厚生労働科学研究費補助金

化学物質リスク研究事業

新規な化学物質の安全性を予測・評価する新規計算手法の開発

平成22年度　総括研究報告書

研究代表者　奥野　恭史

平成22（2011）年　4月

目　次

厚生労働科学研究費補助金（化学物質リスク研究事業）
総括研究報告書

新規な化学物質の安全性を予測・評価する新規計算手法の開発

研究代表者　奥野　恭史　京都大学薬学研究科 教授

研究要旨

化学物質の安全性評価のための実測試験には膨大なコストと時間を要するため、実測試験に代わるアプローチとして、構造活性相関手法などの計算的アプローチの積極的利用が国際的に推進されている。計算的アプローチを用いる際の最大のポイントは、計算予測が実測試験の代用に匹敵する信頼性を有するかという点であるが、とりわけ、新規物質の予測能を持たせた上での適用範囲の設定が大きな問題となっている。しかしながら、現状の予測モデルは、実測データを有する既知化学物質という限られた化学物質データを用いて構築されたものであり、モデル構築に用いられなかった真に新規な化学物質に対する予測性能の評価や新規物質そのものの予測に主眼をおいた方法論の研究は国内外で皆無に等しい。そこで本研究では、OECDのSARバリデーション原則に基づいた、新規な化学物質に対する予測能の評価方法の開発と新規化学物質の予測性能を最大限発揮する計算アルゴリズムの開発を目指す。また、ケミカルゲノミクス情報やトキシコゲノミクス情報を用いることにより反復毒性、遺伝毒性などヒトの安全性に関する毒性のメカニズム解析を行う。平成22年度では下記1と2について着手した。

1.　化学物質および安全性評価実測データの収集（平成22年度）
2.　ケミカルスペースを用いた既存化学物質と新規化学物質との特徴空間比較手法の開発（平成22年度、23年度）
3.　エンドポイント予測と新規化学物質予測に最適なトレーニングセットのサンプリング手法と特徴抽出アルゴリズムの開発による構造活性相関モデルの構築（平成23年度、24年度）
4.　化学物質の特徴解析と標的生体分子予測によるメカニズム解析（平成24年度）

期待される効果としては、本研究により真に新規な化学物質の安全性評価に適用できる計算手法が確立され、より多くの化学物質の総合的かつ合意的な評価が迅速に実現されることが挙げられる。これにより厚生労働行政における化学物質の安全性審査に的確な評価を与え、国民の健康保持におおいに資するものと考える。加えて、計算により、化学物質の実測試験に必要な膨大なコストと時間の劇的削減を実現することから、我が国の財政的貢献も大いに期待できる。

A．研究目的

化学物質の安全性評価のための実測試験には膨大なコストと時間を要するため、実測試験に代わるアプローチとして、構造活性相関手法などの計算的アプローチの積極的利用が国際的に推進されている。計算的アプローチを用いる際の最大のポイントは、計算予測が実測試験の代用に匹敵する信頼性を有するかという点であり、これらを評価する基準としてOECDでは5項目からなる「SARバリデーション原則：①エンドポイントの定義 、②曖昧さのないアルゴリズム、③化学物質の適用範囲の定義、④予測性能（計算精度、頑健性、新規予測能）の適切な評価、⑤活性メカニズムに関する解釈（可能な化合物のみを対象）」を定め、世界各国で各エンドポイントにおける構造活性相関モデルの開発が行われている。しかしながら、OECDバリデーション原則の適用範囲（項目③）とその予測性能（項目④）とはトレードオフの関係にあり（例えば、適用範囲を狭めれば予測精度は確実に向上するが新規物質の予測には適用できない事態に陥ることになる）、新規物質の予測能を持たせた上での適用範囲の設定が大きな問題となっている。実際、現状の予測モデルは、実測データを有する既知化学物質という限られた化学物質データを用いて構築されたものであり、モデル構築に用いられなかった真に新規な化学物質に対する予測性能の評価や新規物質そのものの予測に主眼をおいた方法論の研究は国内外で皆無に等しい。本研究では、OECDのSARバリデーション原則に基づいた、新規な化学物質に対する予測能の評価方法の開発と新規化学物質の予測性能を最大限発揮する計算アルゴリズムの開発を目指す。また、ケミカルゲノミクス情報やトキシコゲノミクス情報を用いることにより反復毒性、遺伝毒性などヒトの安全性に関する毒性のメカニズム解析を行う。初年度である22年度は、化学物質および安全性評価実測データの収集と、ケミカルスペースを用いた既存化学物質と新規化学物質との特徴空間比較手法の開発に着手した。

B．研究方法

1．化学物質および安全性評価実測データの収集

本研究に用いる評価用化学物質データとして、CTD、HGNC、MATADOR、PharmGKB、OMIM、GnD、DrugBank、Label、AERS、MedDRA、GVK_MBT、Entrez geneの公的データベースから、化学物質の構造情報、標的タンパク情報、関連生物活性情報を抜き出し、独自のデータベース構築を行った。また、新規化学物質として、㈱ナミキ商事が取り扱う世界各国のサプライヤーが提供する市販化合物ライブラリーや米国NCBIのPubChemデータベースに集積された5000万以上の化学物質を収集し、内部データベースを構築した。

2．ケミカルスペースを用いた既存化学物質と新規化学物質との特徴空間比較手法の開発

ケミカルスペースを用いた既存化学物質と新規化学物質との特徴空間比較手法として、多変量解析法の一つである正準相関分析の可能性を検討した。上記で収集したDrugBankデータベース由来の化学物質と標的タンパク質の相互作用関係における正準相関分析を行い、構築されたケミカルスペースにPubChemデ

ータベース由来の化学物質1000万種（ランダム抽出）を投射した。

3．化学物質の活性情報（相互作用情報）のデータ表現:化合物とタンパク質の異種データの統合

化合物とタンパク質の相互作用関係を機械学習するためには、化合物とタンパク質の異種のデータ表現を統合し、相互作用関係を定量化する数理的枠組みが必要となる。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問題となる。我々は、化合物ベクトルとタンパク質ベクトルを統合する有力な手段としてカーネル法を用いた。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。

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

ここで、⊗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、$\boldsymbol{\Phi}(c)$と$\boldsymbol{\Psi}(p)$のそれぞれdc，dp次元の特徴ベクトルとしてexplicit に与えた場合、ペアはdc,×dp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、計算量が膨大となり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

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

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

$$kchem(c,c') \equiv \boldsymbol{\Phi}(c)^{\intercal}\boldsymbol{\Phi}(c') \qquad (2)$$

$$kprot(p,p') \equiv \boldsymbol{\Psi}(p)^{\intercal}\boldsymbol{\Psi}(p') \qquad (3)$$

とした場合に、化合物-タンパク質ペアのカーネルは、

$$\begin{aligned}k\big((c,p),(c',p')\big) &\equiv \boldsymbol{\Pi}(c,p)^{\intercal}\boldsymbol{\Pi}(c',p') \\ &= kchem(c,c') \times kprot(p,p')\end{aligned}$$

