情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

化合物-タンパク質活性空間における特徴選択

Feature selection in chemical-protein binding activity space

新島　聡*　　奥野　恭史†
Satoshi Niijima　　Yasushi Okuno

Abstract: In this paper, we address the issue of feature selection for chemical genomics. In particular, we propose an efficient feature selection algorithm for identifying chemical features that contribute to prediction of binding activity between chemicals and proteins. Notably, this algorithm allows feature selection in binding activity space, into which chemicals are mapped jointly with proteins by means of kernel methods. We apply the algorithm to a dataset on Cytochrome P450 (CYP), illustrating its capability of selecting a small subset of predictive features, which are also found to be indicative of CYP inhibitors. Although this study is directed toward the selection of chemical features within the context of chemical genomics, the proposed algorithm has the potential to find wide applications in real-world problems.

Keywords: Kernel methods, feature selection, regression, chemical genomics

1　まえがき

大量の遺伝子・タンパク質情報を膨大な数の化合物と関連付け，化合物-タンパク質相互作用を包括的に明らかにする「ケミカルゲノミクス」において，近年，機械学習を用いて，相互作用をゲノムワイドに予測するインシリコ創薬技術の開発が盛んになっている [1, 15].

化合物とタンパク質の相互作用の有無を予測する問題は，2クラスの識別問題として，またタンパク質に対する化合物の活性を予測する問題は，回帰問題として扱えるため，様々な識別・回帰手法が適用可能である．その中でも，カーネル法は，非線形な予測を効率的におこなえることや，化合物やタンパク質の非ベクトル（構造）表現を許容できることに加え，化合物とタンパク質のペアを効率的に合成できることから，相互作用や活性の予測において重要な役割を果たしている [9]．しかし，これまでの研究は，予測そのものに焦点をあてたものが多く，その一方で，ケミカルゲノミクスデータから，相互作用や活性に関与する特徴を抽出するための数理的手法については，研究がなされていないのが現状である．

*京都大学大学院薬学研究科, 606-8501 京都市左京区吉田下阿達町 46-29, tel. 075-753-4559, e-mail niijima@pharm.kyoto-u.ac.jp, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501
†京都大学大学院薬学研究科, 606-8501 京都市左京区吉田下阿達町 46-29, tel. 075-753-4559, e-mail okuno@pharm.kyoto-u.ac.jp, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501

このような背景から，著者らは，相互作用予測に寄与する化合物属性を同定するために，化合物-タンパク質の相互作用空間において特徴（属性）選択をおこなう手法を，先に提案している (unpublished work)．その手法は，後述の Hilbert-Schmidt Independence Criterion (HSIC) [6] を応用したものであり，カーネル関数を介して構成される相互作用空間において特徴選択をおこなう点が，従来にない特徴である．

同様に，回帰問題として扱える活性予測においても，HSIC を応用できるが，計算量の大きさが実用上の問題となる．そこで本研究では，化合物-タンパク質の活性空間における特徴選択の効率的アルゴリズムを提案し，活性予測に寄与する化合物属性の同定に適用する．本論文では，薬物代謝において重要な役割を果たす酵素，シトクロム P450 の阻害活性データに対する属性選択の結果を示す．

2　相互作用（活性）空間の表現

相互作用予測は，相互作用の有無が判明している化合物-タンパク質のペアを学習サンプルとして識別器を構築し，相互作用未知のペアについてその有無を予測する問題として設定できる．活性予測も同様に，活性値が判明しているペアを学習サンプルとしてリグレッサーを構築し，活性未知のペアについて活性値を予測する問題と

して設定できる．

ここで，クラスラベル（相互作用の有無や活性値）はペアに対して与えられるため，化合物，タンパク質それぞれの特徴ベクトル（explicit に与えられなくてもよい）を統合して，相互作用（活性）空間を表す特徴ベクトル（同じく explicit に与えられなくてもよい）を合成しなければならない．すなわち，化合物 c の特徴ベクトルを $\boldsymbol{\Phi}(c)$，タンパク質 p の特徴ベクトルを $\boldsymbol{\Psi}(p)$ と表すとき，それらからペア (c,p) の特徴ベクトル $\boldsymbol{\Pi}(c,p)$ をどのように合成するかが最初の問題となる．

