# タンパク質化合物相互作用の網羅的予測手法とデータベースの開発

慶應義塾大学理工学部
榊原　康文

## Development of statistical prediction method for protein-chemical interactions

Yasubumi Sakakibara
Department of Biosciences and Informatics, Keio University

Predictions of interactions between target proteins and potential lead compounds are of great benefit in the drug discovery process. We present a comprehensively applicable statistical prediction method for interactions between any proteins and chemical compounds, which requires only protein sequence data and chemical structure data and utilizes the statistical learning method of support vector machines. We show the usefulness of our approach in predicting potential ligands binding to human androgen receptors from more than 19 million chemical compounds and verifying these predictions by in vitro binding. Moreover, we utilize this experimental validation as feedback to enhance subsequent computational predictions, and experimentally validate these predictions again.

## 1. はじめに

創薬の初期ステップであるリード化合物の探索において、計算機によるタンパク質化合物間相互作用予測は有用な手法であり、本課題ではより汎用性が高く、入手が容易であるアミノ酸配列データ、化合物構造データ及びマススペクトルデータを用いたタンパク質化合物相互作用予測手法の開発を行う。膨大なデータの中から注目する特性を識別して説明する特徴や規則を発見し、未知のデータに対して意義のある予測を行う統計学的学習手法の一つであるSVMを適用して、網羅的な相互作用予測を行う。本課題を通じて、インシリコ予測がいかに創薬におけるリード化合物スクリーニングに貢献し得るかを示す。

さらに、網羅的インシリコ予測、ウエット実験検証、ウエット検証実験結果のインシリコ予測へのフィードバックというプロセスを繰り返すフィードバック戦略を提案して、その手法をがんに関わる特定のタンパク質の網羅的結合リガンド予測に適用する。

この研究課題開発においては、予測対象が膨大となる網羅的適用を考慮して、偽陽性予測数の削減を重要課題の一つとする。また、ウエット実験検証を実行することを考慮して、実験結果のフィードバック戦略を提案する。開発手法の統計的予測の利点であるフィードバックを用いた再学習を利用することで，従来あまり行われてこなかったインシリコ予測とウエット実験検証の効率的統合の有効性を示すことを目標とする。

