—Reviews—

バイオインフォマティクスとケモインフォマティクスの融合による インシリコ創薬研究

奥野恭史

In silico Drug Discovery Based on the Integration of Bioinformatics and Chemoinformatics

Yasushi OKUNO

Department of PharmacoInformatics, Centre for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University, 46–29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606–8501, Japan

(Received May 28, 2008)

With the near completion of the human genome sequencing, bioinformatics and chemoinformatics are expected as promising tools in genome-based drug discovery. The emerging field of chemical genomics is accumulating large-scale assay data on compound-protein interactions. We are now developing new mining methods for the chemical genomics data based on the integration of bioinformatics and chemoinformatics. Here we present a GPCR-ligand database (GLIDA) and a novel *in silico* screening method, which we have developed. GLIDA is a novel public GPCR-related chemical genomics database that is primarily focused on the correlation of information between GPCRs and their ligands. Our *in silico* screening method is based on statistical machine learning of the conserved patterns of molecular recognition extracted from comprehensive compound-protein interaction data. These are promising approaches to accelerating drug discovery processes.

Key words——chemoinformatics; bioinformatics; *in silico* screening

1. はじめに

ヒトゲノムが解読された今日，莫大なゲノム情報から創薬への手掛かりを発見すること，すなわち「ゲノム創薬」に大きな期待が寄せられている．ゲノム創薬は，ゲノム情報を出発点とし創薬の標的遺伝子探索からリード化合物探索を経て臨床段階に至る広範で高度に専門化した複合領域であり，その実践にはこれらの複合領域の橋渡しを実現する統合的なインフォマティクス基盤「創薬インフォマティクス」が必須となる．われわれは，創薬インフォマティクスという新たな研究分野の創成に向け，バイオ情報を扱うバイオインフォマティクスとケミカル情報を扱うケモインフォマティクスという独立に発展してきた2つの情報科学分野の統合を図り，バイオ情報とケミカル情報の両者を同時に統合的にマイニングする新しい情報技術の開発に着手している．なお本研究は，現在，国内外で注目されているケミカルゲノミクス・ケミカルバイオロジーのための有力な情報基盤ともなり得るものと考えられる．

京都大学大学院薬学研究科統合薬学フロンティア教育センター（〒606–8501 京都市左京区吉田下阿達町 46–29）
e-mail: okuno@pharm.kyoto-u.ac.jp

本総説は，平成20年度日本薬学会奨励賞の受賞を記念して記述したものである．

2. ケミカル空間とケミカルゲノミクス

2004年12月のNature誌において，Chemical Space特集号が発表された.[1] そこでは，化合物の種類は10^{60}個を超える天文学的なバリエーションを有しており，化合物空間を探索することは宇宙探索と同様に壮大な課題であることが提示されている．このことは医薬品の候補化合物となり得る新規な活性化合物を見つけ出すことがいかに困難でセレンディップなことであるかを示唆するものである．

これらケミカル空間の探索の基礎研究としてケミカルゲノミクス・ケミカルバイオロジー研究が近年注目されている．ケミカルゲノミクスでは，その命題として「莫大な数の化合物と生体系（タンパク質や細胞など）との相互作用を包括的に明らかにする

こと」が挙げられている．実際，米国では，ケミカルゲノミクスプロジェクトを掲げ，数百万もの膨大な化合物に関する情報を収集し，有用化合物の探索に国策として取り組んでいる．

しかしながら，広大な化合物空間から生物活性を有する化合物を探し当てる化合物探索には，天文学的な数量に対応できる新たなインフォマティクス技術とハイスループット技術の研究開発が必須である．そこで，われわれは，莫大な化合物群とタンパク質群との相互作用様式をゲノムスケールで解析することを目的とした情報学的技術，すなわちケミカルゲノミクスのためのインフォマティクス技術の研究開発を行っている．

