厚生労働科学研究研究費補助金

萌芽的先端医療技術推進研究事業

トキシコゲノミクスのための遺伝子ネットワーク解析法の開発

平成１７年度　総括研究報告書

主任研究者　奥野　恭史

平成１８（２００６）年　４月

目　　次

厚生労働科学研究費補助金（萌芽的先端医療技術推進研究事業）
総括研究報告書

トキシコゲノミクスのための遺伝子ネットワーク解析法の開発

主任研究者　奥野　恭史　京都大学薬学研究科 助教授

研究要旨
本研究は、化合物による生体系への影響を、薬物作用遺伝子群や毒性関連遺伝子群の遺伝子発現ネットワーク（分子ネットワーク）の変動として解析する高精度な薬物安全評価アルゴリズムの開発と実用化を目的としている。すなわち、化合物を作用させた各種細胞のDNA マイクロアレイ実験による網羅的遺伝子発現データから、バイオインフォマティクス手法によって薬物毒性特有の遺伝子発現ネットワークを構築し、薬物毒性を反映する遺伝子ネットワークのパターンとして薬物安全性を評価するトキシコゲノミクス計算法の確立を目指す。一般に遺伝子発現ネットワーク解析手法は、生命現象と分子メカニズムを繋げる有力なバイオインフォマティクス手法であり、本手法をトキシコゲノミクスへ適用することにより毒性と遺伝子発現パターンの相関を鮮明にし、予測精度の向上と毒性分子メカニズムの解明という成果をもたらすものと予想される。近年、マイクロアレイ解析のトキシコゲノミクスや薬理ゲノミクスへの適用に対する期待が国内外を問わず高まって来ているが、トキシコゲノミクスや薬理ゲノミクスへのネットワーク解析手法の適用・成功事例は今のところ世界的に皆無である。従って、本研究は、ネットワーク解析法のトキシコゲノミクスへの応用という試み自身の独創性を有しており、薬物の毒性評価とその毒性発現の分子メカニズムに関する知見をも同時に提供する強い特色を有している。初年度である本年の研究進捗は当初の計画通り極めて順調に進捗した。具体的には、ヒト肝細胞への糖尿病薬作用におけるマイクロアレイ実験データの収集、および薬物とタンパク質との相互作用データベースの構築はすでに完了した。また、遺伝子ネットワーク構築手法として、ベイジアンネットワーク法とグラフィカルモデリング手法の開発に着手し従来よりも高性能なプロトタイプの開発にも成功した。本研究によって開発される高精度な毒性予測手法は、医薬品開発における早期毒性予測による医薬品開発期間・コストの軽減化と、国民における医薬品使用の安全性の向上を実現するものと期待できる。

A．研究目的

ポストゲノム時代の今日、生命科学研究の関心は遺伝子・蛋白質の単一の機能解明にとどまらず、これらを統合した生命システム全体の機能解明へと移行しつつある。医薬の観点においても、化合物と直接作用する単一のタンパク（遺伝子）の機能変化からその薬理活性の全てを語り尽くそうとする従来の考え方では不十分であり、化合物と特定のタンパク質との直接的な作用が

周辺や下流の遺伝子たちにどのような影響を及ぼすのかという化合物と生体系との作用を多種多様な分子からなる生命システムの変動とみなす新しい概念の導入および解析法の開発が必須である。そこで、本研究は、薬物標的分子や毒性原因遺伝子などの単一遺伝子(タンパク質)を対象にした従来の解析手法から逸脱し、化合物による生体系への影響を薬物作用遺伝子群や毒性関連遺伝子群の遺伝子発現ネットワーク(分子ネットワーク)の変動として解析する高精度な薬物安全評価アルゴリズムの開発と実用化を目的としている。すなわち、化合物を作用させた各種細胞のDNA マイクロアレイ実験による網羅的遺伝子発現データから、バイオインフォマティクス手法によって薬物毒性特有の遺伝子発現ネットワークを構築し、薬物毒性を反映する遺伝子ネットワークのパターンとして薬物安全性を評価するトキシコゲノミクス計算法の確立を目指す。一般に遺伝子発現ネットワーク解析手法は、生命現象と分子メカニズムを繋げる有力なバイオインフォマティクス手法であり、本手法をトキシコゲノミクスへ適用することにより毒性と遺伝子発現パターンの相関を鮮明にし、予測精度の向上と毒性分子メカニズムの解明という成果をもたらすものと予想される。研究計画(3年間)としては、初年度(平成17年度)にはアルゴリズム開発のための実験データを収集し、2 年次(平成18年度)である今年にはアルゴリズムの実質的開発と追加マイクロアレイ実験を行い、最終年度(平成19年度)にはトキシコゲノミクスプロジェクト作成中データベース内の全データに対して、本手法を用いた解析を行うことにより研究成果の結実とする。本研究によって開発される高精度な毒性予測手法は、医薬品開発における早期毒性予測による医薬品開発期間・コストの軽減化と、国民における医薬品使用の安全性の向上を実現するものと期待できる。

B．研究方法

①基準細胞と基準薬物の選択

基準薬物の選定のための基盤データベースとして、GLIDAデータベースを構築し、公開している。トキシコゲノミクスのインフォマティクス展開には、薬物の作用基点となる遺伝子との相互作用様式を情報学的に処理する基盤技術の整備は必須であり、本データベース構築によりその基盤は確立された。GLIDAは、GPCRのバイオ情報、リガンドのケミカル情報、および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る。GPCRのエントリはヒト、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およびMDL ISIS/Base 2.5 などの公共または商用のデータベースから取得した。GLIDAはLAMP(Linux, Apache, MySQL & PHP)プラットフォームで制作され、現在Web公開を行っている(http://gdds.pharm.kyoto-u.ac.jp/services/glida)。

②マクロアレイ実験データの収集

[細胞培養]:10% (v/v) fetal bovine serum (GIBCO)、100 U/mLのpenicillin (GIBCO) および100 μg/mLのstreptomycin (GIBCO) 含有のDulbecco's Modified Eagle Medium (GIBCO) を用い、コラーゲンコートディッシュにて37℃、5% CO_2 の条件で細胞培養を行った。

[WST-1細胞増殖測定]:WST-1を基質としたNADH (Nicotinamide adenine dinucleotide) 還元酵

素の活性を指標に、細胞増殖を測定した。HepG2細胞をトログリタゾン(vehicle (0.1% dimethylsulfoxide), 0.1, 1, 10, 3, 30, 100 μM) で刺激し、各０，２，６，１２，２４時間後に回収した。1x10^4 cells のHepG2細胞に飽和WST-1溶液 (Roche) を加え、37℃、5 % CO_2の条件下で2時間反応させた後、還元型WST-1 (Formazan) の460nm / 650nmの吸光度をマイクロプレートリーダーで測定した。