化合物とタンパク質のデータ表現は，ケモインフォマティクスやバイオインフォマティクスの分野において様々なものが用いられているが，一般に表現方法が異なるため，それらをうまく統合できるような枠組みが望ましい．それを可能にする有効な手段の一つがカーネル法である．

本研究では，特に有効性が知られているテンソル積カーネルを用いた合成方法に焦点をあて，それを介して相互作用（活性）空間を構成することにする．具体的には，ペアの特徴ベクトルを以下の式で定義する．

$$\boldsymbol{\Pi}(c,p) = \boldsymbol{\Phi}(c) \otimes \boldsymbol{\Psi}(p) \tag{1}$$

ここで，$\otimes$ はテンソル積を表す．このとき，$\boldsymbol{\Pi}(c,p)$ の要素は $\boldsymbol{\Phi}(c)$ と $\boldsymbol{\Psi}(p)$ の各要素の積となる．たとえば，$\boldsymbol{\Phi}(c), \boldsymbol{\Psi}(p)$ をそれぞれ d_c, d_p 次元の特徴ベクトルとして explicit に与えた場合，ペアは $d_c \times d_p$ 次元のベクトルとなる．したがって，化合物とタンパク質の ”交差 ” を十分にとらえることができると考えられるが，一方では，計算量が現実的ではなくなる．

しかし，カーネル法を用いて予測をおこなう場合，いわゆるカーネルトリックによって効率的な計算が可能となる．実際に，ペア同士の内積は

$$\begin{aligned}\boldsymbol{\Pi}(c,p)^\top \boldsymbol{\Pi}(c',p') &= \big(\boldsymbol{\Phi}(c) \otimes \boldsymbol{\Psi}(p)\big)^\top \big(\boldsymbol{\Phi}(c') \otimes \boldsymbol{\Psi}(p')\big) \\ &= \boldsymbol{\Phi}(c)^\top \boldsymbol{\Phi}(c') \times \boldsymbol{\Psi}(p)^\top \boldsymbol{\Psi}(p')\end{aligned}$$

と計算できるため，化合物，タンパク質のカーネルをそれぞれ

$$k_{chem}(c,c') \equiv \boldsymbol{\Phi}(c)^\top \boldsymbol{\Phi}(c'), \tag{2}$$

$$k_{prot}(p,p') \equiv \boldsymbol{\Psi}(p)^\top \boldsymbol{\Psi}(p') \tag{3}$$

とすると，化合物-タンパク質ペアのカーネルは，

$$\begin{aligned}k((c,p),(c',p')) &\equiv \boldsymbol{\Pi}(c,p)^\top \boldsymbol{\Pi}(c',p') \\ &= k_{chem}(c,c') \times k_{prot}(p,p')\end{aligned}$$

となる．つまり，(1) を直接計算する必要はなく，ペア同士の類似度は，ペアを構成する化合物，タンパク質それぞれの類似度の積として計算できることが分かる．そして，化合物，タンパク質の特徴ベクトルは explicit に与える必要がなく，配列，グラフなどの非ベクトル表現を許容できるため，相互作用（活性）空間の多様な表現が可能となる．

3　相互作用空間における特徴選択

先に提案した，相互作用予測のための特徴選択法は，化合物のみを属性ベクトルで表現し，それをタンパク質とカップリングした相互作用空間において，化合物属性の選択をおこなうものである．先述のテンソル積カーネルを介して空間を構成する場合は，kernel-induced feature space（以降，カーネル空間とよぶ）において特徴選択をおこなう必要がある．特徴選択は，機械学習，パターン認識において極めて重要な技術であり，特徴の評価基準や探索アルゴリズムは様々なものが考案されているが，カーネル空間においてそれを可能にする方法が研究されるようになったのは最近のことである（**5節**）．著者らは，特に計算の効率性を考慮し，Hilbert-Schmidt Independence Criterion (HSIC) を評価基準として用いている．本節では，HSIC および，それに基づく後方探索アルゴリズム BAHSIC について概略を示す.（詳細は [6, 18] を参照されたい.）

