことができないが，局所的モデルでは少サンプル問題が起因して十分な汎化能力が得られないことが多々ある．本研究は，転移学習の考え方を導入して，この２つの問題を同時に解決する記述子自動獲得アルゴリズムを開発した．

研究成果の概要（英文）：

This study tackled the problem for ligand prediction. The problem is to predict whether specified chemical compounds are interacted with specified target proteins. To achieve accurate prediction of ligands interacting with target proteins, effective descriptors that describe the properties of chemical compounds are crucial. Global models cannot deal with descriptors specific to each local prediction problem. Local models suffer from small sample size problems. This study introduced a concept of transfer learning to develop automatic descriptor extraction algorithms that solve the above two problems simultaneously.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	1,100,000	330,000	1,430,000
２０１１年度	800,000	240,000	1,040,000
年度			
年度			
年度			
総　計	1,900,000	570,000	2,470,000

研究分野：バイオインフォマティクス
科研費の分科・細目：情報学・生体生命情報学
キーワード：非線形合成記述子，サポートベクトルマシン，転移学習，重みつき経験分布，リガンド予測，一般化固有値問題，P450 アイソザイム，嗅覚受容体

１．研究開始当初の背景
新薬発見に際してバーチャルスクリーニングは不可欠である．ヒトの遺伝子２万～２万５千個の中で，約３千個がターゲットになると考えられており，そのうち薬理工学上で調査されているのは 800 個程度である．そのうち半数程度は GPCR である．一方，化学空間に１千万程度の非冗長な化学構造があるが，そのうち薬剤として承認されたものは 1,000 程度である．このようにターゲット空間とリガンド空間は広大であるが，そのうちほんの一部だけが使われているのが現状である．この広大な２つの空間の積空間には薬剤になりうる組み合わせがまだ多く眠っていると考えられ，これを計算技術によって発掘することが計算化学における主要な課題の一つになっている．
リガンド予測とは，リガンド-ターゲットの行列を完成させる作業に他ならない．この表を埋める方法には，局所的モデルによる方法と大域的モデルによる方法の２つがある．局所的モデルによる方法とは各列ごとに独立に欠損要素を予測するものである．基本的には，その列の中ですでに相互作用するとわかっているリガンドとの類似度が近いリガンドを見つけ，そのリガンドも相互作用すると予測するやりかたである．類似度を計算するために記述子と呼ばれる数値列でリガンドの特徴を表す．その数値列を比較することでどれほど似ているか算出する．従来は，タニモト係数と呼ばれる２値ベクトル間の類似度を測る指標を使うために，数値列を２値のビット列に限定していたが，近年はベイジアンモデリングを応用した方法など通常の統計手法で予測することが多くなったため２値ベクトルに限る必要はなくなった．サポートベクトルマシンなどカーネル法と呼ばれる統計手法を用いる場合は，ベクトルの概念をさらに拡張したカーネルと呼ばれる数量で予測することもできる．局所的モデルはその列の中で，相互作用することが既知のリガンドが最低一つは必要とする．まして，統計手法を用いる場合，十分な量の既知のリガンドがなければ精度よく予測することはできない．
大域的モデルによる方法は，リガンド-ターゲットの行列全体を一つのモデルで表し，行列の穴埋めを同時に予測するものである．そのために，リガンドの記述子とターゲットの記述子を合成して，行列内の一つ一つの要素の記述子を構成する．合成した記述子を一つのサンプルとみて統計手法を適用して予測を行うのである．大域的モデルの長所は，相互作用するリガンドを一つももたないターゲットに対するリガンドも予測できることである．大域的モデルの短所は２つある．短所の一つは，モデルがいわば大ざっぱになってしまうことである．すべての列を同時に扱うため，原理的には似た列の情報が学習に取り込まれ，局所的モデルにおける少サンプルの問題が解決されるはずである．しかし，無関係な列の情報も同時に取り込まれてしまいこれがかえって悪影響を与えてしまう．もう一つの短所は，局所問題ごとに記述子を選ぶことができないことである．局所的モデルの場合は，原理的にはターゲットごとに記述子を再設計することが可能である．しかし，大域的モデルの場合，行列全体で一つのモデルを作るために，どのターゲットに対しても共通の記述子を使わなくてはならない．