[マイクロアレイ解析]：acid guanidiniumthiocyanate-phenol-chloroform法ならびにRneasy mini kit (Qiagen)を用いて、トログリタゾン (vehicle, 0.1, 3, 100 μM) で刺激したHepG2細胞(刺激後０，２，６，１２，２４時間に回収)からtotal RNAを抽出・精製した。これにT7- dT_{24} プライマーをアニールさせ、1st strand DNAを合成し、引き続き2nd strand DNAを合成した。この二本鎖DNAを鋳型として、T7 RNA ポリメラーゼによりビオチン標識dNTPを用いてビオチン標識antisense RNAを合成した。このantisense RNAを200塩基以下に断片化し、Human Genome U133 Plus 2.0 Array (Affymetrix)に対して、４５℃にてハイブリダイゼーションを行った。１６時間後、ストレプトアビジンーフィコエリスリンにて蛍光ラベル化を行い、Genome U133 Plus 2.0 Arrayの蛍光イメージを取得した。解析においては、まず0 hでのデータに対し、いずれかのマイクロアレイ解析データにおいて発現が２倍以上変動した遺伝子を抽出した。

[毒性遺伝子の同定]：上記抽出した遺伝子から、経時的な遺伝子発現パターンが上記WST-1細胞増殖の実験から得られた経時的な増殖阻害効果の変動パターンと正もしくは負の相関関係0.8以上を示す遺伝子のみを抽出し、主成分分析を行った。

③遺伝子ネットワーク解析法のプロトタイプ開発

Gaussian Graphical Modelの基本的なプロセスは回帰分析理論に基づく繰り返し作業である。すなわち、全ての変数の相関係数（Correlation Coefficient）行列から偏相関係数（Partial Correlation Coefficient）を導き、偏相関係数行列を求める。そして、共分散選択（Covariance Selection）法により、いくつかの偏相関係数を０におき、観測された標本相関係数行列を近似する相関縮約モデル（Reduced Model）を求める。得られた縮約モデルがデータによく適合しているかどうかを定量的に評価するための指標として、逸脱度（deviance）を用いる。逸脱度の値は小さければ小さいほど、縮約モデルがデータに適合していると判断し、予め設定した逸脱度の閾値を超えなければ、上記の探索のプロセスを繰り返し実行する。

ここでGraphical Modelingにおけるモデル選択作業において、偏相関係数行列の中にどの偏相関係数を０におくのか（探索作業）が最も重要である。更に、逸脱度の閾値が推定したネットワークの精度に直接関係する。本研究では、Gaussian Graphical Modelの基本プロセスである探索プロセスと逸脱度閾値を改良する４つのアルゴリズムを開発した。また、実際のマイクロアレイデータを用いて、これらのアルゴリズムの予測性能を評価した。具体的には、従来手法である「（１）Gaussian Graphical Modelの基本アルゴリズム」を開発するとともに、これらを改良して「（２）既知の情報を利用するアルゴリズム」、「（３）逸脱度と偏相関係数の閾値を探索プロセスの測定指標として用いるアルゴリズム」、「（４）逸脱度を最後まで検索するアルゴリズム」、と「（５）最小の逸脱度から探索するアルゴリズム」を開発した。また基本ア

ルゴリズムを含めた５つのアルゴリズムの予測率の評価はMAPKパスウェイを用いた。

C．研究結果

①基準細胞と基準薬物の選択

本研究で開発した薬物とGPCRの相互作用データベースであるGLIDAは、http://gdds.pharm.kyoto-u.ac.jp/services/glidaより公開している。各エントリの検索は、GPCR（又はリガンド）のキーワード検索およ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た。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、階層型クラスタリングを行ったGLIDA独自の分類木を作成した。検索された各GPCR（又はリガンド）のページには、バイオ情報（又はケミカル情報）、及びそれらに結合するリガンド（又はGPCR）のリストが同時に表示される。さらに、GLIDAのGPCR(又はリガンド)のページはGPCRー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（又はリガンド）と最も高い類似性を持つ２５個のGPCR（又はリガンド）リストを表示するとともに、これら２５個のエントリと結合するリガンド（又は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。

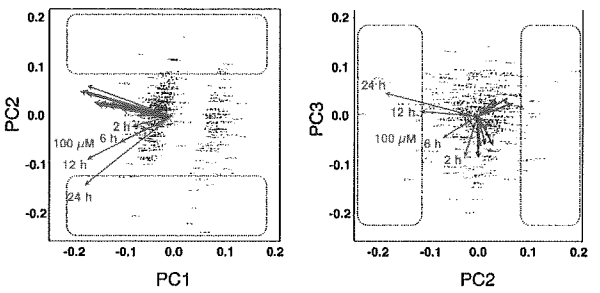
②マクロアレイ実験データの収集

肝毒性を有する薬物としてトログリタゾン、肝毒性の予測を目的としたモデルとしてヒト肝癌由来細胞株HepG2を用いて、遺伝子発現ネットワーク解析アルゴリズム開発のために必要な網羅的な遺伝子発現データの集積を行った。トログリタゾンはperoxisome-proliferator activated receptor gamma (PPARγ)のリガンドであり、チアゾリジン骨格を有するインスリン抵抗性改善剤として開発されたが、劇症肝炎などの肝障害を引き起こすとして臨床での使用が中止された薬物である。

まず、トログリタゾンのHepG2細胞に対する毒性用量を同定するため、WST-1法を用いてトログリタゾンがHepG2細胞の増殖に及ぼす影響を調べた。その結果、vehicleと比較して100 μM刺激群に置いて、刺激後6時間から増殖の抑制傾向が検出され、刺激後２４時間において最も顕著な増殖抑制が検出された。10 μM、30 μM刺激群に置いても刺激後２４時間において増殖抑制が検出されたが、それ以下の濃度ではvehicleと大きな差は検出されなかった。以上のことから、HepG2細胞に対するトログリタゾンの毒性用量は10 μM以上であることが示唆された。

次に、WST-1法を用いた増殖阻害実験の結果から毒性濃度と判断されたトログリタゾン100 μM、非毒性濃度と判断された0.1 μM及び3 μM、および溶媒である0.1% dimethylsulfoxide (vehicle) における経時的なマイクロアレイ実験およびそれらのデータ解析を行った。まず0 hでのマイクロアレイ実験データと比較して発現が２倍以上変動した遺伝子のみ抽出し、トログリタゾン刺激により発現が増減する遺伝子を明らかにしたところ、経時的な遺伝子の増減数パターンに刺激濃度によって違いが見られた。また、vehicleやトログリタゾン0.1 μM及び3 μMで刺激し