3. ケモインフォマティクスとバイオインフォマティクス

ケミカルゲノミクス・ケミカルバイオロジーでは，化合物のケミカル情報と生体系のバイオ情報の2種の異なる情報が対象となる．したがって，ケミカルゲノミクスのための情報処理技術には，ケミカル情報を処理するケモインフォマティクスとバイオ情報を処理するバイオインフォマティクスを融合する新たなインフォマティクス技術の開発が必須となる．しかしながら，化学と生物学という異なる分野を背景に持つ2つのインフォマティクスは，独立して発展してきており現状では互いに相容れない．そこで，われわれはケモインフォマティクスとバイオインフォマティクスにおける方法論的なアナロジーに着目しその融合を図った．すなわち，ケモインフォマティクスもバイオインフォマティクスもともに，個体（化合物やタンパク質）の特徴量を数値やベクトルで表現することにより，各個体の相対的な特性の違いを探索空間上の個体間の距離として定量的に算出する方法論を基本としている．例えば，ケモインフォマティクスでは，データベースに集積された膨大な化合物エントリーは化学構造や特性を定量的に表すベクトルとして表現され，その相対的な違いを距離の尺度として持つ座標空間（探索空間）をコンピュータ内部に構築する．データベース検索はこの探索空間において距離が近接する化合物を類似化合物として選出してくることになる．また，バイオインフォマティクスでも同様の考え方であり，遺伝子・タンパク質エントリーは配列や構造として表現され，それぞれの相同性（類似度）を尺度として持つ探索空間（バイオデータの場合，探索空間は系統樹やネットワーク構造になっている場合もある）が構築され，データベース検索にはこの探索空間に基づき，類似（類縁）遺伝子・タンパク質が選出される．

一方，ケミカルゲノミクスとは，ケミカル空間の個体（化合物）とバイオ空間の個体（遺伝子・タンパク質）との相互作用関係を網羅的に明らかにする研究であり，Fig. 1の赤線に示す対応関係を付加したモデルであると考えられる．ここで，われわれは，ケミカル情報とバイオ情報を統合的に処理するために，ケミカル空間（緑色）とバイオ空間（黄色）を独立して扱うのではなく，2つの空間を融合したモデルをケミカルゲノミクスのためのインフォマティクスモデルとして考案した．

4. ケミカル空間とバイオ空間の融合モデル

情報科学的アプローチによる化合物探索は，これまで化合物のケミカル情報のみを用いたケモインフォマティクス手法が用いてきた．これに対し，われわれの手法は，このケミカル情報のみの従来手法にバイオインフォマティクス技術を融合させ，バイオ情報を考慮に入れた化合物探索を実現する新しいインフォマティクス手法と言える（Fig. 1）．

例えば，化合物について構造や特性の類似性を相対的な位置関係として表現したものをケミカル空間（赤が化合物，緑領域がケミカル空間）として定義するとともに，タンパク質についても類似関係（配列や構造の相同性）を相対的な位置関係として表現したものをバイオ空間（青がタンパク質，黄色領域がバイオ空間）として定義する．さらに個々の化合物とタンパク質の結合をリンク（黒線）することによって，これらケミカル空間とバイオ空間を融合した単純なモデルを構築できる（Fig. 2）．

ここで，標的タンパク質に作用する化合物候補を探索する *In silico* スクリーニングにこの融合モデルを適用する場合を考えると，

1） 標的タンパク質（青星）の配列構造から，そのタンパク質がバイオ空間座標にマッピングされる．

2） バイオ空間にマッピングされた標的タンパク質の近隣タンパク質からのケミカル空間へのリンク情報をたどること（青矢印）により，その標的タンパク質が関係するケミカル空間のエリア（青円内）を指定することができる．