さらに、相互作用予測システムを実際の問題、がんに関わる特定のタンパク質の一つとしてアン

ドロゲンレセプターの網羅的結合リガンド予測に適用し、本システムの有効性を実証する。

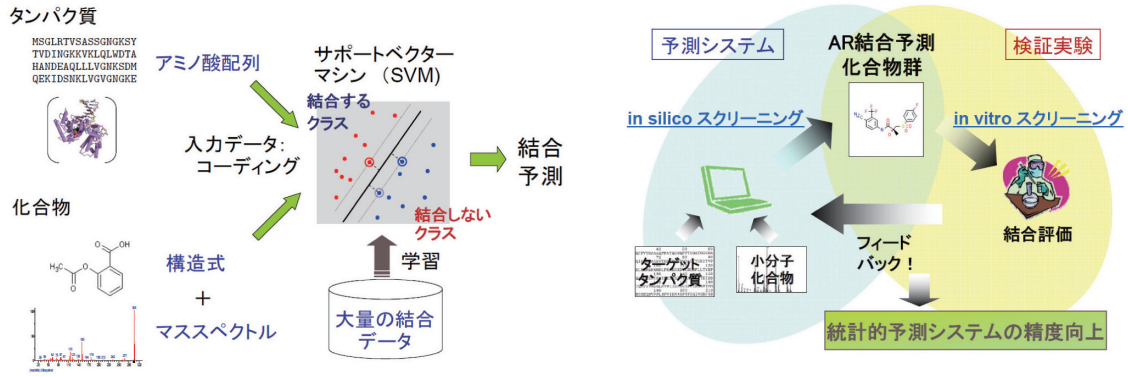



図１　統計的相互作用予測システム（左）とウエット検証実験のフィードバック戦略（右）

## 2. 研究開発の成果

### 2.1 タンパク質化合物相互作用の網羅的予測手法の開発

入力されたタンパク質化合物ペアについて、まず、タンパク質を、アミノ酸の物理化学的特性、および、アミノ酸文字列の出現頻度に基づいて数値化した。 一方、化合物は、化合物構造式を用いる場合にはパスの出現頻度、マススペクトルデータを用いる場合にはピークの位置と強度をもとに数値化を行った。数値化、ベクトル化されたタンパク質化合物ペアに対して、SVMモデルを適用し、与えられたタンパク質化合物ペアが結合するか否かを判定した。SVMは一度モデルが作成できれば、以降はあらゆるタンパク質化合物ペアの相互作用の有無を迅速に判定可能であり、本課題の目的とする網羅的、汎用的利用の実現に適している。その結果、964種類のFDA認可薬物とレセプター、酵素、イオンチャネルなどを含むターゲットタンパク質456種類を用いて、作成された汎用の結合予測モデルは約85%の正解率を達成することができた[1]（**図２**）。

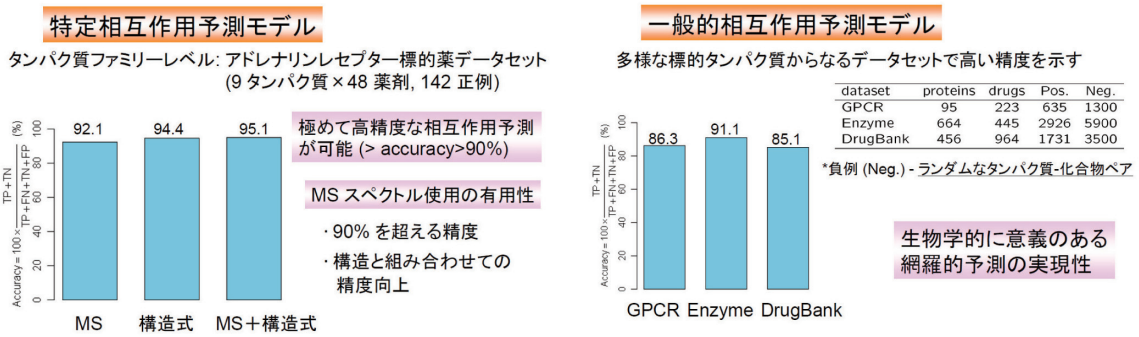



図２　タンパク質ファミリー特定の予測精度と（左）と汎用の相互作用予測の精度（右）

統計的学習手法を利用することの利点の一つに、新たに得られた実験結果のフィードバックを行えることが挙げられる。本課題においてもウエット検証実験結果をいかに新たな予測に反映させるかが重要な課題となる。生物学的実験結果のフィードバックとして以下の３つの手法を比較検討した[2]（**図３（左）参照**）：

①　１層SVM及び２層SVMの学習データの正例・負例に新たなフィードバックデータを追

加する。

②（１に加えて）新たなフィードバックデータから付加モデルを作成し、このモデルの出力を２層SVMで利用する。

③（２に加えて）付加モデルの出力に重みをつけて２層SVMで利用する。

PDSP Ki データベース (http://pdsp.med.unc.edu/pdsp.php) などから、アンドロゲンレセプター結合化合物35個の情報を収集した。この情報を付加的情報として、フィードバック手法の比較を行った。**図３（右）**が示すように、付加モデルを一定値以上の重みとともに利用することで、偽陽性率を低く保ちつつ、予測範囲の拡大が可能なことが示された[2]。