3.1　HSIC

簡潔には，HSIC は特徴（属性）とクラスラベルの独立性を測る基準であり，値が 0 に近いほど独立性が高いことを意味する．x_i をサンプル，y_i をそのクラスラベルとし，n 個の学習サンプル $(x_1,y_1),\ldots,(x_n,y_n)$ が与えられているとする．このとき，HSIC の経験推定量は，$x_i, y_i\ (i=1,\ldots,n)$ に対するカーネル行列 $\boldsymbol{K}, \boldsymbol{L} \in \mathbb{R}^{n\times n}$ を用いて以下の式で与えられる．

$$\mathrm{HSIC} = \frac{1}{(n-1)^2}\mathrm{Tr}\,(\boldsymbol{K}\boldsymbol{L}) \tag{4}$$

ここで，Tr は行列のトレースを表す．ただし，$\boldsymbol{K}, \boldsymbol{L}$ は中心化されているものとする．HSIC は収束性などにおいて，良い性質を持つことが知られる．

化合物-タンパク質相互作用の予測においては，$x_i = (c_i, p_i)$ が相互作用するペアであれば $y_i = 1$, そうでなければ $y_i = -1$ であり，$L_{ij} = y_i y_j$ で与えられる．一方，$\boldsymbol{K}$ は化合物-タンパク質ペア間の類似度を表し，テンソル積カーネルを用いる場合，

$$\boldsymbol{K} = \boldsymbol{K}_{chem} \circ \boldsymbol{K}_{prot} \tag{5}$$

と表される．ここで，$\circ$ はアダマール積を表す．すなわち，$\boldsymbol{K}_{chem}, \boldsymbol{K}_{prot} \in \mathbb{R}^{n\times n}$ の各要素は，(2), (3) で計算

され，$\boldsymbol{K}$ の要素はそれらの積となる．したがって，(4), (5) から分かるように，化合物とタンパク質のカーネルさえ計算できれば，HSIC の計算は容易である．

3.2 BAHSIC アルゴリズム

相互作用予測に有用な属性とは，クラスラベルとの依存性が高い（独立性が低い）ものである．したがって，与えられた属性集合の中から，(4) を最大化するような属性部分集合をどのように探索するかが次の問題となる．HSIC を評価基準とする探索には，原理的には様々なアルゴリズムを利用できるが，ここでは後方探索アルゴリズム BAHSIC を用いる．簡潔には，各属性に関して leave-one-out 方式で HSIC を計算し，HSIC の値が高く（依存性が高く）なるように属性を除外していき，少数の属性にまで絞り込む．

本研究では，相互作用に寄与する化合物属性を同定することに関心があるので，化合物は explicit にベクトル表現するが，タンパク質は非ベクトル表現でもかまわない．相互作用予測のための属性選択アルゴリズムは以下のとおりである．

入力：化合物の属性集合 $\mathcal{S}$ および
$\big((c_1,p_1),y_1\big),\ldots,\big((c_n,p_n),y_n\big)$
出力：属性の順序リスト $\mathcal{L}$

1: $L \leftarrow \emptyset$
Repeat 2-4 until $\mathcal{S}=\emptyset$
2: $\mathcal{I} \leftarrow \arg\max_{\mathcal{I}} \sum_{i\in\mathcal{I}\subset\mathcal{S}}$ HSIC $(\mathcal{S}\setminus\{i\})$
3: $\mathcal{S} \leftarrow \mathcal{S}\setminus\mathcal{I}$
4: $\mathcal{L} \leftarrow \mathcal{L}\cup\mathcal{I}$

ここで，HSIC $(\mathcal{S}\setminus\{i\})$ は，$\mathcal{S}$ に含まれる属性のうち i 以外を用いて HSIC を計算すること，すなわち leave-one-out 方式を意味する．

4 活性空間における特徴選択

4.1 回帰における HSIC

前節では，相互作用予測，すなわち 2 クラスの識別のための特徴選択に HSIC を応用した．カーネル空間において特徴選択を可能とする評価基準の中で，HSIC は計算の効率性が高い．さらに，HSIC は回帰にも容易に拡張でき，その意味で汎用性が高い．したがって，前節と同じ議論が活性予測においても成り立つ．