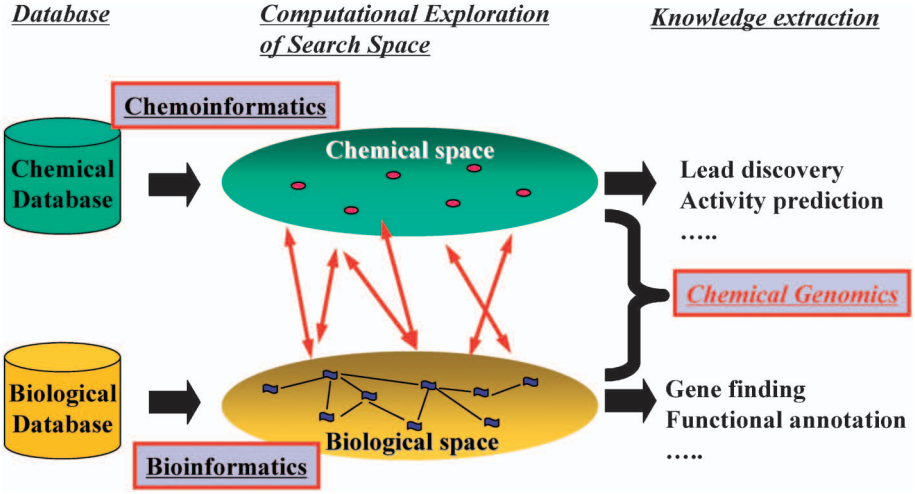

Fig. 1. Bioinformatics and Chemoinformatics

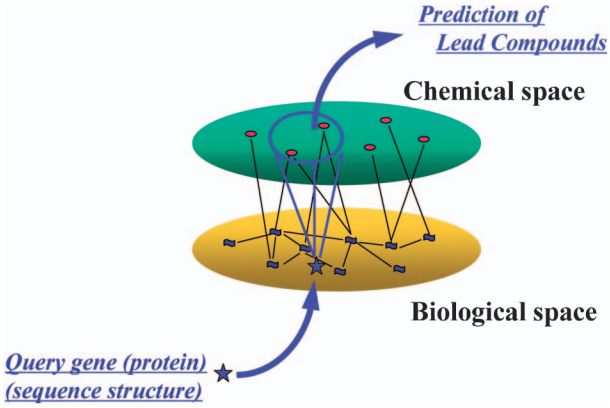

Fig. 2. *In silico* Screening for Chemical Genomics Data

3) 上記エリア内の化合物群が，標的タンパク質に相互作用する可能性のある化合物群と推定される（ここでは，類似のタンパク質は，類似の化合物を結合するという前提を基にしている）．

われわれは，このケミカル空間とバイオ空間の融合モデルを用いた探索を，GPCR ファミリーとそのリガンド化合物の探索に適用し，GLIDA データベース[2-4]として Web サービスを行っている．GLIDA は，GPCR のバイオ情報，そのリガンドのケミカル情報，及び GPCR とリガンドの相互作用情報の 3 種類の情報より構成されている．GPCR のエントリーはヒト，ラット，マウスに限定し，バイオ情報は GPCRDB から取得した．また，GPCR と結合するリガンドのエントリーとそのケミカルデータ（化学名，構造式，分子量，MDL Mol ファイルなど）は IUPHAR Receptor Database, PubMed, PubChem 及び MDL ISIS/Base 2.5 などの公共又は商用のデータベースから取得した．具体的には，2008 年 1 月現在，24077 件のリガンドエントリー，及び 39140 件のリガンド–GPCR の相互作用エントリーの登録に至っている．

各エントリーの検索は，GPCR（またはリガンド）のキーワード検索及びクラス分類テーブルから行うことが可能である．ここで，GPCR 分類は，GPCRDB に定義されている進化系統樹由来の分類に従っている．またリガンド分類は，KEGG で定義されている原子タイプの原子数／結合数に基づいた頻度プロファイルから距離行列を計算し，主成分分析（PCA）に基づいて GLIDA 独自のリガンド分類を行っている．検索された各 GPCR（またはリガンド）のページには，バイオ情報（またはケミカル情報），及びそれらに結合するリガンド（または GPCR）のリストが同時に表示される．さらに，GLIDA の GPCR（またはリガンド）のページは GPCR–リガンド相互作用の解析機能を有している．すなわち，検索された GPCR（またはリガンド）と最も高い類似性を持つ 25 個の GPCR（またはリガンド）リストを表示するとともに，これら 25 個のエントリーと結合するリガンド（または GPCR）との相互作用様式を 2 次元マップ表示する．このマップの 2 軸に並ぶ GPCR とリガンドの順番は，各々 GPCR とリガンドのクラスタリング結果を反映している．したがって，GPCR，リガンドの類似性情報と相互作用情報を同時に視覚化し，このパターンを分析して GPCR とリガンドの相互

作用予測を実現し，薬物と作用基点の相互作用に関する情報を得ることができる（Fig. 3, Fig. 4）.