た際と比較して、毒性濃度である100 μMではより多くの遺伝子の発現が変動していた。次に、0 hでのマイクロアレイ実験データと比較して発現が２倍以上変動した遺伝子のうち、WST-1を用いた実験から得られた経時的な増殖阻害効果の変動パターンと、正もしくは負の相関関係0.8以上になる発現変動パターンを示した遺伝子のみ抽出し、主成分分析を行った（下図)。これにより３変数に情報を縮約したところ、PC2において毒性濃度である100 μMで刺激した際のベクトルとそれ以外の濃度で刺激した際のベクトルが逆の方向性を示した。さらに、トログリタゾンによる細胞毒性の亢進に伴い、ベクトルも負の方向へ伸びることから、PC2はトログリタゾンによる毒性を表していると判断された。従って、下図 において点線で囲った領域（PC2軸の両端）に存在する遺伝子はトログリタゾンによる細胞毒性に強く関与することが示唆された。

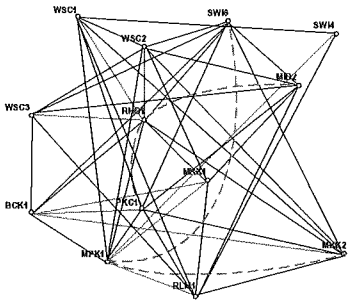


③遺伝子ネットワーク解析法のプロトタイプ開発

（１）Gaussian Graphical Modelの基本アルゴリズム

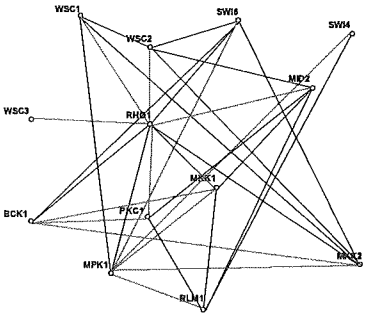
MAPKデータから、PKCパスウェイの代表的遺伝子13個（WSC1, WSC2, WSC3, MID2, RHO1, PKC1, BCK1, MKK1, MKK2, MPK1, SWI4, SWI6, and RLM1）を用いて、基本アルゴリズムの性能評価を行った。ここで、逸脱度の閾値は3.0にした。

推測したネットワークは下図に示す。PKCパスウェイに存在すべき遺伝子間の関係をカラーでハイライトしている。青い線は推定できたもので、赤い線はできなかったものである。既知の知見では存在すべきでない遺伝子間の関係をプログラムで推定したものを黒い実線で表示している。予測率はこうした赤い線の数とそうでないもの（青い線と黒い線）の数と比較した結果である。



（２）既知の情報を利用することで、既知パスウェイを保存するアルゴリズム

基本アルゴリズムの拡張として、すでに証明された遺伝子間の関係をネットワークの推定を行う前に設定することによって、存在すべき遺伝子間の関係を必ず最後の収束ネットワークに残すようにした。すなわち、基本方法（１）で推定できなかった遺伝子間の関係（（RHO1&PKC1）、（RHO1&MID2）、（MPK1&MKK2）、及び（MPK1& SWI6））をここで保留変数間関係として設定する。これらの４つの関係は探索の繰り返し作業の全てのステップに必ず削除されない。得られた結果を右図に示す。図に示すように、保存したパス以外の結果は（１）と比べて大きな変化はなかった。

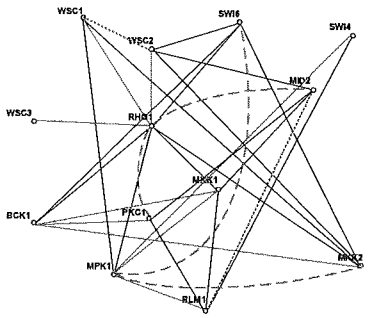


（３）逸脱度と偏相関係数の閾値を探索プロセスの測定指標として用いるアルゴリズム

逸脱度だけで遺伝子ネットワークを推定する

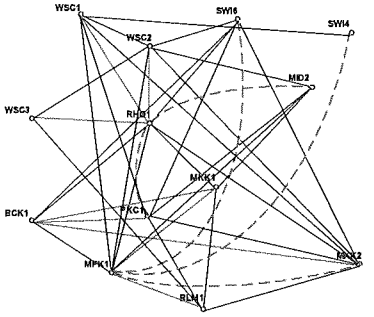
ことは不十分であると予想し、偏相関係数の閾値もGaussian Graphical Modelの探索プロセスの測定指標として用いた。すなわち、逸脱度が設定した閾値を超えても、偏相関係数が設定した閾値を超えなければ、探索プロセスを続行する。ここで注意すべきは、探索プロセスを続行させるかどうかの主な判断基準は依然として逸脱度であり、逸脱度が続行の条件に満たしている間は偏相関係数による探索は行わない。

得られた結果は下図に示す。図に示すように、基本手法（１）より、存在すべき遺伝子間の関係を推定できたことに加え、存在すべきでない関係（ピンク色の点線）を削除することができ、予測精度を高めることに成功した。



（４）局所探索により逸脱度を最後まで探索していくアルゴリズム

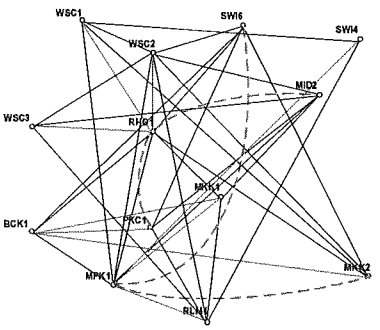
従来のGraphical modeling法では、常に最小の偏相関係数から逸脱度を求めてきた。しかし、この逸脱度は真の「最小」であると言い切ることができない。最小の偏相関係数でない係数から求めた逸脱度は最小かもしれないからである。そこで、従来通りのプロセスを先に行い、逸脱度が閾値を越える時点で探索を中止させずに、引き続き他の偏相関係数から逸脱度を求めていく。ここで、閾値より小さい逸脱度が算出された場合、繰り返し作業に戻る。残りの全ての偏相関係数を確認し、閾値より小さい逸脱度がない場合は探索を終了する。得られた結果は下図に示す。図には、本来あるべき遺伝子間の関係（MPK1&SWI4）がなくなっている。この関係は前述したアルゴリズムで正しく推定したものである。この結果から、基本アルゴリズムで削除した遺伝子間の関係が、次の逸脱度の算出にも影響を及ぼしていることがわかった。