具体的には，テンソル積カーネルを介して相互作用（活性）空間を構成するとき，相互作用予測ではサンプルに対するカーネル行列 $\boldsymbol{K}$ を (5) で与えたが，活性予測では以下の式で再定義すればよい[1]．（さらに一般的な議論については [4] を参照されたい．）

$$\boldsymbol{K} = (\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n)^{-1}\,(\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}) \quad (6)$$

ここで，λ は正則化パラメータ，$\boldsymbol{I}_n\in\mathbb{R}^{n\times n}$ は単位行列を表す．ただし，$\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}$ は中心化されているものとする．一方，ラベルに対するカーネル行列 $\boldsymbol{L}$ については，y_i が連続変数になるのみで，前節と同様，$L_{ij}=y_iy_j$ で与えられる．したがって，(4) より

$$\mathrm{Tr}\big((\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n)^{-1}\,(\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot})\,\boldsymbol{L}\big) \quad (7)$$

の最大化が主たる問題となる．

ところが，(7) では $n\times n$ 行列 $\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n$ に対する逆行列計算 $(\cdot)^{-1}$ が必要となるため，(5) の場合と比べて計算量が増大する．特に，BAHSIC アルゴリズムでは，各属性に関して leave-one-out 方式で HSIC を計算する必要があるため，(6) を直接計算するのは現実的ではない．

4.2 アルゴリズムの効率化

この問題に対処するため，効率的な BAHSIC アルゴリズムを提案する．本節では，概略のみを示す．

簡単な計算により，(7) の最大化問題は

$$\mathrm{Tr}\big((\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n)^{-1}\,\boldsymbol{L}\big)$$

の最小化問題に帰着できる．したがって，

$$(\boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n)^{-1}$$

に関して leave-one-out

$$\big((\boldsymbol{K}_{chem}-\boldsymbol{f}_i\boldsymbol{f}_i^\top)\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n\big)^{-1} \quad (8)$$

を効率的に計算できればよい．$\boldsymbol{f}_i\in\mathbb{R}^n$ は除外する化合物属性 i のベクトルを表す．

簡単のため，

$$\begin{aligned}\boldsymbol{P} &= \boldsymbol{K}_{chem}\circ\boldsymbol{K}_{prot}+\lambda\boldsymbol{I}_n \quad \in\mathbb{R}^{n\times n}\\ \boldsymbol{Q} &= (\boldsymbol{f}_i\circ\boldsymbol{g}_1,\ldots,\boldsymbol{f}_i\circ\boldsymbol{g}_k) \quad \in\mathbb{R}^{n\times k}\end{aligned}$$

とする．ここで，

$$\boldsymbol{K}_{prot}=\boldsymbol{G}\boldsymbol{G}^\top,\quad \boldsymbol{G}=(\boldsymbol{g}_1,\ldots,\boldsymbol{g}_k)\in\mathbb{R}^{n\times k} \quad (9)$$

である．このとき，(8) は Sherman-Morrison-Woodbury の公式 [5] により以下の式に変形できる．

$$\boldsymbol{P}^{-1}+\boldsymbol{P}^{-1}\boldsymbol{Q}(\boldsymbol{I}_k-\boldsymbol{Q}^\top\boldsymbol{P}^{-1}\boldsymbol{Q})^{-1}\boldsymbol{Q}^\top\boldsymbol{P}^{-1}$$

[1] 厳密に言えば，リッジ回帰の場合がそうである．[18]

ゆえに，(9) によりタンパク質のカーネル行列をランク $k \ll n$ の行列で近似すれば，各 $i \in \mathcal{S}$ について $n \times n$ 行列ではなく，$k \times k$ 行列 $\boldsymbol{I}_k - \boldsymbol{Q}^\top \boldsymbol{P}^{-1} \boldsymbol{Q}$ に対する逆行列計算をおこなえばよく，効率化を図ることができる[2]．たとえば，**6節**で用いるデータのサンプル数は $n = 798$ であるが，タンパク質の種類は 14 であるため，$k \leq 14$ に削減できる．なお，低ランク行列による近似は，不完全コレスキー分解 [4] などにより可能である[3]．