5. ケミカルゲノミクスに基づくバーチャルスクリーニング

活性化合物を効率よく迅速に探し出すために，計算機を用いた候補化合物の絞り込み手法「バーチャルスクリーニング（VS)」が開発されてきた．現在よく用いられている VS として，既知リガンドとの構造類似性に基づく「Ligand-based virtual screening（LBVS)」と標的タンパク質の立体構造に基づく「Structure-based virtual screening（SBVS）」がある.[5,6] この 2 つの手法は，近年の情報技術の進歩と相まって，この 10–20 年で著しい発展を遂げ，ゆるぎない地位を確立した．しかしながら現在，VS のヒット確率は 1–10％もあればよしとされており（例えば，LBVS では，既知活性化合物の骨格構造に強く影響される嫌いがあるし，SBVS では，パラメータの恣意性，予測的中率の低さなどが指摘されている），さらなる技術的な改良や革新的技術の開発が切望されていることは間違いない.

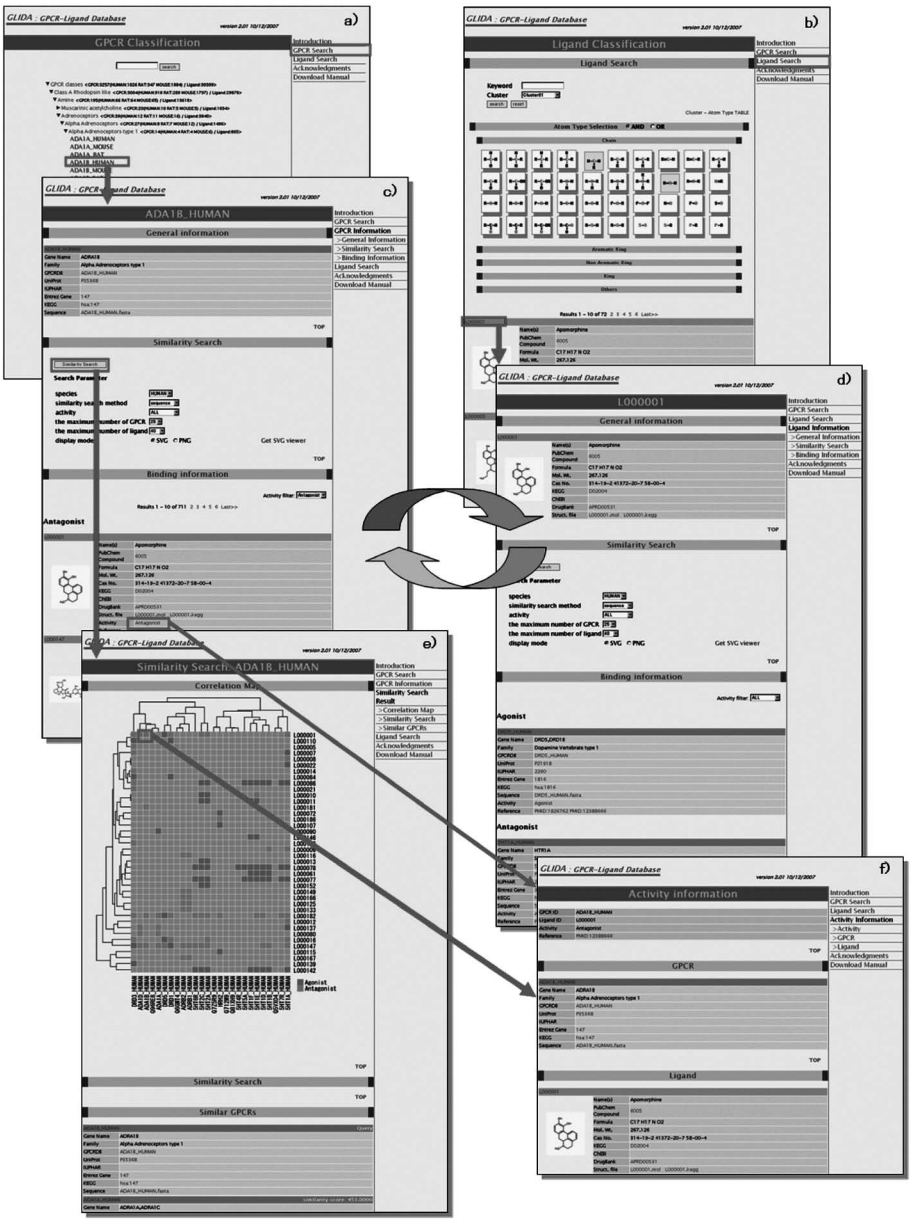

Fig. 3. A Screenshot of GLIDA Showing Linked Relations among Search Pages (a, b), Result Pages (c, d), an Analytical Report Page (e), and a Binding Information Page (f)