となる。つまり、(eq. 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。本研究では、このようなカーネル表現を導入することで化学物質の活性予測や特徴抽出を試みた。

4．カーネル空間における化学物質の特徴選択

上述のとおり、カーネル法を用いた数理モデル用いた特徴選択を行う場合、カーネル空間上での　特徴選択を行うことが望ましい。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。本研究では，特に計算の効率性を考慮し，Hilbert-Schmidt Independence Criterion (HSIC) を評価基準として採用した。

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。xi をサンプル、yi をそのクラスラベルとし、n 個の学習サンプル$(x1, y1), \ldots\ldots, (xn, yn)$が与えられているとする。このとき、HSIC の経験推定量は、$xi, yi\ (i = 1, \ldots, n)$ 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。

$$\mathrm{HSIC} = \frac{1}{(n-1)^2}\,\mathrm{Tr}(\boldsymbol{KL}) \qquad (4)$$

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、xi =(ci, pi) が相互作用するペアであれば yi = 1、そうでなければ yi = -1 であり、Lij = yiyj 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、

$$\boldsymbol{K} = \boldsymbol{Kchem} \circ \boldsymbol{Kprot} \qquad (5)$$

と表される。ここで、∘ はアダマール積を表す。すなわち、$\boldsymbol{Kchem}, \boldsymbol{Kprot} \in \mathbb{R}^{n \times n}$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。

5．HSICの回帰モデルへの拡張

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。

$$\boldsymbol{K} = (\boldsymbol{Kchem} \circ \boldsymbol{Kprot} + \lambda \boldsymbol{I}_n)^{-1}(\boldsymbol{Kchem} \circ \boldsymbol{Kprot})$$

ここで、, 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、$\boldsymbol{Kchem} \circ \boldsymbol{Kprot}$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、yiが連続変数になるのみで、前節と同様、Lij = yiyj で与えられる。したがって、(4) より

$$\mathrm{Tr}((\boldsymbol{Kchem} \circ \boldsymbol{Kprot} + \lambda \boldsymbol{I}_n)^{-1}(\boldsymbol{Kchem} \circ \boldsymbol{Kprot})\boldsymbol{L})$$

の最大化が主たる問題となる。

（倫理面への配慮）

本年度実施した研究は計算機アルゴリズムの開発のみであり、倫理面に関する問題は一切無い。

C．研究結果と考察

1．化学物質および安全性評価実測データの収集

本研究に用いる評価用化学物質データとして、CTD、HGNC、MATADOR、PharmGKB、OMIM、GnD、DrugBank、Label、AERS、MedDRA、GVK_MBT、Entrez geneの公的

データベースから、標的タンパク質情報、薬理活性情報、毒性情報などの活性情報を有する化合物の収集を行った。収集したデータベースの統計を表１に示す。表１に示す通り、141,475件の化学物質数に関する多種多様な活性情報の収集に成功した。
また、これらの大量データを整理、検索することを意図した独自データベースを開発した。当該データベースでは、集積した化学物質14,475件について、化学物質検索、標的遺伝子情報検索、化学物質活性情報検索や検索結果のクラスタリング等のデータマイニング機能も付与した。（図１、図２）本データベース、http://cgs.pharm.kyoto-u.ac.jp/services/ddida/より公開予定である。

２．ケミカルスペースを用いた既存化学物質と新規化学物質との特徴空間比較手法の開発

図３－１は、DrugBankデータベース由来の化学物質を正準相関座標系にプロットしたケミカルスペースである。（緑色の細かな点が各化学物質に対応している。）これに対し、図３－２はPubChemデータベース由来の化学物質をプロットしたものである。（ここでは白色の細かな点が各化学物質に対応している。）2図の比較からわかるように、PubChem由来の化学物質の分布は、DrugBank由来の化学物質の分布の全域に重なる化学物質群のほかに、異なる分布を示す化学物質のクラスター（矢印）が存在することがわかる。DrugBankは医薬品由来物質を中心に集積されたデータベースであることから、生物にとって安全性の高い化学物質が集積されているものと考えられる。従って、DrugBank由来のケミカルスペースに重ならないPubChem由来の化学物質は、医薬品とは成り得ない化学物質群が集積している可能性が高いものと考えられる。

３．カーネル空間における化学物質の特徴選択

予測モデルの性能評価をするために、ここでは、シトクロムP450（CYP）の阻害活性データへの適用を試みた。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。
CYPの阻害活性データはから入手した。化合物とタンパク質（CYP）のペアに対して阻害活性の値(IC50) 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 (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、タンパク質のカーネルとして、以下の3 種類を用いた。

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

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、リグレッサーとして、サポートベクトル回帰(SVR)を用いて評価した。
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値を用いた。検証試験は、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。SVRを用いた予測結果を図４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。

D. 結論

本年度は初年度であることからも、3年間の研究期間を通じて用いる化学物質の評価データセットの収集・整理と、ケミカルスペースを用いた既存化学物質と新規化学物質との特徴空間比較手法の開発に着手した。前者では、141,475件もの化学物質に関する情報を集積し、標的タンパク質や活性情報とのリレーションを実現するデータベースを構築した。当該データベースは、http://cgs.pharm.kyoto-u.ac.jp/services/ddida/より公開する予定である。また、後者に関しては、特徴空間比較手法の開発として、多変量解析の一つである正準相関分析法の適用性について検討した。またこれらの要素技術として、カーネル空間に基づく化学物質の特徴抽出アルゴリズムの開発を行った。このように、当該年度では、当初計画通りの成果を得るに至った。

E. 健康危険情報

特記事項無し

F. 研究発表

1．論文発表

1. Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., Okuno, Y. "Analysis of multiple compound-protein interactions reveals novel bioactive molecules" *Mol. Syst. Biol.* 7, 472, 2011

2. Niijima, S., Yabuuchi, H., Okuno, Y. "Cross-target view to feature selection: identification of molecular interaction features in ligand-target space" *J. Chem. Inf. Model.*, 51, 15-24, 2010

3. Kusaka, M., Katoh-Fukui, Y., Ogawa, H., Miyabayashi, K., Baba, T., Shima, Y., Sugiyama, N., Sugimoto, Y., Okuno, Y., Kodama, R., Iizuka-Kogo, A., Senda, T., Sasaoka, T., Kitamura, K., Aizawa, S., Morohashi, KI. "Abnormal Epithelial Cell Polarity and Ectopic Epidermal Growth Factor Receptor (EGFR) Expression Induced in Emx2 KO Embryonic Gonads" *Endocrinology*, 151, 5893-5904, 2010

4. van der Horst, E., Peironcely, JE., Ijzerman, AP., Beukers, MW., Lane, JR., van Vlijmen, HW., Emmerich, MT., Okuno, Y., Bender, A. "A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization" *BMC Bioinformatics*, 11, 316, 2010