5 関連研究

特徴選択のアプローチは，評価基準と予測器（識別器やリグレッサー）との関係において，filter, wrapper, embedded に大別できる [7]．カーネル空間における特徴選択に関する研究は，本論文のベースとなっている [18] の他，[14, 23] などが報告されているが，HSIC や [14, 23] で提案されているクラス分離度基準はいずれも，予測器とは独立して特徴を評価する filter である．予測器の性能に基づいて特徴を評価する wrapper や，SVM-RFE [8] に代表される embedded アプローチは，カーネル空間においても原理的には適用可能であるが，計算量が大きくなる．一方，filter は予測器に依存しないため，一般に計算量の観点から実用的である．

HSIC との関連研究としては他に， Kernel Target Alignment (KTA) [3] が挙げられる．KTA はカーネルに関するモデル選択基準として以前から知られているが，カーネル空間における特徴選択の基準として用いることもできる．特に，2 クラス識別においては，HSIC との密接な関連が示されている [19]．また，HSIC はクラス分離度基準とも関連づけられる [22]．

なお，本研究で扱っている回帰のための特徴選択は，[4] の一例とみなすことができる．

6 実データへの適用

前節で提案した効率的アルゴリズムを，活性予測に寄与する化合物属性の同定に適用した．同定した化合物属性は，予測性能および化学的知見の観点から評価した．ここでは，シトクロム P450 (CYP) の阻害活性データへの適用結果を示す．

CYP は薬物代謝において重要な役割を果たす酵素であり，ヒトではおよそ 60 のアイソフォーム[4]が知られている．CYP は多様な化合物の代謝に関わっており，分子認識の特異性が広いため，分子認識に関与する特徴を特定するのは容易ではない．ケミカルゲノミクスデータからの知識発見は，これに対する有望なアプローチと考えられ，活性に関与する化合物属性を抽出することは，分子認識メカニズムの理解に役立つと期待される．

6.1 データおよび予測性能の評価

CYP の阻害活性データは [10] から入手した．化合物とタンパク質（CYP）のペアに対して阻害活性の値 (IC_{50}) が付与されている．化合物の種類は 371，CYP の種類（アイソフォーム）は 14，そしてペアの数 n は 798 である．

各化合物の構造データファイルから，DragonX [20] (http://www.talete.mi.it/) を用いて化合物の属性値を算出し，そのうち，物理化学的特性や官能基の数など，解釈が比較的容易な 345 種類を選出した．さらに 371 すべての化合物において属性値が同一のものを除外した．その結果，139 種類が解析の対象となった．

ここでは，化合物属性を同定することが目的であるため，化合物はベクトル表現するが，タンパク質は非ベクトル表現も許容できる．そこでタンパク質のカーネルとして，以下の 3 種類を用いた．

- PROFEAT [13] の特徴ベクトル＋ RBF カーネル
- ミスマッチカーネル (Mismatch) [12]
- 局所アラインメントカーネル (LA) [16]

予測性能は，すべての属性を用いる場合と，徐々に属性数を削減した場合について，以下の 2 種類のリグレッサーを用いて評価した．

- サポートベクトル回帰 (SVR) [21]
- カーネルリッジ回帰 (KRR) [17]

これらはカーネルに基づくリグレッサーとして代表的なものであり，諸問題において高い予測性能が報告されていることから，本問題においても有用であると考えられる．なお，属性選択の効率的な計算が可能となるのは，化合物のカーネルが線形カーネルの場合に限定されるが，属性選択とリグレッサーによる予測は独立したプロセスであるので，予測には線形カーネルに加えて RBF カーネルも用いた．予測性能の指標としては，

$$r^2 = \frac{(n\sum_{i=1}^n \hat{y}_i y_i - \sum_{i=1}^n \hat{y}_i \sum_{i=1}^n y_i)^2}{(n\sum_{i=1}^n \hat{y}_i^2 - (\sum_{i=1}^n \hat{y}_i)^2)(n\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2)}$$