（５）全域探索により最小値の逸脱度を探索するアルゴリズム

アルゴリズム（４）と同様に、従来の手法で最小の偏相関係数から求めた逸脱度は最小でない可能性があるため、ここで、探索の繰り返し作業の初めから、全ての偏相関係数について逸脱度を求め、真の最小逸脱度を算出する。この最小逸脱度が閾値を超えなければ、繰り返し作業を続ける。

得られた結果を下図に示す。図より、得られた遺伝子ネットワークは他のアルゴリズムでの結果と同様でないが、推定の精度はあまり上がらなかった。また、処理途中結果を確認したところ、探索の繰り返し途中の個々のアウトプットは従来法とは異なることがわかった。更に、（４）で間違って削除した遺伝子間の関係（MPK1&SWI4）は推定することができた。



D．考察

①基準細胞と基準薬物の選択

GLIDAデータベースは市販の医薬品の半分以上の標的分子となっているGPCRとそれに作用する薬物の相互作用に関する知識データベースであるとともに、その相互作用メカニズムの解明に関する知識を提供し得るケミカルゲノミクスのためのデータベースである。トキシコゲノミクスのためのインフォマティクス基盤として、薬物の作用基点となる遺伝子との相互作用様式を情報学的に処理する基盤技術の整備は必須であり、本データベース構築によりその基盤は確立された。現在は、GPCRタンパク質以外の他の標的タンパク質の情報を集積したデータベースへの拡張を行っており、このインフォマティクス基盤を用いて、基準薬物と基準細胞の選択を適宜行っていく。

②マクロアレイ実験データの収集

WST-1法を用いた実験結果から、HepG2細胞に対するトログリタゾンの毒性用量は10 μM以上であることが示唆され、これは既知の論文報告の結果とよく一致していた。主成分分析を行った結果、トログリタゾンによる毒性発現への関与が強く示唆される遺伝子群が同定された。この遺伝子群には、アポトーシスの感受性に関わる遺伝子や細胞周期に関わる遺伝子などが含まれており、興味深い。今後、現在開発中の遺伝子発現ネットワーク解析手法を用いて、トログリタゾンによる肝毒性特有の遺伝子発現ネットワークを構築し、上記の遺伝子群がどのようなネットワークを形成しているのか明らかにする予定である。また、上記の遺伝子群の遺伝子産物に対する活性化剤および抑制剤の処理や、上記の遺伝子群の遺伝子の過剰発現やsiRNAによる発現低下を行い、トログリタゾンによる細胞毒性に対する影響を調べることを計画している。

③遺伝子ネットワーク解析法のプロトタイプ開発

上述、基本アルゴリズムを含めた５つのアルゴリズムでのMAPKパスウェイの予測率を比較した。予測率が一番高いのはアルゴリズム（３）で、偏相関係数の閾値を0.25に設定した時である。全ての改良アルゴリズムは従来法である基本アルゴリズムより、高い予測率を得ることができた。結果、基本アルゴリズムを改良することが成功していると判断できる。

開発した幾つかのアルゴリズムを融合的に利用することによって、ネットワークの推定精度を上げる可能性がある。例えば、方法（２）と方法（５）を組み合わせて、結果となるグラフのEDGEの数を更に減らすことができると考えている。つまり、遺伝子ネットワークの規模に柔軟に対応したほうがよいと考えている。従って、様々なモデルの弱点を補うため、実験データのタイプに応じてモデルを組み合わせ、段階的にネットワークを同定することで、詳細な遺伝子制御ネットワークを推定することが期待できる。本アルゴリズムの精密な改良に加えるとともに、今後の課題は、ネットワーク構築が可能な遺伝子数の評価であり、これらを克服し、肝細胞のトログリタゾン投与遺伝子ネットワークの構築を実践する。

E．結論

①基準細胞と基準薬物の選択

トキシコゲノミクスのインフォマティクス基盤となる、薬物とタンパク質との相互作用データベースGLIDAを開発し、http://gdds.pharm.kyoto-u.ac.jp/services/glidaより公開した。

②マクロアレイ実験データの収集

HepG2細胞及びトログリタゾンを用いた解析から、トキシコゲノミクスのための遺伝子ネットワーク解析アルゴリズム開発を確立する上で基礎となる網羅的な遺伝子発現データを取得し、毒性発現に強く関与することが示唆される遺伝子群を同定した。

③遺伝子ネットワーク解析法のプロトタイプ開発

遺伝子ネットワーク構築手法として、ベイジアンネットワーク法とグラフィカルモデリング手法の開発に着手し従来よりも高性能なプロトタイプの開発にも成功した。

F．健康危険情報

特記事項無し

G．研究発表

1. 論文発表

1. Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H., Tsujimoto, G, "GLIDA: GPCR-Ligand database for Chemical Genomic Drug Discovery" **Nucleic Acids Research,** 34, D673-7, 2006

2. Zhu, S., Okuno, Y., Tsujimoto, G., Mamitsuka, H., "A probabilistic model for mining implicit 'chemical compound-gene' relations from literature" **Bioinformatics,** 21(s2), ii245-ii251, 2005

3. Sugimoto Y, Fukada Y, Mori D, Tanaka S, Yamane H, Okuno Y, Deai K, Tsuchiya S, Tsujimoto G, Ichikawa A., "Prostaglandin E2 Stimulates Granulocyte Colony-Stimulating Factor Production via the Prostanoid EP2 Receptor in Mouse Peritoneal Neutrophils" **J. Immunol.** 175(4), 2606-12, 2005

4. Adachi T., Okuno Y, Takenaka S, Matsuda K, Ohta N, Takashima K, Yamazaki K, Nishimura D, Miyatake K, Mori C, Tsujimoto G. "Comprehensive analysis of the effect of phytoestrogen, daidzein, on a testicular cell line, using mRNA and protein expression profile." **Food Chem. Toxicol.** 43(4): 529-35, 2005

5. Yamada M, Katsuma S, Adachi T, Hirasawa A, Shiojima S, Kadowaki T, Okuno Y, Koshimizu TA, Fujii S, Sekiya Y, Miyamoto Y, Tamura M, Yumura Y, Nihei H, Kobayashi M, Tsujimoto G. "Inhibition of protein kinase CK2 prevents the progression of glomerulonephritis." **Proc. Natl. Acad. Sci. U. S. A.**, 102(21): 7736-41, 2005

2. 学会発表

1. 奥野恭史、「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向けたケミカルーゲノムネットワークの構築」、第７８回日本薬理学会年会　シンポジウム：ケミゲノミクスの新しい展開とゲノム創薬科学、横浜（2005）

2. Okuno, Y., "Bioinformatics for Chemical Genomics", The 2005 AAPS Annual Meeting, Cutting-Edge Bioinformatics from Nanotechnology to Toxicogenomics, Nashville, USA