5. Tamba, S., Yodoi, R., Morimoto, K., Inazumi, T., Sukeno, M., Segi-Nishida, E., Okuno, Y., Tsujimoto, G., Narumiya, S., Sugimoto, Y. "Expression profiling of cumulus cells reveals functional changes during ovulation and central roles of prostaglandin EP2 receptor in cAMP signaling" *Biochimie*, 92, 665-675, 2010

6. Hagihara, M., Yoneda, K., Yabuuchi, H., Okuno, Y., Nakatani, K. "A reverse transcriptase stop assay revealed diverse quadruplex formations in UTRs in mRNA" *Bioorg. Med. Chem. Lett.*, 20, 2350-2353, 2010

7. Terada, N., Shimizu, Y., Kamba, T., Inoue, T., Maeno, A., Kobayashi, T., Nakamura, E., Kamoto, T., Kanaji, T., Maruyama, T., Mikami, Y., Toda, Y., Matsuoka, T., Okuno, Y., Tsujimoto, G., Narumiya, S., Ogawa, O. "Identification of EP4 as a potential target for the treatment of castration-resistant prostate cancer using a novel xenograft model" *Cancer Res.*, 70, 1606-1615, 2010

8. Takahashi, J., Hijikuro, I., Kihara, T., Murugesh, M.G., Fuse, S., Kunimoto, R., Tsumura, Y., Akaike, A., Niidome, T., Okuno, Y., Takahashi, T., Sugimoto, H. "Design, synthesis, evaluation and QSAR analysis of N(1)-substituted norcymserine derivatives as selective butyrylcholinesterase inhibitors" *Bioorg. Med. Chem. Lett.*, 20, 1718-1720, 2010

9. Doi, M., Takahashi, Y., Komatsu, R., Yamazaki, F., Yamada, H., Haraguchi, S., Emoto, N., Okuno, Y., Tsujimoto, G., Kanematsu, A., Ogawa, O., Todo, T., Tsutsui, K., van der Horst, G.T., Okamura, H. "Salt-sensitive hypertension in circadian clock-deficient Cry-null mice involves dysregulated adrenal Hsd3b6" *Nat. Med.*, 16, 67-74, 2010

2．学会発表（招待講演のみ）

1. The International Chemical Congress of Pacific Basin Societies (PacifiChem 2010) "Reverse Polypharmacology to explore novel bioactive molecules" (2010.12.18)

2. 第 33 回情報化学討論会（日本化学会　情報化学部会　主催）「Data mining of multiple compound-protein interactions for drug discovery」(2010.10.31)

3. 株式会社ペプチド研究所 フィッシャー祭（ペプチド研究所　彩都研究所　主催）「Chemical Space Travel: インフォマティクスによる包括的な活性化合物探索」(2010.10.15)

4. CBI 学会 2010 年大会（情報計算化学生物学会　主催）「分子標的創薬からシステム創薬へ」(2010.9.16)

5. 生物有機科学研究所セミナー・コロキウム（(財)サントリー生物有機科学研究所　主催）「ケミカルゲノミクスがもたらす多分子対多分子の多重相互作用とその創薬への可能性」(2010.5.25)

3．著書

1. 奥野恭史他「Handbook of Systems Toxicology」Wiley-Blackwell　2011 年

2. 奥野恭史他「薬学の展望とロードマップ」日本薬学会　2010年

3. 奥野恭史他「医薬ジャーナル　増刊号 新薬展望2010」㈱医薬ジャーナル社　2010年

4．新聞掲載

1. 2010年8月10日　朝日新聞（25面）「生物研究コンピュータ時代－がんの原因、薬の候補を解析－」

G. 知的財産権の出願・登録状況

特記事項無し

表1．公共データベースより集積した化学物質情報の内訳

Resources	Target genes	Chemicals	ADRs	Usages	Chemical-target gene interactions	Chemical-clinical observation relationships		Target gene-ADR relationships
						Chemical-ADR	Chemical-Usage	
Entrez gene	45,338							
CTD	10,243	135,367	6,687	496	48,811	1,686	2,211	7,088
HGNC	26,392							
MATADOR	2,902	801			15,843			
PharmGKB	25,997	2,332	4,203		1,654	1,241		3,176
OMIM	7,116		4,678					5,211
GnD	113		95					
DrugBank	2,537	6,824			8,779			121
Label		1,617	1,526	2,426		107,608	15,044	
AERS			10,984			210,000		
MedDRA			67,645					
GVK_MBT		13,209	787			3,764		
Unique	**46,262**	**141,475**	**76,293**	**2,777**	**63,137**	**233,032**	**17,213**	**13,611**