を用いた．ここで，$\hat{y}_i$ は x_i の予測値である．r^2 値が大きいほど，真の値と相関が高く，予測性能が高いことを示す．

$n = 798$ のサンプルを，学習サンプルとテストサンプルにランダムに 6:1 に分割し，学習サンプルのみを用い

[2]なお $\boldsymbol{P}^{-1}$ の計算は 1 回だけでよい．
[3](9) の計算も 1 回だけでよい．
[4]機能は類似しているが，アミノ酸配列の異なるタンパク質のこと．

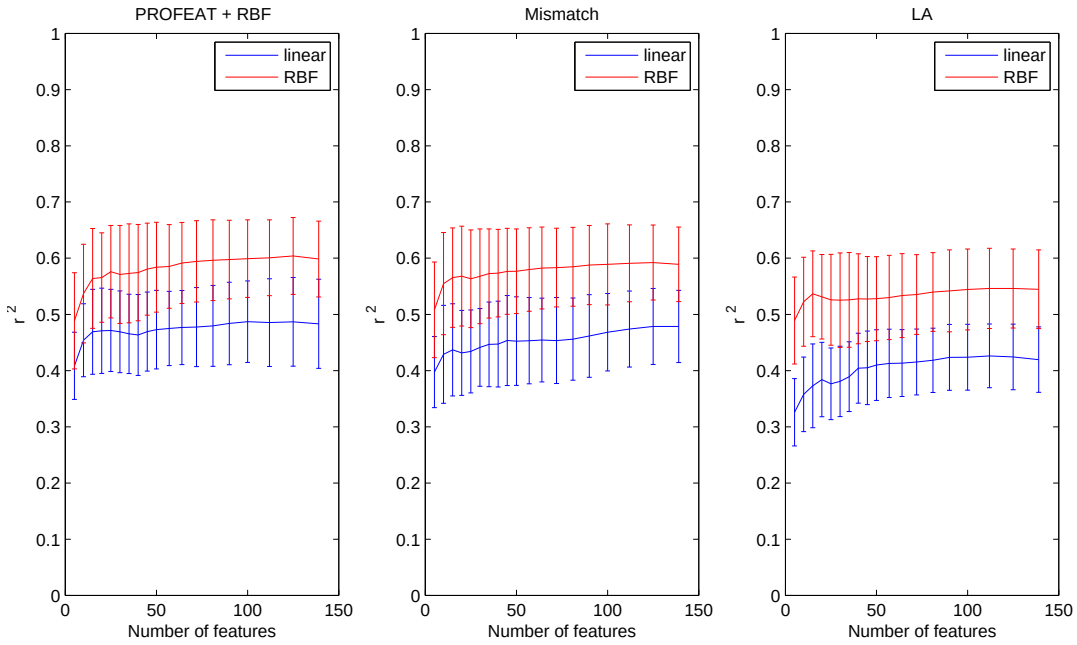

図 1: SVR の予測性能

て属性選択およびリグレッサーの構築をおこない，テストサンプルに適用した．このプロセスを 20 回繰り返し，r^2 値の平均，標準偏差を算出した．

6.2 結果

SVR を用いた予測結果を図 1 に示す．タンパク質のカーネルは，PROFEAT+RBF, Mismatch, LA それぞれについて，属性選択，予測ともに一貫して同じものを用いている．一方，化合物のカーネルについては，属性選択では線形カーネル，予測では線形カーネルと RBF カーネルを用いている．タンパク質カーネルによって，性能に多少の違いはあるが，r^2 値の平均値としては，線形カーネルでは最大 0.48, RBF カーネルでは最大 0.60 程度の比較的高い値が得られた．

本研究の目的は，特徴選択アルゴリズムを用いて予測に有効な属性を同定することにあるので，すべての属性を用いた場合と，属性を絞り込んだ場合において，予測性能がどのように変化するかに関心がある．図 1 から，属性数を 20〜30 に絞り込んでも，139 すべてを用いた場合と同程度の予測性能を維持できることが分かる．見方を変えれば，予測に無関係あるいは冗長な属性を排除できたと考えられる．この傾向は，いずれのカーネルを用いた場合においても同様であった．