3. 奥野恭史, 種石慶, 土屋創健, 王質輝, 辻本豪三:ゲノム創薬支援システムの開発, 日本薬学会第１２６年会, 仙台, 2006年3月28日

H. 知的財産権の出願・登録状況

1. 特許取得

１. 特願２００５−１９２６７５、「データ処理装置、データ処理プログラム、それを格納したコンピュータ読み取り可能な記録媒体、およびデータ処理方法」、平成１７年６月３０日出願、出願者　京都大学、発明者　奥野恭史、辻本豪三、梁　智允、種石　慶

2. 実用新案登録

無し

3. その他

無し

研究成果の刊行に関する一覧表

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
Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H., Tsujimoto, G.	GLIDA: GPCR-Ligand database for Chemical Genomic Drug Discovery	**Nucleic Acids Research**	34	D673-7	2006
Zhu, S., Okuno, Y., Tsujimoto, G., Mamitsuka, H.	A probabilistic model for mining implicit 'chemical compound-gene' relations from literature	**Bioinformatics**	21(s2)	245-51	2005
Sugimoto Y, Fukada Y, Mori D, Tanaka S, Yamane H, Okuno Y, Deai K, Tsuchiya S, Tsujimoto G, Ichikawa A.	Prostaglandin E2 Stimulates Granulocyte Colony-Stimulating Factor Production via the Prostanoid EP2 Receptor in Mouse Peritoneal Neutrophils	**J. Immunol.**	175(4)	2606-12	2005
Adachi T., Okuno Y, Takenaka S, Matsuda K, Ohta N, Takashima K, Yamazaki K, Nishimura D, Miyatake K, Mori C, Tsujimoto G.	Comprehensive analysis of the effect of phytoestrogen, daidzein, on a testicular cell line, using mRNA and protein expression profile	**Food Chem. Toxicol.**	43(4)	529-35	2005

Yamada M, Katsuma S, Adachi T, Hirasawa A, Shiojima S, Kadowaki T, Okuno Y, Koshimizu TA, Fujii S, Sekiya Y, Miyamoto Y, Tamura M, Yumura Y, Nihei H, Kobayashi M, Tsujimoto G.	Inhibition of protein kinase CK2 prevents the progression of glomerulonephritis	**Proc. Natl. Acad. Sci. U. S. A.**	102(21)	7736-41	2005

研究成果の刊行物・別刷

Nucleic Acids Research, 2006, Vol. 34, Database issue ***D673–D677***
doi:10.1093/nar/gkj028

GLIDA: GPCR-ligand database for chemical genomic drug discovery

Yasushi Okuno*, Jiyoon Yang, Kei Taneishi, Hiroaki Yabuuchi and Gozoh Tsujimoto

Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Received August 15, 2005; Revised and Accepted September 22, 2005

ABSTRACT

G-protein coupled receptors (GPCRs) represent one of the most important families of drug targets in pharmaceutical development. GPCR-LIgand DAtabase (GLIDA) is a novel public GPCR-related chemical genomic database that is primarily focused on the correlation of information between GPCRs and their ligands. It provides correlation data between GPCRs and their ligands, along with chemical information on the ligands, as well as access information to the various web databases regarding GPCRs. These data are connected with each other in a relational database, allowing users in the field of GPCR-related drug discovery to easily retrieve such information from either biological or chemical starting points. GLIDA includes structure similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs). By analyzing the correlation patterns between GPCRs and ligands, we can gain more detailed knowledge about their interactions and improve drug design efforts by focusing on inferred candidates for GPCR-specific drugs. GLIDA is publicly available at http://gdds.pharm.kyoto-u.ac.jp:8081/glida. We hope that it will prove very useful for chemical genomic research and GPCR-related drug discovery.

INTRODUCTION

The superfamily of G-protein coupled receptors (GPCRs) forms the largest class of cell surface receptors. These molecules regulate various cellular functions responsible for physiological responses (1). GPCRs represent one of the most important families of drug targets in pharmaceutical development (2). A large majority of human-derived GPCRs still remain 'orphans' with no identified natural ligands or functions, and thus a key goal of GPCR research related to drug design is to identify new ligands for such orphan GPCRs.

With the unprecedented accumulation of the genomic information, databases and bioinformatics have become essential tools to guide GPCR research. The GPCRDB (http://www.gpcr.org/7tm/) (2) and IUPHAR (http://iuphar-db.org/iuphar-rd/index.html) (3) receptor databases are representatives of widely used public databases covering GPCRs. These databases, which provide substantial data on the GPCR proteins and pharmacological information on receptor proteins containing GPCRs, are mainly focused on biological aspects of the gene products or proteins. In spite of the significance of ligand compounds as drug leads, the relationships between GPCRs and their ligands and/or chemical information on the ligands themselves are not yet fully covered.

On the other hand, there is increasing interest in collecting and applying chemical information in the post-genome era. This new trend is called 'chemical genomics', in which biological information and chemical information are integrated on the genome scale (4,5). PubChem (http://pubchem.ncbi.nlm.nih.gov/) (6), KEGG/LIGAND (http://www.genome.jp/kegg/ligand.html) (7) and ChEBI (http://www.ebi.ac.uk/chebi/) (8) have been developed as databases related to chemical genomics. KEGG/LIGAND and ChEBI contain primarily biochemical information on reported enzymatic reactions. Recently, NIH (the National Institutes of Health) opened PubChem, a public database providing information on the chemical structures of small molecules. However, one cannot retrieve direct information relating these chemical structures to gene or protein entries. Although chemical genomic approaches have thrown new light on relationships between receptor sequences and compounds that interact with particular receptors, the GPCR-ligand information is not well represented in these large-scale databases for chemical genomics.

There are still very few publicly available databases or tools for GPCR-specialized drug discovery from the viewpoint of chemical genomics. Herein, we have developed a novel relational database, GLIDA (GPCR-LIgand DAtabase) (9).

*To whom correspondence should be addressed. Tel: +81 75 753 9264; Fax: +81 75 753 4544; Email: okuno@pharm.kyoto-u.ac.jp

Table 1. The current numbers of GLIDA ligands and GPCRs and their respective links

Information item	Number of entries
GPCR entries	3738
Links to Entrez Gene	3073
Links to GPCRDB	3738
Links to UniProt	3738
Links to IUPHAR	389
Links to KEGG	595
Ligand entries	649
Cas registry number	320
Lolecular structure	364
Links to PubChem	242
Links to ChEBI	28
Links to KEGG	109
GPCR-ligand pair entries	1989
GPCR entries	281
Ligand entries	632

GLIDA contains biological information on GPCRs and chemical information on their ligand compounds. Furthermore, it provides various analytical data on GPCR-ligand correlations by incorporating bioinformatics and chemoinformatics methods, and thus it should prove very useful for chemical genomic research in GPCR-related drug discovery.