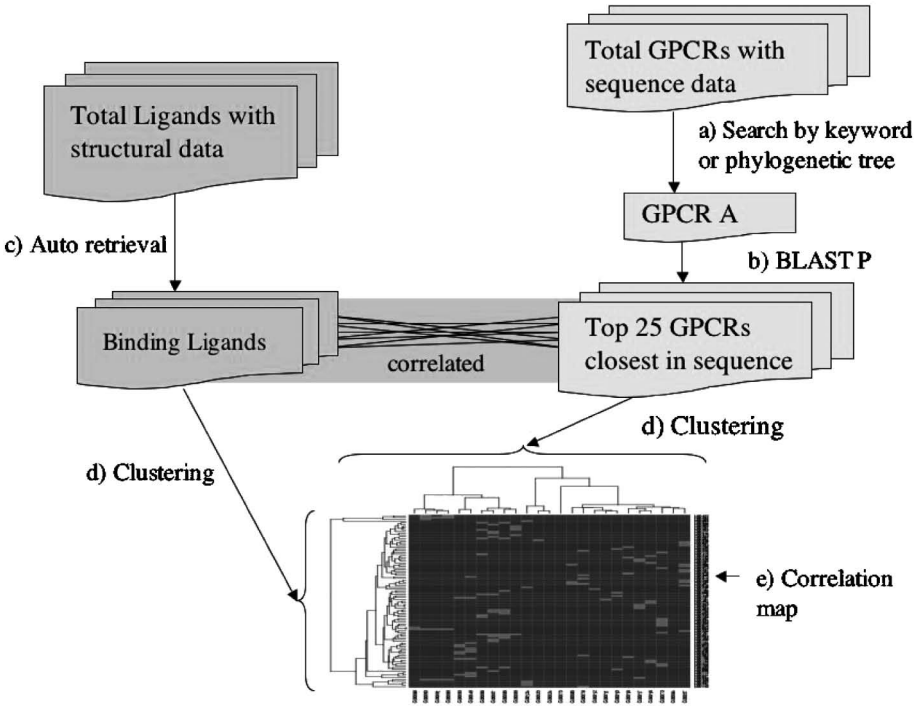

Fig. 4. A Schematic Example of the Search and Analysis Process Showing GPCR-ligand Correlations Produced from a GPCR Query Using GLIDA

ここでは，LBVS や SBVS とは概念の異なる第 3 の VS として，われわれが開発している「ケミカルゲノミクスに基づく VS 手法（Chemical Genomics-based virtual screening; CGBVS)」を紹介する．「ケミカルゲノミクス」は，興味を持つ化合物が生物に与える影響についてゲノム規模で研究する学問であり，[7,8] マイクロアレイやハイスループットスクリーニングなどの同時大量解析技術の革新に後押しされ，近年，注目を浴び始めている．それに伴い，化合物と遺伝子の関連性について，膨大な実験データが蓄積されている．そこでわれわれは，情報科学技術の 1 つであるパターン認識技術を用いて，タンパク質と化合物との結合情報（ケミカルゲノミクス情報）から抽出したタンパク質のリガンド認識パターンに基づいて活性化合物を効率的に発見する新たな VS，「CGBVS」を開発している．

タンパク質とリガンドとの相互作用パターンの認識とそのリガンド予測を開発するために，われわれは学習アルゴリズムの一種であるサポートベクターマシン（SVM)[9] を用いた．SVM は，2 クラス分類器の一種であり，与えられた 2 つのグループに属する特徴ベクトルを最大マージンで分離するような超平面を構築する．ここで，最大マージンとは，分離した超平面から各サンプル間までの最短距離を指す．

われわれは，この SVM を用いて，化合物–タンパク質相互作用の有／無を判別する手法を開発した．その手法の流れを Fig. 5 に示す．まず，収集した相互作用をベクトルとして表現するために，各化合物の化学構造，各タンパク質のアミノ酸配列について，様々な属性（記述子と呼ぶ）を計算する．次に，正例（相互作用する化合物–タンパク質ペア）及び負例（相互作用しない化合物–タンパク質ペア）に対応する記述子をそれぞれ組み合わせて特徴ベクトルを構成し，SVM を用いて学習モデルを構築する．このモデルが得られると，（未知の化合物–タンパク質ペアに相当する）新しいベクトルが相互作用有／無 のどちらのクラスに属するかを予測することができる．