図 2 に KRR を用いた予測結果を示す．r^2 値の平均値としては，線形カーネルでは最大 0.49, RBF カーネルでは最大 0.58 程度が得られ，SVR と同程度の予測性能を示した．ここでも注目すべきは，少数の属性だけでも，高い予測性能を維持できることである．つまり，活性予測に寄与する属性を同定できたことが示唆される．

同定した属性が予測に有用であることは示唆されたが，実際にそれらが活性に関与するものであるか，化学的知見の観点から妥当性を評価することが次の重要なステップである．一般に，どのタンパク質カーネルを用いるかによって，得られる化合物属性の順位リストは異なってくるが，今回用いた CYP データに関しては，上位 20 の化合物属性は高い一致を示した．たとえば，PROFEAT + RBF を用いた場合，分子の疎水性の指標であるオクタノール-水分配係数に関わる属性が 1, 2, 7, 9 位に位置した．CYP1A2 や CYP3A4 の阻害剤は高い脂溶性を示すことが知られており [2, 11]，実際に，データに含まれる 14 種類の CYP のうち，CYP1A2, CYP3A4 の阻害活性データは 34 % を占めることから，この結果は妥当と考えられる．また，分極率に関わる属性が 8, 13 位に位置したが，CYP3A4 の阻害には分極率が関与していることが知られており [11]，矛盾のない結果と言える．このように，上位の属性のうち，先行研究と関連付けられるものは少なくなかったが，むしろ，関与が明らかにされていない特徴を見つけ出し，タンパク質の分子認識メカニズムとの接点を探っていくことが，今後の方向性として重要となるであろう．

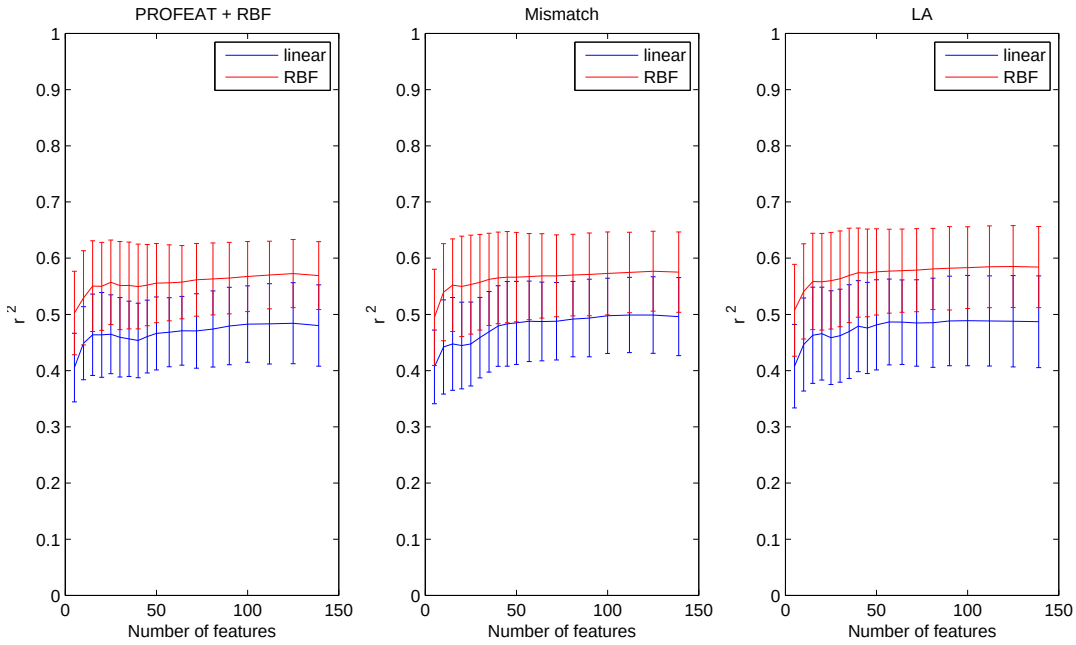

図 2: KRR の予測性能

7 おわりに

本研究では，活性予測に寄与する化合物属性を同定するために，化合物-タンパク質の活性空間において属性選択をおこなう手法について論じた．特に，回帰のための特徴選択において生じる計算量の問題に対処するために，効率的アルゴリズムを提案した．実際に，CYP の阻害活性データに適用し，予測に有効な属性を同定できることおよび，それらは化学的知見と照合しても妥当であることを示した．本論文では，ケミカルゲノミクスの領域において特徴選択およびその効率化について論じたが，同様のアプローチは様々な実問題に適用できると考えられる．