図1．独自化学物質データベースのホーム画面（http://cgs.pharm.kyoto-u.ac.jp/services/ddida/）

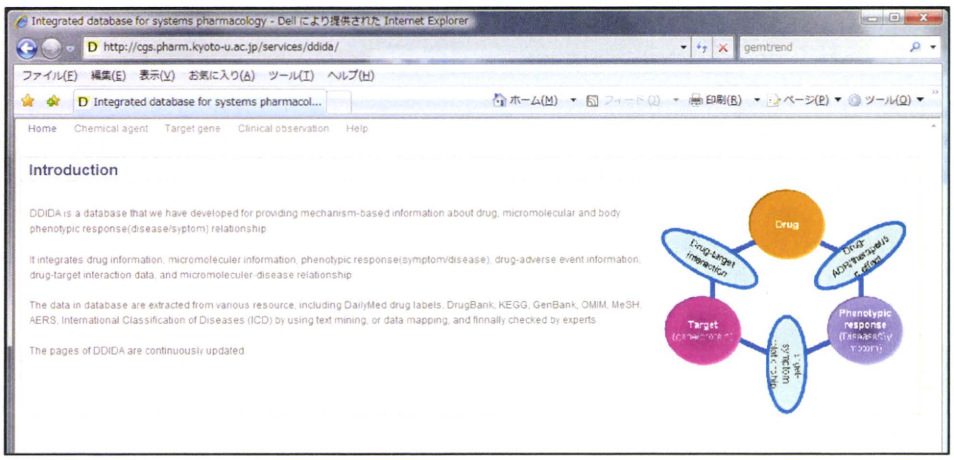


図２．独自化学物質データベースの検索結果例

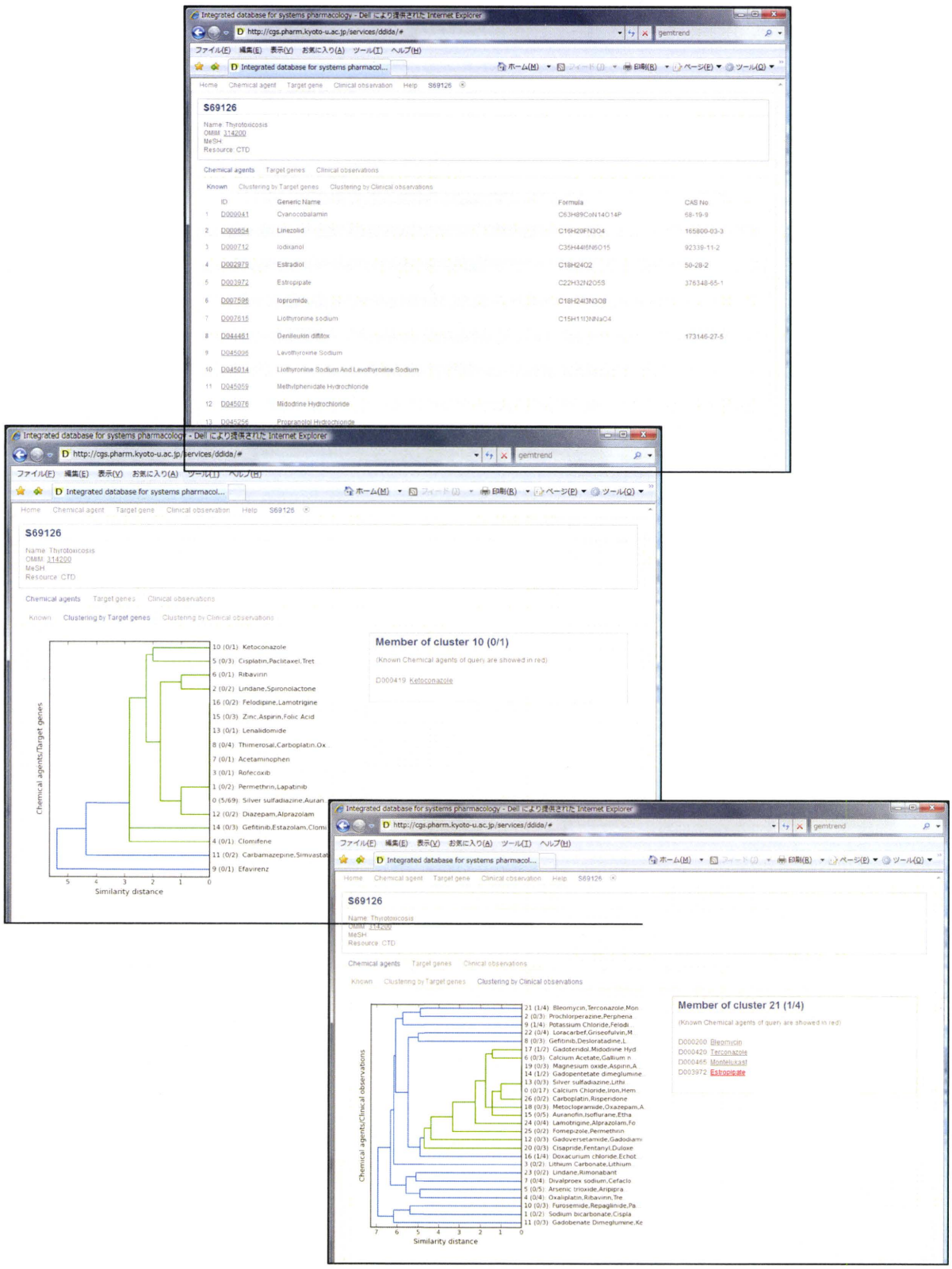


図３－１．正準相関座標上のDrugBank化学物質のケミカルスペース

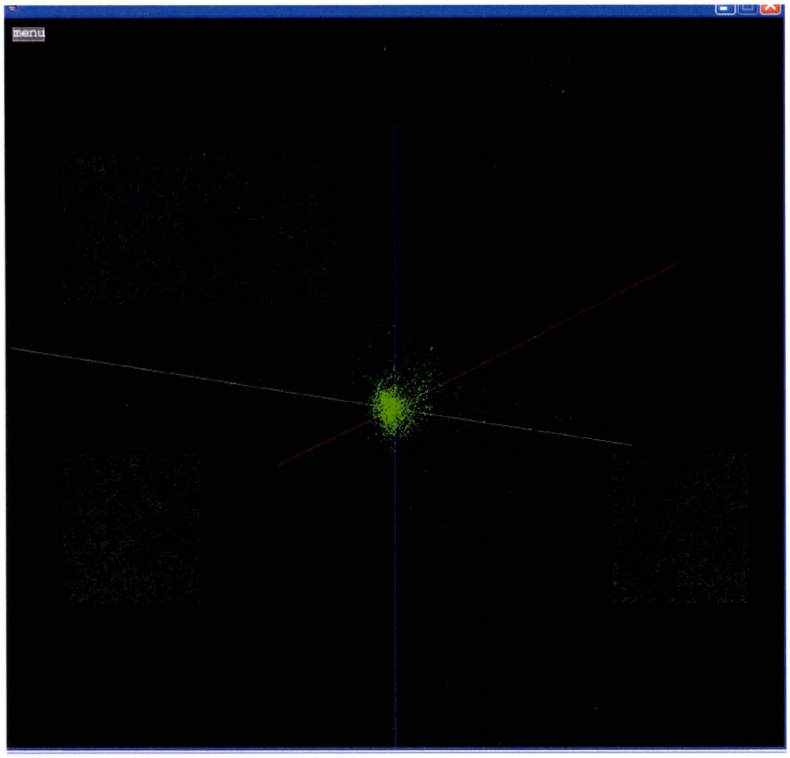


図３－２．正準相関座標上のPubChem化学物質のケミカルスペース

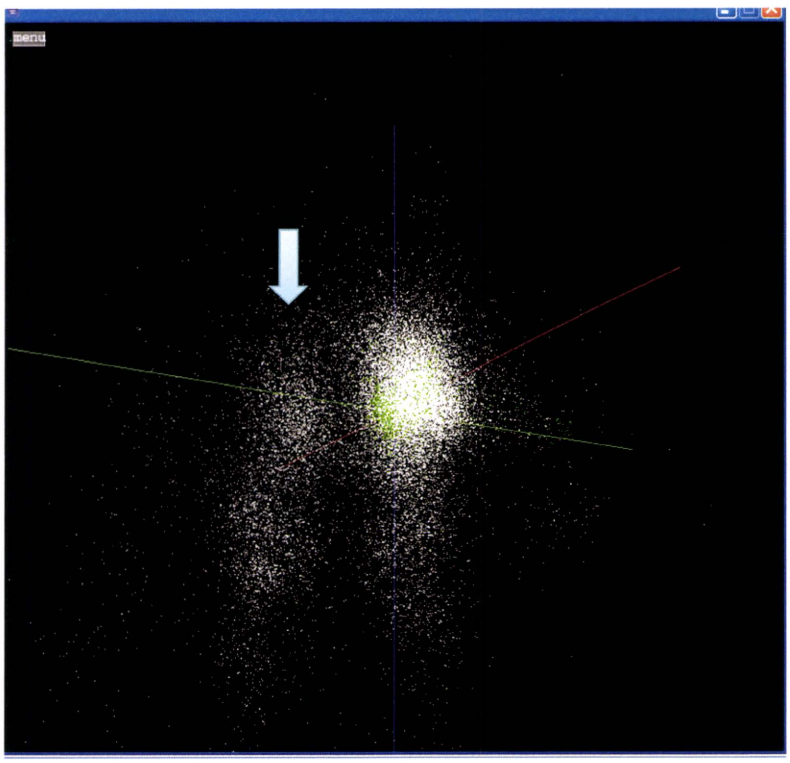
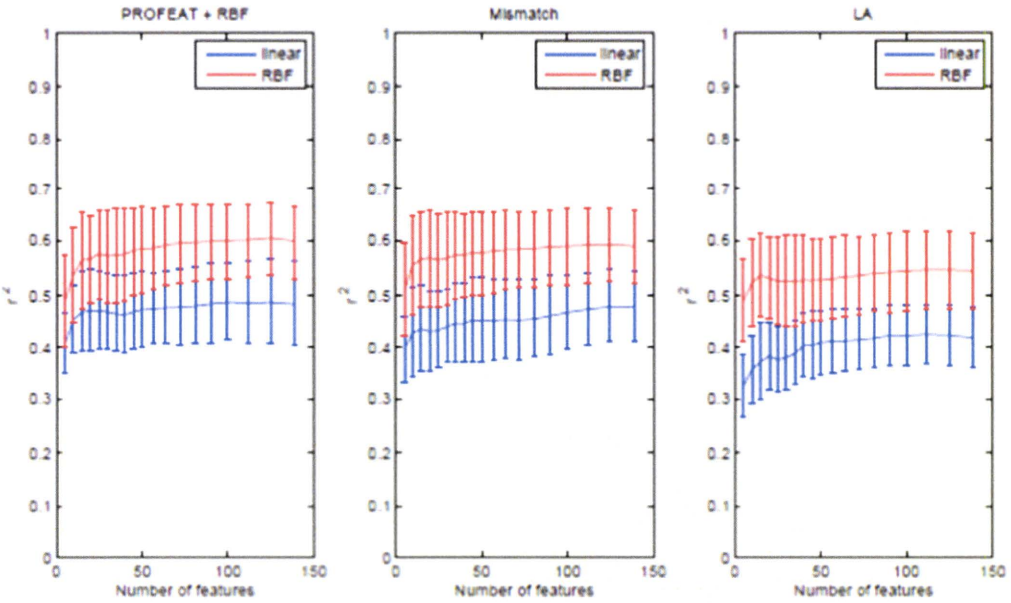


図４．サポートベクター回帰法の予測性能比較



研究成果の刊行に関する一覧表

書籍

著者氏名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年
奥野恭史ほか	Daniel A. Casciano, Dr Saura C. Sahu	Handbook of Systems Toxicology	Wiley-Blackwell	日本	2011
奥野恭史ほか	日本薬学会将来展望委員会	薬学の展望とロードマップ	日本薬学会	日本	2010
奥野恭史ほか	－	医薬ジャーナル　増刊号 新薬展望2010	㈱医薬ジャーナル社	日本	2010

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G., Okuno, Y.	Analysis of multiple compound-protein interactions reveals novel bioactive molecules	**Molecular Systems Biology**	7	472	2011
Niijima, S., Yabuuchi, H., Okuno, Y.	Cross-target view to feature selection: identification of molecular interaction features in ligand-target space	**J. Chem. Inf. Model.**	51	15-24	2010
Kusaka, M., Katoh-Fukui, Y., Ogawa, H., Miyabayashi, K., Baba, T., Shima, Y., Sugiyama, N., Sugimoto, Y., Okuno, Y., Kodama, R., Iizuka-Kogo, A., Senda, T., Sasaoka, T., Kitamura, K., Aizawa, S., Morohashi, KI.	Abnormal Epithelial Cell Polarity and Ectopic Epidermal Growth Factor Receptor (EGFR) Expression Induced in Emx2 KO Embryonic Gonads	**Endocrinology**	151	5893-5904	2010
van der Horst, E., Peironcely, JE., Ijzerman, AP., Beukers, MW., Lane, JR., van Vlijmen, HW., Emmerich, MT., Okuno, Y., Bender, A.	A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization	**BMC Bioinformatics**	11	316	2010