DATA CONTENTS

GLIDA contains three types of primary data: biological information on GPCRs, chemical information on their ligands and information on binding of specific GPCR-ligand pairs. The GPCR entries were acquired from the deposits of human, mouse and rat entries in the GPCRDB because these three species include sufficient information regarding ligands, and rats and mice are representative model animals for drug discovery. The ligand information was manually collected and curated using various public web sites and commercial DBs, such as the IUPHAR Receptor Database, PubMed, PubChem and MDL ISIS/Base 2.5. Table 1 indicates the size and scope of the GLIDA database.

GPCR and ligand data

The database lists general information on GPCR and ligand data, respectively. The general information table of GPCR contains gene names, family names, protein sequences and links to other biological databases, such as GPCRDB, UniProt, IUPHAR, Entrez Gene and KEGG. The ligand result page provides a general information table containing names, molecular structures, CAS registry numbers, formulas, molecular weights, MOLfiles and links to the other chemical databases KEGG, PubChem and ChEBI.

Information on binding of GPCR-ligand pairs

The correlation information relating GPCRs to particular ligands, a key issue for GPCR-related drug discovery, is stored in a relational database. GLIDA allows users to retrieve GPCR-ligand binding information dynamically and continuously. When users retrieve a GPCR (or ligand) entry, its result page displays all entries showing the corresponding ligands (or GPCR entries) with their binding activity types, as well as references. The references are hyperlinked with the corresponding PubMed literature or the IUPHAR pages that were used to collect the information regarding GPCR-ligand binding. The activity types include agonist, inverse agonist, antagonist and so on. An agonist will bind to and activate the corresponding GPCRs, whereas an antagonist will bind to and block the activity of the corresponding GPCRs. An inverse agonist binds to GPCRs and reduces the fraction of them that are in an active conformation, and a partial agonist is an agonist that in a given tissue, under specified conditions, cannot elicit as large an effect as another agonist acting through the same GPCRs in the same tissue can.

WEB INTERFACE AND APPLICATION

GLIDA was constructed on the LAMP (Linux, Apache, MySQL and PHP) platform. GLIDA is available at http://gdds.pharm.kyoto-u.ac.jp:8081/glida. The web interface of GLIDA includes a GPCR search page (Figure 1a) and a ligand search page (Figure 1b). Each page consists of a classification table and a keyword search box. The user can search a GPCR (or ligand) manually from the guide-tree of the classification table, or automatically by using the keyword search function of MySQL. Every GPCR (or ligand) has its own result page (Figure 1c or d) containing a general information table for a GPCR (or ligand), a table of its correlated ligands (or GPCRs) and a button to carry out a similarity search and correlation analysis. Clicking the button starts the calculation, and an analytical report page (Figure 1e) then appears with a list of the top 25 entries that are most similar to the GPCR (or ligand) and a correlation map of the 25 GPCRs (or ligands) and their corresponding binding pairs. A search starting from ligand retrieval proceeds in the same way.

Hierarchical classification

The GPCR classification table on the search page was adapted from the phylogenetic tree of the GPCRDB information system (http://www.gpcr.org/7tm/phylo/phylo.html). As for the ligand classification table, GLIDA offers an original one (Figure 1b) that is based on a cluster analysis of the ligand structures as follows. We converted the structural images of the ligands into computational MDL Mol files using ISIS/Draw software. Next, we calculated distance metrics among all of the ligands using the frequency profiles of the atoms and the bonds of the KEGG atom types (10), and carried out complete-linkage clustering. We manually defined subclusters based on their common structural skeletons. Both the GPCR and ligand classification tables display the entries of the corresponding GPCRs or ligands at the end of the tree, and these are hyperlinked with their respective result pages.

Similarity search and GPCR-LIGAND correlation maps

GLIDA has a structure similarity search function on its result pages. Alignment scores of protein sequences generated by the BLAST algorithm provide similarity measures for GPCRs. Ligand similarity is defined by the dissimilarity (distance) of frequency profile patterns generated from the constitutive atoms and bonds of the chemical structure, using the KEGG atom types (10,11). From this similarity search, the 25 most