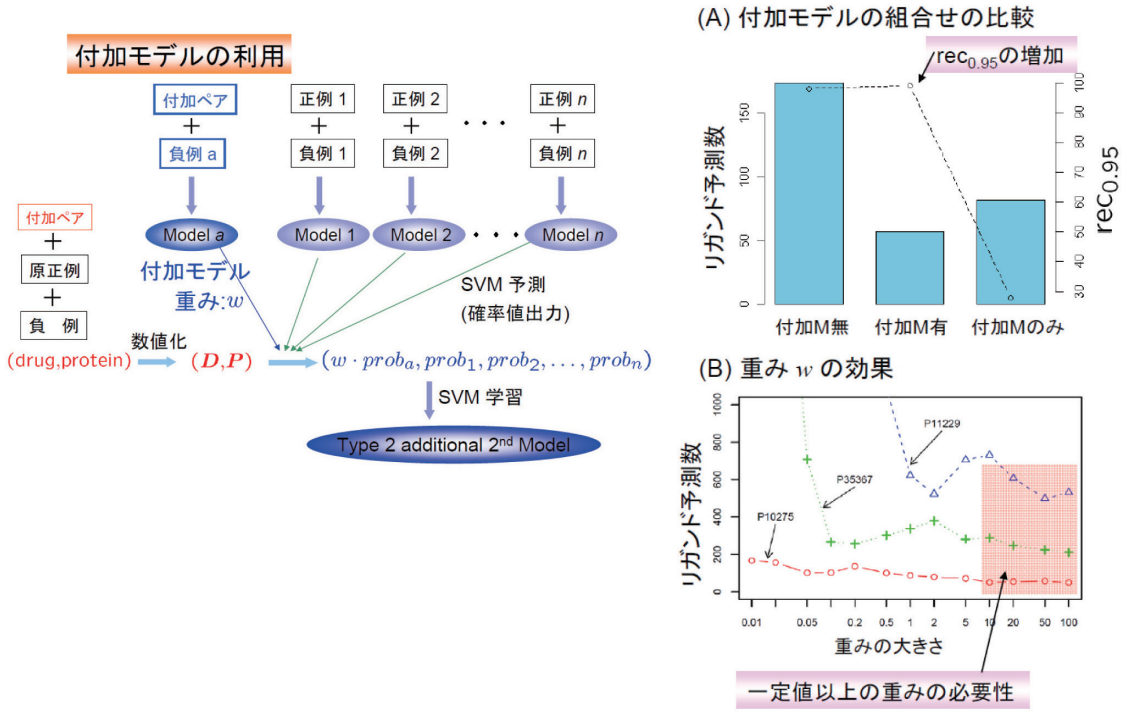



**図３　付加モデルを利用する２層SVMモデル（左）と付加データ利用法の効果（右）**

## 2.2 がん治療薬リード化合物網羅的探索への応用―インシリコ予測とウエット検証の統合

アンドロゲンレセプターとPubChemデータベース中の約1,900万個の全化合物に対する予測の結果、464個の化合物を結合リガンドと予測した[2]。次に、464予測結合化合物及び結合すると予測されなかった化合物の中から、購入可能な17個の化合物について，ケミカルバイオロジーの手法を用いて結合予測の検証を行った。レセプターバインディングアッセイによる結合評価系が構築できたので、相互作用予測システムによりARに結合すると予測された11化合物（positive）と、結合しないと予測された６化合物（negative）について結合評価を行った。その結果、MR300,000までにIC50 値が求められた化合物は15化合物、MR 300,000まで50％阻害に到達しなかった化合物は３化合物であった。また、IC50値をARと小分子化合物の結合の強さの指標としたとき、３化合物を除いてはIC50値 MR 60,000を境として、positiveとnegativeに分かれることが分かった。以上の結果から、相互作用予測システムは高精度に予測できたといえる。よって予測が誤った３化合物の結果を修正して相互作用予測システムに再学習させることで予測システムのさらなる精度の向上が図れると考えられる。

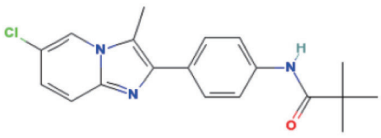
本研究課題が提案する実験的検証結果のフィードバックを用いた再学習による第２次インシリコ予測を行った。ここで、第１次実験的検証結果をフィードバックすると同時に、ステロイド骨格を有する化合物を負例として設定するアンタゴニスト志向フィードバックを適用して、「付加モデルを利用する２層SVMモデル」を用いて予測を行った。その理由は、単純な結合・非結合の識別が目的ではなくアンタゴニストであるリード化合物の探索が目的であるため、ステロイドホルモン受容体であるアンドロゲンレセプターに対してステロイド骨格を有する化合物はアゴニストの可能性が高いこと、構造の異なるアンタゴニストが薬剤耐性克服につながる可能性があること等の生物学的知見を反映した。

第２次インシリコ予測によって得られた化合物のうち５つの購入可能な化合物について，再びケミカルバイオロジーの手法を用いて結合予測の検証を行った。５個の予測結合化合物のうち３化合物が実際に結合することが確認された（**図４（A）**）.ここで，非常に興味深い発見は、アンタゴニスト志向フィードバックを伴う第２次インシリコ予測は，フィードバックのねらい通り，ステロイド骨格とは構造が大きく異なるアンドロゲンレセプター結合リガンドT5853872（**図４（C）**）を発見した[2]。このT5853872は、E-Dragon（既存ソフトウェア）による数値化及び主成分分析の適用で構成されたケミカルスペース上で，ステロイド骨格を持つ結合リガンド群（**図４（B）**の上側の点線内の領域）と、アンタゴニスト薬剤として知られるflutamide骨格を持つ結合リガンド群（**図４（B）**の下側の点線内の領域）のいずれからも離れた位置に存在する。