参考文献

[1] J. Bajorath, “Computational analysis of ligand relationships within target families”, *Curr Opin Chem Biol*, **12**, 352–358, 2008

[2] K. K. Chohan, S. W. Paine, J. Mistry, P. Barton, A. M. Davis, “A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries”, *J Med Chem*, **48**, 5154-5161, 2005

[3] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, J. S. Kandola, “On Kernel-Target Alignment”, *Advances in Neural Information Processing Systems 14*, 367–373, 2001

[4] K. Fukumizu, F. R. Bach, M. I. Jordan, “Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces”, *J Mach Learn Res*, **5**, 73–99, 2004

[5] G. H. Golub, C. F. Van Loan, *Matrix Computations, 3rd edition*, Johns Hopkins University Press, Baltimore, 1996

[6] A. Gretton, O. Bousquet, A. J. Smola, B. Schölkopf, “Measuring statistical dependence with Hilbert-Schmidt norms”, *Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory*, 63–78, 2005

[7] I. Guyon, A. Elisseeff, “An Introduction to Variable and Feature Selection”, *J Mach Learn Res*, **3**, 1157–1182, 2003

[8] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, “Gene Selection for Cancer Classification Using Support Vector Machines”, *Mach Learn*, **46**, 389–422, 2002

[9] L. Jacob, J.-P. Vert, “Protein-ligand interaction prediction: an improved chemogenomics approach”, *Bioinformatics*, **25**, 2149–2156, 2008

[10] A. Kontijevskis, J. Komorowski, J. E. S. Wikberg, "Generalized Proteochemometric Model of Multiple Cytochrome P450 Enzymes and Their Inhibitors", *J Chem Inf Model*, **48**, 1840–1850, 2008

[11] J. M. Kriegl, T. Arnhold, B. Beck, T. Fox, "Prediction of Human Cytochrome P450 Inhibition Using Support Vector Machines", *QSAR Comb Sci*, **24**, 491-502, 2005

[12] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, W. S. Noble, "Mismatch string kernels for discriminative protein classification", *Bioinformatics*, **20**, 467–476, 2004

[13] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, Y. Z. Chen, "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence", *Nucleic Acids Res*, **34**, W32–W37, 2006

[14] S. Niijima, S. Kuhara, "Gene subset selection in kernel-induced feature space", *Pattern Recognit Lett*, **27**, 1884–1892, 2006

[15] D. Rognan, "Chemogenomic approaches to rational drug design", *Br J Pharmacol*, **152**, 38–52, 2007

[16] H. Saigo, J.-P. Vert, N. Ueda, T. Akutsu, "Protein homology detection using string alignment kernels", *Bioinformatics*, **20**, 1682–1689, 2004

[17] C. Saunders, A. Gammerman, V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", *Proceedings of the Fifteenth International Conference on Machine Learning*, 515–521, 1998

[18] L. Song, J. Bedo, K. M. Borgwardt, A. Gretton, A. Smola, "Gene selection via the BAHSIC family of algorithms", *Bioinformatics*, **23**, i490–i498, 2007

[19] L. Song, A. Smola, A. Gretton, K. M. Borgwardt, J. Bedo, "Supervised feature selection via dependence estimation", *Proceedings of the Twenty-Fourth International Conference on Machine Learning*, 823–830, 2007

[20] R. Todeschini, V. Consonni, M. Pavan, *Dragon*, Milano Chemometrics and QSAR Research Group, Milan, 2007

[21] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc., New York, 1998

[22] H. Xiong, M. N. S. Swamy, M. O. Ahmad, "Optimizing the kernel in the empirical feature space", *IEEE Trans Neural Netw*, **16**, 460–474, 2005

[23] L. Wang, "Feature selection with kernel class separability", *IEEE Trans Pattern Anal Mach Intell*, **30**, 1534–1546, 2008