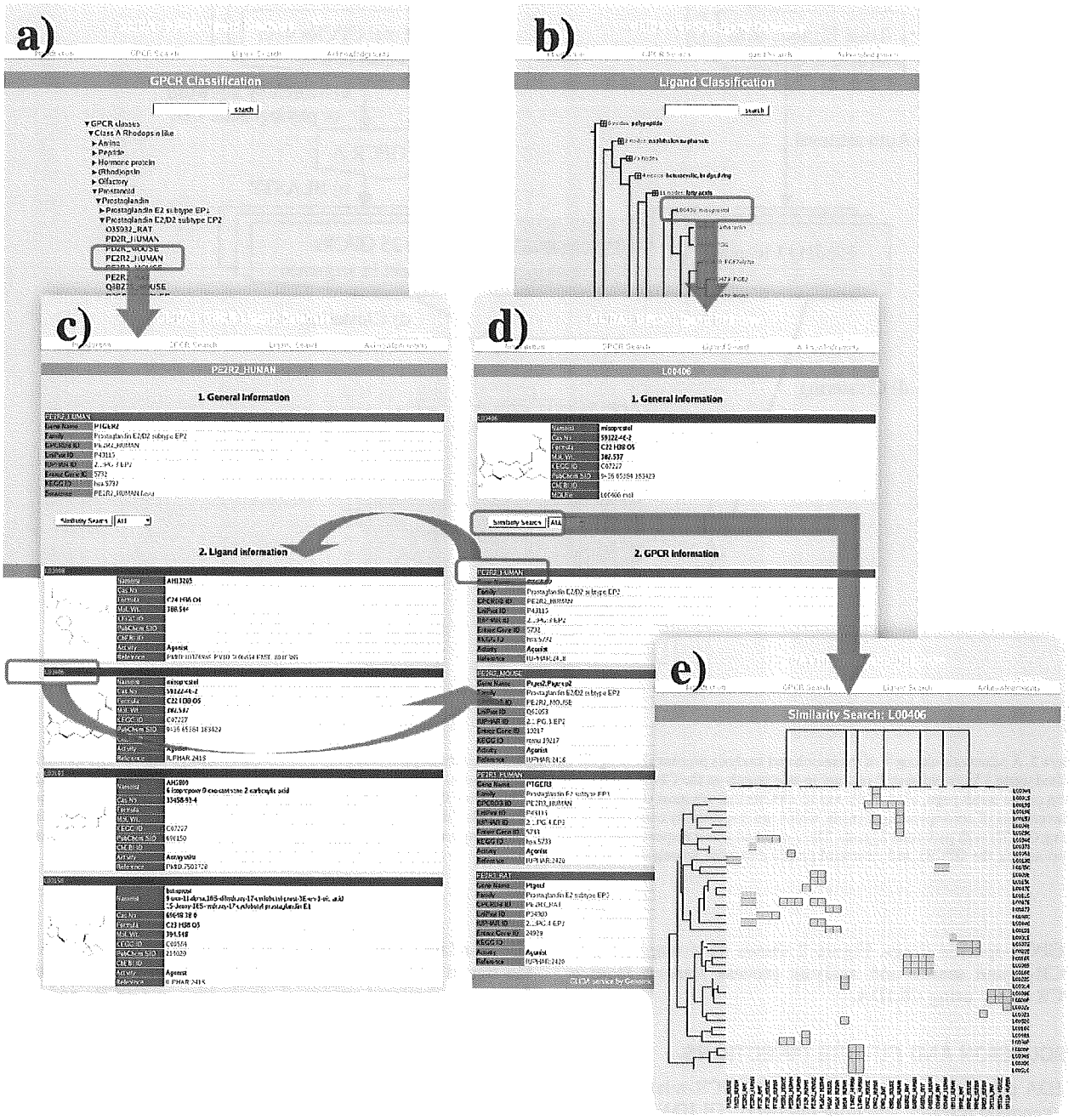


Figure 1. A screenshot of GLIDA showing its linked relations among search pages (**a**, **b**), result pages (**c**, **d**) and an analytical report page (**e**).

similar GPCRs (or 40 ligands) are retrieved and listed with their similarity scores on an analytical report page.

As the similarity search calculation is proceeding, GLIDA illustrates the correlation map (Figure 2e) showing the homologous GPCRs (or ligands) and their ligands (or GPCRs) that are retrieved. This map shows spots that match the GPCRs and their ligands in a two-dimensional matrix. The ordering along the *x*-axis and the *y*-axis are calculated respectively by two-way clustering of the GPCRs and the ligands based on their similarities. In particular, the ordering along the *x*- and *y*-axis allows users to evaluate information regarding similarities and correlations between GPCRs and ligands simultaneously. By analyzing the correlation patterns between GPCRs and ligands that are illustrated by these maps, we can gain detailed knowledge about their interactions and utilize this information to infer possible candidates for development

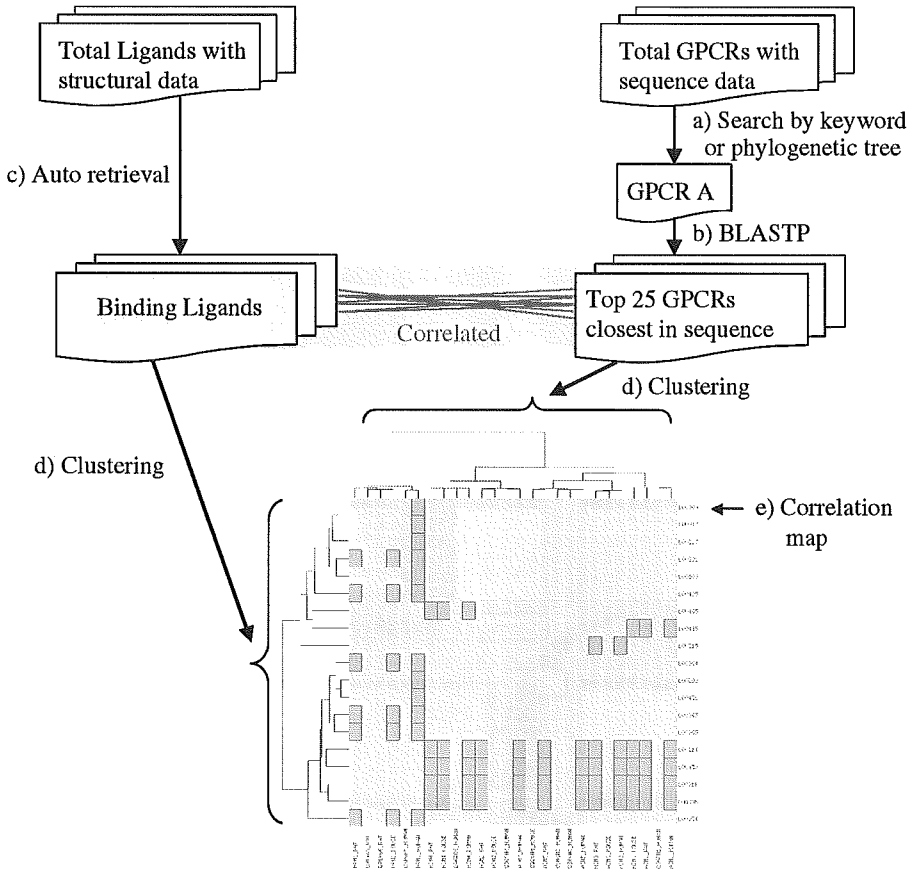


Figure 2. A schematic example of the search and analysis process showing GPCR-ligand correlations produced from a GPCR query using GLIDA. (**a**) If GPCR A is selected using a keyword search or a guide-tree search on the GPCR search page, its retrieved data will be displayed in its result page, (**b**) By clicking an analysis button on the result page, a list of the top 25 GPCRs that are most similar in sequence, including GPCR A, are obtained by the BLASTP calculation. (**c**) The server retrieves a list of corresponding ligands, which are respectively correlated with the 25 GPCRs. (**d**) Finally, a map is displayed to help visualize the matching spots linking GPCRs with particular ligands. The *x*-axis and *y*-axis respectively indicate the clustering results for GPCRs and ligands, calculated using sequence alignment scores among the GPCRs and structural profile distances among the ligands.

of GPCR-specific drugs. Figure 2 shows an example of the GPCR-ligand search and analysis process starting from a GPCR query using GLIDA.

DISCUSSION AND FUTURE DIRECTIONS

GLIDA provides a unique database for GPCR-related chemical genomic research and drug discovery. GLIDA is distinct from other public chemical genomic databases because it contains original, GPCR-specific chemical entries, although the total scale of its contents is not yet large (Table 1). GLIDA provides several advantages over other databases, in that a search can be started either from a GPCR or from a ligand. Thus, searches may be carried out in a dynamic and user-friendly way. GLIDA's coverage of chemical and biological information simultaneously also provides an advantage to users by saving them the time and labor required to search multiple databases. The ligand search page is another distinct characteristic of GLIDA in that it displays the structural distribution of ligands, and thereby facilitates research on GPCR-related drugs by incorporating structural aspects of the ligand compounds. The analytical report pages resulting from the calculated structural similarities of GPCRs and ligands can give the user deep insights into the GPCR-ligand relationships. The lists of neighboring ligands (or GPCRs) and the correlation maps are useful visualizing tools for analyzing correlations among their structural features and their GPCR-ligand binding properties. Because the GLIDA algorithms can be applied to proteins other than the GPCR family, it may also be considered as a promising database for chemical genomics research.