Tamba, S., Yodoi, R., Morimoto, K., Inazumi, T., Sukeno, M., Segi-Nishida, E., Okuno, Y., Tsujimoto, G., Narumiya, S., Sugimoto, Y.	Expression profiling of cumulus cells reveals functional changes during ovulation and central roles of prostaglandin EP2 receptor in cAMP signaling	**Biochimie**	92	665-675	2010
Hagihara, M., Yoneda, K., Yabuuchi, H., Okuno, Y., Nakatani, K.	A reverse transcriptase stop assay revealed diverse quadruplex formations in UTRs in mRNA	**Bioorg. Med. Chem. Lett.**	20	2350-2353	2010
Terada, N., Shimizu, Y., Kamba, T., Inoue, T., Maeno, A., Kobayashi, T., Nakamura, E., Kamoto, T., Kanaji, T., Maruyama, T., Mikami, Y., Toda, Y., Matsuoka, T., Okuno, Y., Tsujimoto, G., Narumiya, S., Ogawa, O.	Identification of EP4 as a potential target for the treatment of castration-resistant prostate cancer using a novel xenograft model	**Cancer Research**	70	1606-1615	2010
Takahashi, J., Hijikuro, I., Kihara, T., Murugesh, M. G., Fuse, S., Kunimoto, R., Tsumura, Y., Akaike, A., Niidome, T., Okuno, Y., Takahashi, T., Sugimoto, H.	Design, synthesis, evaluation and QSAR analysis of N(1)-substituted norcymserine derivatives as selective butyrylcholinesterase inhibitors	**Bioorg. Med. Chem. Lett.**	20	1718-1720	2010
Doi, M., Takahashi, Y., Komatsu, R., Yamazaki, F., Yamada, H., Haraguchi, S., Emoto, N., Okuno, Y., Tsujimoto, G., Kanematsu, A., Ogawa, O., Todo, T., Tsutsui, K., van der Horst, G.T., Okamura, H.	Salt-sensitive hypertension in circadian clock-deficient Cry-null mice involves dysregulated adrenal Hsd3b6	**Nature Medicine**	16	67-74	2010

その他

新聞	記事タイトル名	掲載紙面	掲載日時
朝日新聞	生物研究コンピュータ時代－がんの原因、薬の候補を解析－	25面	2010年8月10日

研究成果の刊行物・別刷

Molecular Systems Biology 7; Article number 472; doi:10.1038/msb.2011.5
Citation: *Molecular Systems Biology* 7:472

www.molecularsystemsbiology.com

molecular systems biology

Analysis of multiple compound–protein interactions reveals novel bioactive molecules