既存の VS 手法との比較検討を行うため，今回開発した CGBVS と LBVS との予測性能を比較した．収集した化合物–GPCR 相互作用を用いて，負例を交換しながら 5 分割交差検定法（5 fold cross-validation）を試行した．交差検定の結果，最近傍法を用いた LBVS では 84.4±0.3%，CGBVS では 91.6±0.2% の相互作用を正しく予測した．また，ROC 曲線からも，CGBVS の予測性能の高さが確

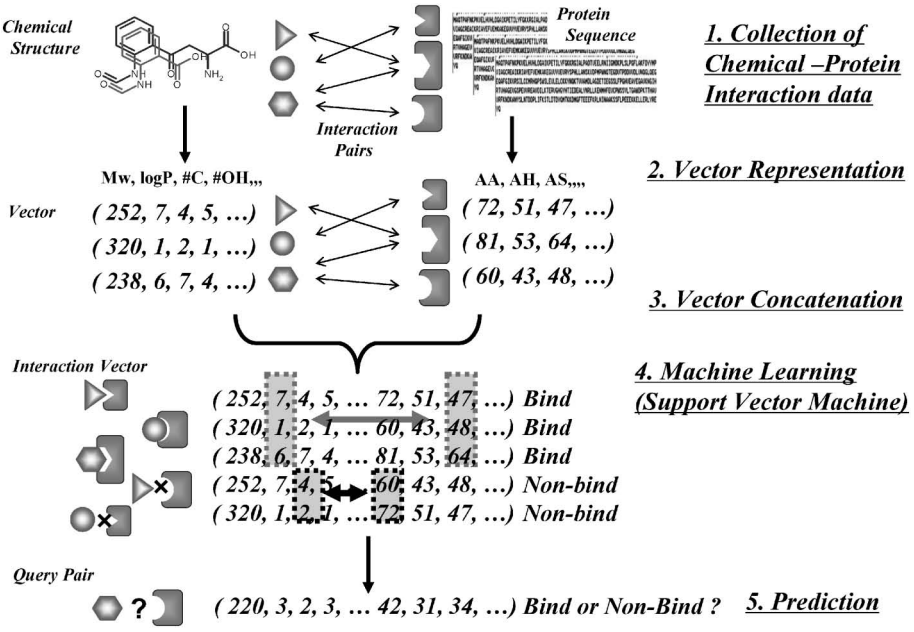

Fig. 5. Overview of Chemical Genomics-based Virtual Screening

認された（Fig. 6）．したがって，ケミカルゲノミクス情報の活用がリガンド予測性能の向上につながったといえる．

さらに，ヒト β2 アドレナリン受容体（β2AR）を標的 GPCR とし，構築した学習モデルを用いてリガンド予測を行い，*in vitro* 実験による検証を行った．リガンド予測の対象化合物は，β2AR 以外の GPCR のリガンドとして知られている 826 化合物とした．ここで，CGBVS により予測された β2AR リガンド候補上位 50 の化合物のうち，文献・特許調査により 14 種の化合物について β2AR との相互作用に関する報告を確認した．さらに，残りの相互作用不明な化合物のうち，入手可能な 21 種類について *in vitro* 結合阻害実験を行ったところ，17 種類の化合物が相互作用（10^{-5} M＜IC_{50}＜10^{-3} M）を示した．結合阻害実験のヒット率は 81％（17/21）に上り，ここにおいても高い予測的中率が示された．また今回新たに発見した化合物の化学構造を精査したところ，典型的な β2AR 作動薬の構造（カテコラミン骨格，イソプレナリン誘導体）及び β2AR 拮抗薬の構造（アリルアルキルアミン誘導体）とは異なる多様な骨格を含んでおり，化合物の構造類似性に基づく従来の方法では発見が困難なリガンド群が含まれることが明らかになった．したがって，ケミカルゲノミクス情報が，リガンド予測精度の向上