GLIDA will be updated continuously. In particular, we are planning to computationally extract GPCR-ligand information from the literature and from patents using a text-mining tool, and to increase the number of ligand entries immediately. Further information on ligands from various computable chemical descriptors is currently being incorporated, and GLIDA will be combined with a system for predicting novel ligands of orphan GPCRs in the future. Furthermore, we also plan to carry out XML publication of GLIDA.

ACKNOWLEDGEMENTS

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, from the Ministry of Health, Labor and Welfare of Japan and from the 21st Century COE program 'Knowledge information infrastructure for Genome Science'. Funding to pay the Open Access publication charges for this article was provided by the Ministry of Health, Labor and Welfare of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. George,S.R., O'Dowd,B.F. and Lee,S.P. (2002) G-protein-coupled receptor oligomerization and its potential for drug discovery. *Nature Rev. Drug Discov.*, **1**, 808–820.
2. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
3. Fredholm,B.B., Fleming,W.W., Vanhoutte,P.M. and Godfraind,T. (2002) The role of pharmacology in drug discovery. *Nature Rev. Drug Discov.*, **1**, 237–248.
4. Lipinski,C. and Hopkins,A. (2004) Navigating chemical space for biology and medicine. *Nature*, **432**, 855–861.
5. Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
6. Zerhouni,E. (2003) Medicine: the NIH roadmap. *Science*, **302**, 63–72.
7. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic. Acids Res.*, **30**, 402–404.
8. Brooksbank,C., Cameron,G. and Thornton,J. (2005) The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.*, **33**, D46–D53.
9. Yang,J., Okuno,Y. and Tsujimoto,G. (2004) GLIDA: GPCR and Ligand Database. *Genome Informatics*, **15**, P057.
10. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
11. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.

Vol. 21 Suppl. 2 2005, pages ii245–ii251
doi:10.1093/bioinformatics/bti1141

Text Mining

A probabilistic model for mining implicit 'chemical compound–gene' relations from literature

Shanfeng Zhu[1], Yasushi Okuno[2], Gozoh Tsujimoto[2] and Hiroshi Mamitsuka[1,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, Japan and
[2]Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

Motivation: The importance of chemical compounds has been emphasized more in molecular biology, and 'chemical genomics' has attracted a great deal of attention in recent years. Thus an important issue in current molecular biology is to identify biological-related chemical compounds (more specifically, drugs) and genes. Co-occurrence of biological entities in the literature is a simple, comprehensive and popular technique to find the association of these entities. Our focus is to mine implicit 'chemical compound and gene' relations from the co-occurrence in the literature.

Results: We propose a probabilistic model, called the mixture aspect model (MAM), and an algorithm for estimating its parameters to efficiently handle different types of co-occurrence datasets at once. We examined the performance of our approach not only by a cross-validation using the data generated from the MEDLINE records but also by a test using an independent human-curated dataset of the relationships between chemical compounds and genes in the ChEBI database. We performed experimentation on three different types of co-occurrence datasets (i.e. compound–gene, gene–gene and compound–compound co-occurrences) in both cases. Experimental results have shown that MAM trained by all datasets outperformed any simple model trained by other combinations of datasets with the difference being statistically significant in all cases. In particular, we found that incorporating compound–compound co-occurrences is the most effective in improving the predictive performance. We finally computed the likelihoods of all unknown compound–gene (more specifically, drug–gene) pairs using our approach and selected the top 20 pairs according to the likelihoods. We validated them from biological, medical and pharmaceutical viewpoints.

Contact: mami@kuicr.kyoto-u.ac.jp

1 INTRODUCTION

Traditional molecular biology tells us that genetic information is transferred from DNA to protein and ultimately shows up as protein functions. The final goal of molecular biology in this 'central dogma' is to identify and understand biological activities regulated by proteins so that they can be managed. The most important protein function is to catalyze biochemical reactions for the synthesis of one chemical compound from another. Thus the first step of the above goal may be compared with detecting one or more chemical compounds for which each protein can catalyze.

Recently the importance of chemical compounds has been emphasized more in molecular biology, and a new research field, called 'chemical genomics', has attracted a great deal of attention. In fact, one of the five items to be taken up by the National Institute of Health (NIH) roadmap initiative is a chemoinformatics project for building small molecular libraries. This chemoinformatics project will develop a new compound database of chemical structures and their biological activities, with the idea of promoting pharmaceutical research, such as discovering new drugs. This database, called PubChem, will house compound information on the screening and probe data newly obtained by the Molecular Libraries Screening Centers Network (MLSCN) as well as those from the current scientific literature. A related fact is that databases of chemical compounds and their biochemical reactions have also been developed in recent years. For example, the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (Kanehisa *et al.*, 2004), which is a database of metabolic pathways generated by gathering biochemical reactions, has drastically grown both in the size of stored reactions and the number of citations in the biological and medical sciences. European Bioinformatics Institute (EBI) also developed a freely available database of small molecular entities, ChEBI (Brooksbank *et al.*, 2005), which stands for 'Chemical Entities of Biological Interest'.

Thus an important issue in current molecular biology is to identify biological-related chemical compounds and genes, which is a fundamental step of chemical genomics research. Mining biomedical and biological literature databases, such as Medline (Wheeler *et al.*, 2005), for identifying these kinds of biological-related entities has been actively tackled in the last few years (Blasckel *et al.*, 2002; Yandell *et al.*, 2002). Co-occurrence of biological entities in the literature is a simple, comprehensive and popular technique to identify the association of these entities (Stapley *et al.*, 2000; Jenssen *et al.*, 2001; Chang *et al.*, 2004). This technique is based on the following hypothesis: if a biological entity appears with another biological entity in the same document, these two entities should be biologically related with high probability. This hypothesis was already experimentally testified by many researchers (Jenssen *et al.*, 2001; Chang *et al.*, 2004). We will describe the details of this and related issues in Section 2.

Thus we focus on the co-occurrence information in the literature to discover implicit 'chemical compound–gene' relations, being those which are not in existing compound–gene co-occurrences in the literature but could be discovered from the co-occurrence data. All possible combinations of compounds and genes are very large in number, but obviously known co-occurrences of them are very limited, even though the literature is very abundant in size. Thus we attempt to use not only the co-occurrence of a chemical compound

*To whom correspondence should be addressed.