Hiroaki Yabuuchi[1,5], Satoshi Niijima[1,5], Hiromu Takematsu[2], Tomomi Ida[1], Takatsugu Hirokawa[3], Takafumi Hara[4], Teppei Ogawa[1], Yohsuke Minowa[1], Gozoh Tsujimoto[4] and Yasushi Okuno[1,*]

[1] Department of Systems Biosciences for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan, [2] Laboratory of Membrane Biochemistry and Biophysics, Graduate School of Biostudies, Kyoto University, Kyoto, Japan, [3] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan and [4] Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan
[5] These authors contributed equally to this work
* Corresponding author. Department of Systems Biosciences for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan. Tel.: +81 75 753 4559; Fax: +81 75 753 4559; E-mail: okuno@pharm.kyoto-u.ac.jp

Received 21.7.10; accepted 20.1.11

The discovery of novel bioactive molecules advances our systems-level understanding of biological processes and is crucial for innovation in drug development. For this purpose, the emerging field of chemical genomics is currently focused on accumulating large assay data sets describing compound–protein interactions (CPIs). Although new target proteins for known drugs have recently been identified through mining of CPI databases, using these resources to identify novel ligands remains unexplored. Herein, we demonstrate that machine learning of multiple CPIs can not only assess drug polypharmacology but can also efficiently identify novel bioactive scaffold-hopping compounds. Through a machine-learning technique that uses multiple CPIs, we have successfully identified novel lead compounds for two pharmaceutically important protein families, G-protein-coupled receptors and protein kinases. These novel compounds were not identified by existing computational ligand-screening methods in comparative studies. The results of this study indicate that data derived from chemical genomics can be highly useful for exploring chemical space, and this systems biology perspective could accelerate drug discovery processes.
Molecular Systems Biology **7:** 472; published online 1 March 2011; doi:10.1038/msb.2011.5
Subject Categories: bioinformatics; computational methods
Keywords: chemical genomics; data mining; drug discovery; ligand screening; systems chemical biology

Introduction

Experimental perturbations of biological systems, such as genetic mutation and chemical exposure, have been used as powerful approaches to deepen our systems-level understanding of biological processes and to discover unprecedented biological phenomena (Lehár *et al*, 2008). In particular, perturbations by chemical probes provide broader applications not only for analysis of complex systems but also for intentional manipulations of these systems, e.g., a medicine is a small molecule designed for the purpose of clinical therapy that can actively manipulate biological systems from disordered to well-ordered states. Unfortunately, the number of well-characterized chemical probes is highly limited, which has bottlenecked their wide range of application.

The set of all possible small organic molecules, referred to as chemical space, has been estimated to consist of more than 10^{60} compounds (Dobson, 2004). Chemical space is as vast as the diversity of biological systems, and the vastness of the two domains creates difficulty in comprehensive understanding of the interface between chemical space and biological systems (Lipinski and Hopkins, 2004; Renner *et al*, 2009). Recently, chemical genomics has emerged as a promising area of research applicable to exploration of novel bioactive molecules, and researchers are currently striving toward the identification of all possible ligands for all target protein families (Wang *et al*, 2009). Large-scale data sets of compound–protein interactions (CPIs) are being collected, and chemical genomics studies have shown that patterns of protein–ligand interactions are too diverse to be understood as simple one-to-one events. For example, multiple structurally different compounds have been shown to bind the same protein or express similar biological activities (Eckert and Bajorath, 2007; Young *et al*, 2008). In other cases, one drug has been shown to affect multiple targets from different protein families (MacDonald *et al*, 2006; Paolini *et al*, 2006). This phenomenon, termed polypharmacology, is thought to be one critical cause of adverse drug effects (Hopkins, 2008). There-

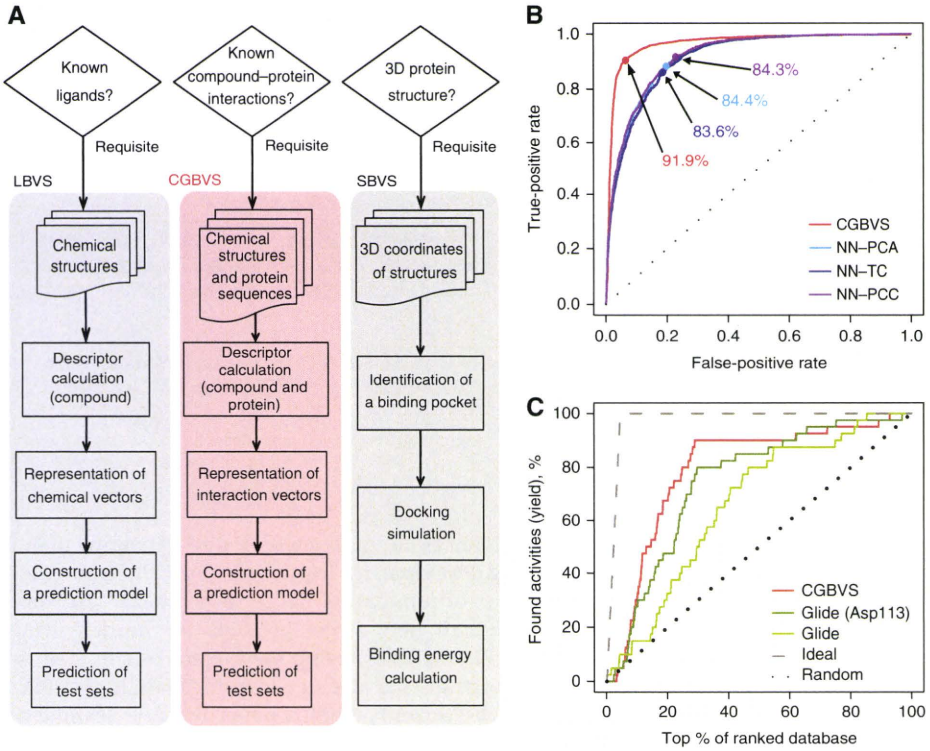


Figure 1 Overview and performance of CGBVS. (**A**) Comparison of the strategies used for CGBVS, LBVS, and SBVS. CGBVS used multiple CPI data, represented the CPI in vector form, and used SVM for CPI pattern learning. (**B**) ROC curves obtained by fivefold cross-validation using compound–GPCR interactions for the CGBVS (red) and LBVS methods. The best accuracy rate for each method is also shown. NN–PCA (light blue), nearest neighbor (NN) method with the Pearson correlation coefficient in the constructed space using principal component analysis (PCA); NN–TC (navy), NN method with the Tanimoto coefficient (TC); and NN–PCC (purple), NN method with the Pearson correlation coefficient (PCC) in the original space. (**C**) Enrichment curves obtained by the CGBVS and SBVS methods. Glide was used both without constraints and with constrained hydrogen bonding between the compounds and Asp113, a residue known to be crucial for ligand binding to ADRB2 (Strader *et al*, 1987). Information regarding interactions between ADRB2 and its ligands was not used in test set for machine learning for CGBVS.

fore, an integrative understanding of multiple interactions among chemical and biological components beyond a one-compound/one-target simplification could open up new opportunities in drug development, but the need to develop appropriate data mining methods for characterizing and visualizing the full complexity of interactions between chemical space and biological systems is urgent (Oprea *et al*, 2007). Recently, mining of multiple CPI data sets has been used to identify new protein targets for known drugs and thereby predict unreported polypharmacology (Keiser *et al*, 2009). However, this approach only identifies additional targets for known drugs. No existing screening approach has so far succeeded in identifying novel bioactive compounds using multiple interactions among compounds and target proteins, and the potential application of analysis of multiple CPIs to identify novel bioactive molecules remains unknown.