(A) 実験結果

| compound | pred. | $IC_{50}$ ($\mu$M) |
|---|---|---|
| DSHS00507SC | Pos.(79%) | 22 |
| 4J-584S | Pos.(54%) | 22 |
| T5853872 | Pos.(84%) | 80 |
| BAS 01279920 | Pos.(79%) | > 200 |
| AN-562/43163258 | Pos.(53%) | > 200 |

(C) T5853872 の構造




(B) ケミカルスペース上の分布

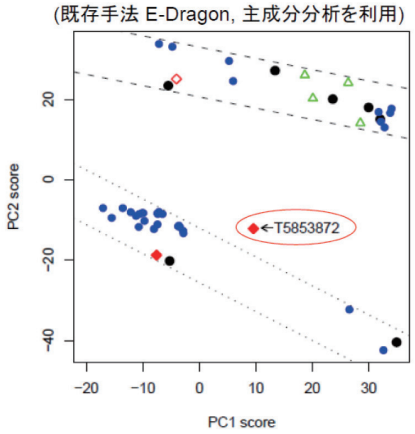



**図４　予測結果の上位化合物（左）とその結合評価実験の結果（右）**

### 2.3 相互作用予測システムCOPICATの実装

これまで、立体構造未知のタンパク質を対象としたリード化合物の探索を可能とするため、SVMを用いた統計的学習に基づくタンパク質化合物間相互作用予測のアルゴリズムを開発し、計算機予測実験を実施してきた。それをベースに、Web上で予測モデルの構築とそれを用いた相互作用予測を行うことのできる「COPICAT」システムの開発を行い、ウェブサイト上で公開した（http://copicat.dna.bio.keio.ac.jp/）。

## 3. まとめ

タンパク質化合物相互作用予測システムCOPICATを開発した。統計的相互作用予測システムとウエット検証実験系を統合するフィードバック戦略手法を提案して、アンドロゲンアンタゴニスト

において実践した。前立腺がん関連タンパク質であるアンドロゲンのアンタゴニスト用新規リード化合物T5853872を発見して、その阻害活性と細胞増殖抑制効果を確認した。網羅的インシリコ予測、ウエット実験検証、ウエット検証実験結果のインシリコ予測へのフィードバックという本課題が提案するプロセスをアンドロゲンレセプター結合リガンド予測に適用し、80%以上の高い精度（**図５**と**図６**）、フィードバックによりケミカルスペースの効率的な探索、構造や物理化学特性の異なる新規リガンドの発見が可能なことを示した.

## 4. 研究開発実施体制

代表研究者　榊原　康文（慶應義塾大学理工学部）

研究開発題目

(1) タンパク質化合物の相互作用の網羅的予測手法の開発
　　グループリーダー　榊原　康文（慶應義塾大学理工学部）

(2) ケミカルバイオロジーの手法を用いた結合予測の検証と実験データの生成
　　グループリーダー　井本　正哉（慶應義塾大学理工学部）

(3) テキストマイニングを用いた相互作用予測の検証とデータ構築
　　グループリーダー　櫻井　彰人（慶應義塾大学理工学部）

## 5. 参考文献


[1] N. Nagamine and Y. Sakakibara. Statistical prediction of protein?chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics*, 23(15), 2004-2012, 2007.

[2] N. Nagamine, T. Shirakawa, Y. Minato, K. Torii, H. Kobayashi, M. Imoto, Y. Sakakibara. Integrating Statistical Predictions and Experimental Verifications for Enhancing Protein-Chemical Interaction Predictions in Virtual Screening. *PLoS Computational Biology*, 5(6), e1000397, 2009.

[3] K. Sasaki, N. Nagamine, and Y. Sakakibara. Support vector machine prediction of N- and O-glycosylation sites using whole sequence information and subcellular localization. *IPSJ Transactions on Bioinformatics*, 2, 25-35, 2009.