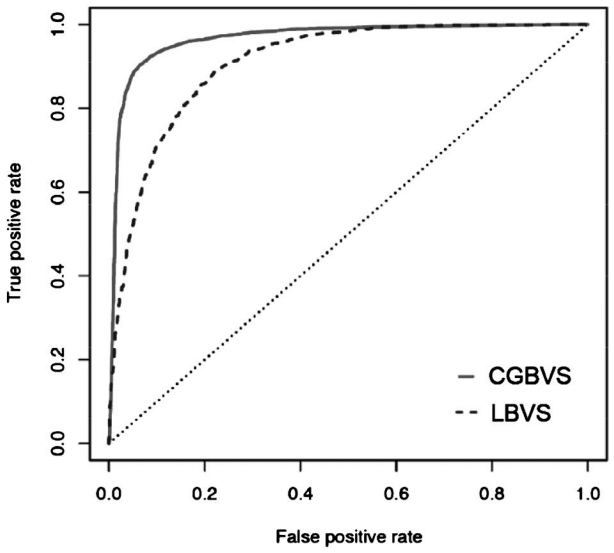

Fig. 6. Comparison of Prediction Performance between Chemical Genomics-based and Ligand-based Virtual Screenings

のみならず，新規骨格を持つリガンドの検出にも有用であることが示唆された．

6. おわりに

従来，生物活性を示す化合物の探索には，化合物の構造類似性がその指標とされてきた．しかし，化合物の構造類似性だけでタンパク質との多種多様な相互作用を見い出すには限界があり，新しい探索手法の開発が望まれていた．われわれは，「ケミカル

ゲノミクス情報の中から相互作用パターンを抽出する」という新しいアプローチを試み，1）化合物–GPCR 質間相互作用データベース GLIDA を開発し，2）学習アルゴリズムを用いた予測モデルにより，GPCR リガンドを精度よく予測できることを実証した．

なお，β2AR の立体構造が 2007 年決定された[10]ことを受けて，われわれは SBVS との予測性能比較も行っており，その結果，CGBVS が SBVS よりも高精度で β2AR リガンドを予測できることを確認している．

われわれが開発している CGBVS は，標的タンパク質のリガンド情報がないケースにも適用可能であり，実際に現在，オーファン GPCR を標的としたリガンド予測を行っている．タンパク質の立体構造情報を必要としない点は，詳細な 3D 構造が不明な GPCR にとって，魅力的なリガンド予測手法といえる．将来的には，結合阻害定数（Ki 値）など定量的な活性データから回帰分析を行うことにより，相互作用の強さを反映させたリガンド予測手法への拡張を行う予定である．ケミカルゲノミクス情報を利用するという新規なアプローチは，GPCR に限らずすべてのタンパク質に適用可能であり，化合物–タンパク質相互作用の理解，ひいては創薬プロセスの効率化につながることが期待される．

謝辞 本研究の一部は，文部科学省，経済産業省（NEDO 若手グラント），厚生労働省の助成金支援によって行われており，深く感謝申し上げる．本総説は，平成 20 年度日本薬学会奨励賞の受賞を記念して記述したものであり，日本薬学会役員，審査員の先生方をはじめご関係の皆様に心より感謝申し上げる．また，これまでご指導，ご鞭撻賜りました学生時代の恩師・杉浦幸雄先生（現同志社女子大学教授），研究員時代の恩師・金久 實先生（京都大学化学研究所教授），助手時代の恩師・辻本豪三先生（京都大学薬学研究科教授），並びに京都大学薬学研究科研究科長・藤井信孝先生に謹んで感謝の意を表します．

REFERENCES

1） *Nature*, **432** (7019) (Insight), 823–865 (2004).
2） Okuno Y., Yang J., Taneishi K., Yabuuchi H., Tsujimoto G., *Nucleic Acids Res.*, **34**, D673–677 (2006).
3） Okuno Y., Tamon A., Yabuuchi H., Niijima S., Minowa Y., Tonomura K., Kunimoto R., Feng C., *Nucleic Acids Res.*, **36**, D907–D912 (2008).
4） 〈http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/〉
5） Muegge I., Oloff S., *Drug Discov. Today Technol.*, **3**, 405–411 (2006).
6） Oprea T. I., Matter H., *Curr. Opin. Chem. Biol.*, **8**, 49–58 (2004).
7） MacBeath G., *Genome Biol.*, **2**, COMMENT2005 (2001).
8） Salemme F. R., *Pharmacogenomics.*, **3**, 257–267 (2003).
9） Vapnik V. N., ‘‘The Nature of Statistical Learning Theory,’’ Springer, New York, 1995.
10） Cerezov V., *Science.*, **318**, 1258–1265 (2007).