High-throughput screening (HTS) and computational screening have greatly aided in the identification of early lead compounds for drug discovery. However, the large numbers of assays required for HTS to identify drugs that target multiple proteins render this process very costly and time-consuming. Therefore, interest in using *in silico* strategies for screening has been increasing. The most common computational approaches, ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS; Oprea and Matter, 2004; Muegge and Oloff, 2006; McInnes, 2007; Figure 1A), have been used for practical drug development. Unfortunately, these methods have important limitations. LBVS aims to identify molecules that are very similar to known active molecules and generally has difficulty identifying compounds with novel structural scaffolds that differ from reference molecules. Attempts to scaffold-hop using LBVS are prone to identification of increased numbers of false positives (Eckert and Bajorath, 2007). Therefore, the primary objective of virtual screening, reduction of the number of candidate compounds to be assayed, remains unachievable using this method. The other popular strategy, SBVS, is constrained by the number of three-dimensional crystallographic structures available and, more importantly, by the difficulty of accurately simulating molecular docking processes for targets, including membrane-spanning G-protein-coupled receptors (GPCRs). To circumvent these limitations, we have shown that a new computational screening strategy, chemical genomics-based virtual screening (CGBVS), has the potential to identify novel, scaffold-hopping compounds and assess their polypharmacology by using a machine-learning method to recognize conserved molecular patterns in comprehensive CPI data sets.

Results

Theoretical framework for CGBVS

The CGBVS strategy is made up of five steps: CPI data collection, descriptor calculation, representation of interaction vectors, predictive model construction using training data sets, and predictions from test data (Figure 1A and Supplementary Figure S1). Importantly, step 1, the construction of a data set of chemical structures and protein sequences for known CPIs, does not require the three-dimensional protein structures needed for SBVS. We chose GPCRs, important pharmaceutical targets (Hopkins and Groom, 2002), as our first target proteins for virtual ligand screening. In total, 5207 CPIs (including 317 GPCRs and 866 ligands) retrieved from the GLIDA database (Okuno *et al*, 2006) were used as experimental data (Supplementary Table S1). In step 2, compound structures and protein sequences were converted into numerical descriptors using 929-dimensional and 400-dimensional feature vectors, respectively. A wide variety of chemical descriptors was used to describe the substructures, as well as the physicochemical and molecular properties of the small molecules. Descriptors for protein sequences were created using a string kernel (see Materials and methods section and Supplementary information for details). These descriptors were used to construct chemical or biological spaces, in which decreasing distance between vectors corresponded to increasing similarity of compound structures or protein sequences. In step 3, we represented multiple CPI patterns by concatenating these chemical and protein descriptors (in 929 + 400 dimensions). Using these interaction vectors, we could quantify the similarity of molecular interactions for compound–protein pairs, despite the fact that the ligand and protein similarity maps differed substantially (Keiser *et al*, 2007). In step 4, concatenated vectors for CPI pairs (positive samples) and non-interacting pairs (negative samples) were input into a support vector machine (SVM; Vapnik, 1995), an established machine-learning technique widely applied to pattern-recognition problems (Schölkopf *et al*, 2004; Shawe-Taylor and Cristianini, 2004). Using training sets, an SVM classifier was generated as a hyperplane dividing positive and negative samples into two distinct classes representing interaction and non-interaction. By mapping the samples into high-dimensional feature space using a nonlinear kernel function, samples that were linearly inseparable in the original input space could be linearly separated in the feature space. As a nonlinear SVM can extract patterns from data sets with nonlinear characteristics, non-intuitive interaction rules can be obtained from multiple CPI patterns, creating the potential to identify novel CPIs. In the final step, the SVM classifier constructed using training sets was applied to test data. Along with providing simple yes/no outputs, the calculated prediction scores also ranked all test compound–GPCR pairs in the order of binding probability.

Computational evaluation of CGBVS

To evaluate the predictive value of CGBVS, we compared its performance with that of LBVS methodologies using respective data sets of 5207 interacting and non-interacting pairs. The performance of each method was tested by repeating fivefold cross-validations 20 times. CGBVS performed with a considerably higher accuracy (91.9 ± 0.3%) than LBVS (84.4 ± 0.3%, at best). We also recorded the number of true-positive interactions as a function of false positives and plotted receiver operating characteristic (ROC) curves (Hanley and McNeil, 1982), as shown in Figure 1B and Supplementary Table S2. ROC analysis revealed that CGBVS performed better than all LBVS methods in terms of the score ranking of CPI pairs.

Recently, the crystal structure of the β2-adrenergic receptor (ADRB2) has been determined (Cherezov *et al*, 2007; Rasmussen *et al*, 2007). Therefore, we were able to compare CGBVS and SBVS in a retrospective virtual screening based on the human ADRB2 using a representative docking program, Glide (Friesner *et al*, 2004). For CGBVS, we constructed a predictive model based on 5167 CPI pairs, excluding 40 known ADRB2-related CPIs to avoid any bias in favor of CGBVS during the machine-learning step. Using both methods, we predicted scores for the same 866 known GPCR ligands, including the 40 known ADRB2 ligands as positive controls. We asked whether the scores for these 40 known positive compounds were higher than scores for other compounds. Figure 1C and Supplementary Table S3 show that CGBVS provided higher enrichment factors (EFs) and hit rates than did SBVS. These results suggest that CGBVS is more successful than conventional approaches for prediction of CPIs.

Polypharmacological interactions for ADRB2

We also evaluated the ability of the CGBVS method to predict the polypharmacology of ADRB2 by attempting to identify novel ADRB2 ligands from the above ligand data set. As an established, well-studied pharmaceutical target (Waldeck, 2002), we expected that novel ADRB2 ligands would be difficult to find, and that searching for novel scaffolds for such a well-known target would be a stringent assessment of the predictive ability of CGBVS. After training an SVM classifier using all 5207 CPIs, we ranked the prediction scores for the interactions of 826 reported GPCR ligands (excluding the 40 known ADRB2 ligands) with ADRB2, and then analyzed the 50 highest-ranked compounds in greater detail. To complement the less-than-comprehensive binding data available in the original GLIDA database, a literature search was performed using SciFinder (Wagner, 2006) and PubMed (http://www.ncbi.nlm.nih.gov/pubmed/). This search identified 15 of the top 50 compounds as known ADRB2 ligands. Importantly, these compounds were identified as ligands of other GPCRs, not of ADRB2, in training sets used in the machine-learning step. ADRB2 ligands already reported in the literature were excluded from analysis, but the remainder were tested in *in vitro* binding assays. From the remaining 35 ligands, 21 were commercially available. Of these 21, 11 were not previously reported, but were discovered to bind to ADRB2 (Figure 2A and Supplementary Table S5). These compounds included ligands for the acetylcholine, serotonin, dopamine, and neuropeptide Y receptors (Figure 2E), indicating the presence of potential polypharmacological interactions with ADRB2.