２．研究の目的
本研究は，高精度なリガンド予測を実現するために記述子を自動獲得することを目的とする．ターゲットと相互作用するリガンドを予測するには効果的な記述子が重要となる．大域的モデルでは，各々のターゲットに特化した記述子を扱うことができないが，局所的モデルでは少サンプル問題が起因して十分な汎化能力が得られないことが多々ある．本研究は，転移学習の考え方を導入して，この２つの問題を同時に解決するアルゴリズムを開発する．

３．研究の方法
本研究ではリガンド予測のために記述子を学習によって獲得するアルゴリズムを開発する．これまでにリガンドの記述子として様々な記述子（１次元記述子，２次元記述子，３次元記述子）がある．このうち，本研究では２次元記述子の可能性に注目する．２次元記述子とは，化学構造式を離散数学でいうところのグラフとみて，グラフに含まれる部分構造をカウントすることによって得られる数量である．このような情報を実際に記述子として表すと膨大な長さのベクトルになってしまうが，サポートベクトルマシンなど多くの統計手法は記述子間の内積(カーネルと呼ばれる)を計算するだけで動作するようにできている．この性質を利用して記述子を陽に表さずに直接内積を計算する効率的なアルゴリズムが存在している．これらに共通し

て言えることは，すべての部分グラフに対して，その頻度を部分グラフの大きさによって系統的に重みづけして計算されるようになっている．しかし，実際にはすべての部分グラフを含めてしまうと余分な情報が入りすぎてしまい，これが学習機械の予測性能を下げる原因となっている．本研究では，２次元記述子生成のためのアプローチとしては，高頻度部分グラフを列挙してから記述子を構成するやり方を採用した．従来の方法は記述子が疎になりすぎてうまく学習できない場合があった．これに対して本研究では，部分グラフ特有の性質として次元間の相関情報の有用性を見出した．開発するアルゴリズムはターゲットごとに最適な記述子を自動的に獲得すると同時にリガンドの予測を高精度化するものである．リガンド予測と記述子獲得は相互に依存している．すなわち，予測精度が悪ければ効果的な記述子は得られず，記述子が悪ければ予測精度も悪くなる．本研究では，局所的モデルにおける少サンプル問題を解決するため転移学習のアルゴリズムを開発した．それぞれのターゲットに対する選択性を予測することを一つのタスクとみると，ターゲット間の類似性によって図１のようなタスク間ネットワークを作ることができる．タスク間ネットワークでは各ノードがタスクでエッジはノード間が関係を持っていることを示している．ターゲット蛋白質のアミノ酸配列を使ってタスク間ネットワークを構築した．転移学習アルゴリズムは高い汎化能力を有するサポートベクトルマシンという予測モデルを拡張して用いた．予測モデルは記述子と同じ長さを持つベクトルとバイアスをパラメータとして持っており，そのベクトルと記述子との内積にバイアス項を加えることで予測に用いるスコアを得ることができる．学習によって，正例はスコアが＋１以上になるように，負例はスコアが－１以下になるようにモデルパラメータの値が決定される．本研究で開発したアルゴリズムでは，タスク間ネットワークによって既定されているタスク間にモデルパラメータ間の類似性を持たせるトリックを用い，類似タスクと同時に学習を行うようにした．そのアルゴリズムは，非線形に統合した記述子からも学習できるように，カーネル化可能になっている．

局所的記述子を得るために，記述子の非線形合成を行った．非線形合成を行うために，化学的空間と相互作用空間を導入した（図２参照）．各々のデータの化学的プロファイルと相互作用プロファイルはそれぞれの空間にある．これらのプロファイルを合成空間に線形射影する．このとき，射影したときの期待ユークリッド距離が最小になるような射影は正準相関分析と呼ばれる古典的な統計手法を利用することで求めることができる．本研究では，局所的記述子を得るために，プロファイル間の類似性から導出される重みつき経験分布を導入して，その経験分布に基づく期待ユークリッド距離が最小になるような射影を計算するようにした．このアルゴリズムはカーネル化できるため，最適な非線形合成を求めることができる．さらに重みつき経験分布に適合するような学習を行う拡張サポートベクトルマシンアルゴリズムを開発した．