To substantiate the novelty of the ligands identified by CGBVS, we compared the predictive scores estimated by the three virtual screening approaches (CGBVS, LBVS, and SBVS).

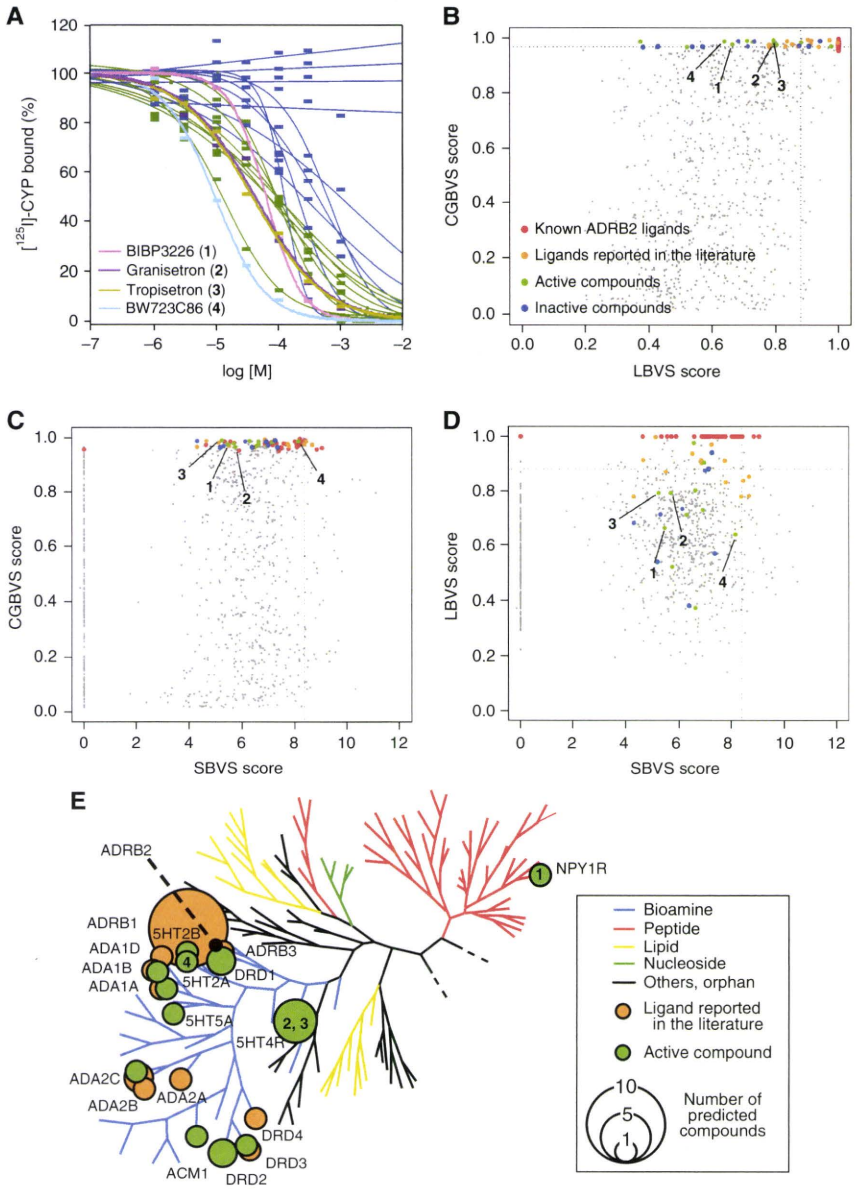


Figure 2 Prediction of ADRB2 ligands. (**A**) Binding curves for 21 of the top 50 compounds ranked by CGBVS available commercially. Pink, BIBP3226 (**1**); purple, granisetron (**2**); dark yellow, tropisetron (**3**); light blue, BW723C86 (**4**) (Supplementary Table S5); green, 11 active compounds screened ($K_i < 100\ \mu M$); and blue, inactive compounds. (**B–D**) ADRB2 ligand prediction using CGBVS, LBVS (NN-PCA), and SBVS (GlideScore, Asp113). The NN–PCA score indicates the Pearson correlation coefficient between each compound and its nearest ligands in principal component space. Red, ADRB2 ligands; orange, ligands reported in the literature; green, screened active compounds; and blue, inactive compounds. The numbers **1–4** are corresponding to BIBP3226 (**1**), granisetron (**2**), tropisetron (**3**), BW723C86 (**4**), respectively. Dotted lines show the score of the fiftieth-ranked compound using each method. (**E**) Polypharmacological relationships of newly identified ADRB2 ligands. Newly identified (green) and known (orange) ADRB2 ligands were mapped on a human GPCR phylogenetic tree (Fredriksson *et al*, 2003). The size of the circles indicates the number of compounds reported to bind to each GPCR on the tree. Target GPCRs of the four compounds **1–4** are shown in the circles. 5HT2A, 5-hydroxytryptamine receptor 2A; 5HT2B, 5-hydroxytryptamine receptor 2B; 5HT4R, 5-hydroxytryptamine receptor 4; 5HT5A, 5-hydroxytryptamine receptor 5A; ACM1, acetylcholine receptor M1; ADA1A, alpha 1-adrenergic receptor type A; ADA1B, alpha 1-adrenergic receptor type B; ADA1D, alpha 1-adrenergic receptor type D; ADA2A, alpha 2-adrenergic receptor type A; ADA2B, alpha 2-adrenergic receptor type B; ADA2C, alpha 2-adrenergic receptor type C; ADRB1, beta 1-adrenergic receptor; ADRB3, beta 3-adrenergic receptor; DRD1, dopamine receptor D1; DRD2, dopamine receptor D2; DRD3, dopamine receptor D3; DRD4, dopamine receptor D4.

We used 26 (orange and green dots in Figure 2B–D) of the top 50 compounds predicted by CGBVS that had ligand activity, but that had not been used in the training data set, to evaluate the other methods. As shown in Figure 2B and D, the known ADRB2 ligands (orange dots) clustered with the highest scores on the LBVS axis, whereas most of the other active compounds identified (green dots) fell outside the top 50 scores. Active compounds (red, orange, and green) were widely scattered along the SBVS axis (Figure 2C and D), and only six were found in the top 50 SBVS scores. This was consistent with the known