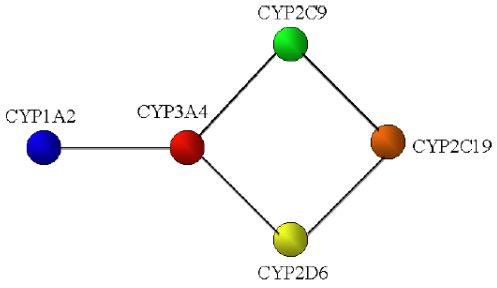

図１．タスク間ネットワーク

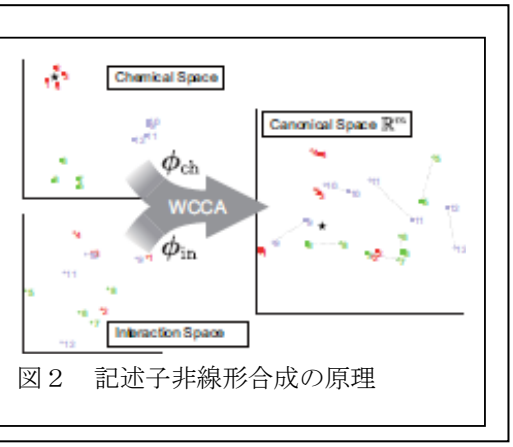

図２　記述子非線形合成の原理

４．研究成果

まず，P450 アイソザイムの選択性を予測する問題を使って，転移学習の予測性能を評価した．具体的には，CYP1A2，CYP2C9，CYP3A4，CYP2D6，CYP2C19 の５個のアイソザイムを用いた．これらのターゲット蛋白質に対して，網羅的スクリーニングを行ったデータセットを使った．このデータセットは 1991 個の基質との相互作用の有無を含んでいる．これをベンチマークとして用いたところ，ROC カーブに基づく AUC は平均 0.865 を得た．通常の SVM の AUC は，0.808 であり，顕著な性能向上が確認できた．そのほかに，Bagging，Weighted kNN，Adaboost，LDA，CART，ナイーブベイズ，Latent Semantic Index といった機械学習アルゴリズムとも比較実験を行い，本研究で開発した手法が統計的有意に性能が上回っていることを確認した．

さらに，非線形合成記述子の性能評価も行った．図３は嗅覚受容体データを用いた時の予測性能を示している．WW は，非線形合成記述子と拡張サポートベクトルマシンアルゴリズムとの組み合わせ，WU は非線形合成記述子と従来の SVM との組み合わせ，KW は通常のユークリッド距離による非線形合成と拡張サポートベクトルマシンアルゴリズムとの組み合わせ，KU は従来の SVM との組み合わせ，S はそのまま従来の SVM を用いた場合，SGL は線形カーネルによる大域的モデル，SGR は RBF カーネルによる大域的モデルを示す．ROC カーブによる AUC による評価でも，F-measure による評価でも提案手法が最高予測性能を示した．

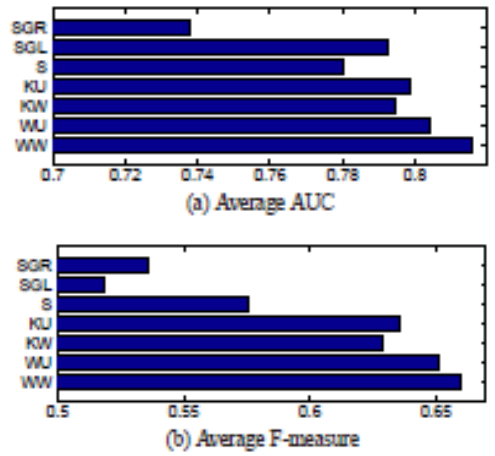

図３．非線形合成記述子の性能評価

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　1 件）
①Reiji Teramoto and Tsuyoshi Kato, Transfer learning for cytochrome P450 isozyme selectivity prediction, Journal of Bioinformatics and Computational Biology, Vol. 9, No. 4, pp.521--540, 2011.

〔学会発表〕（計　0 件）

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等

現在，非線形合成記述子の成果に関して論文投稿中である．掲載が決まり次第，公開する．

６．研究組織
(1)研究代表者
加藤　毅（KATO　TSUYOSHI）
群馬大学・大学院工学研究科・准教授
研究者番号：４０４０１２３６

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：