(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5448447号
(P5448447)
(45)発行日　平成26年3月19日(2014.3.19)　(24)登録日　平成26年1月10日(2014.1.10)

(51)Int.Cl.　　F I
G06F 19/16 (2011.01)　　G06F 19/16
G06F 19/24 (2011.01)　　G06F 19/24

請求項の数 2　(全 32 頁)

(21)出願番号　特願2008-517917(P2008-517917)
(86)(22)出願日　平成19年5月25日(2007.5.25)
(86)国際出願番号　PCT/JP2007/060736
(87)国際公開番号　WO2007/139037
(87)国際公開日　平成19年12月6日(2007.12.6)
審査請求日　平成22年5月24日(2010.5.24)
(31)優先権主張番号　特願2006-147433(P2006-147433)
(32)優先日　平成18年5月26日(2006.5.26)
(33)優先権主張国　日本国(JP)

前置審査

(73)特許権者　504132272
国立大学法人京都大学
京都府京都市左京区吉田本町３６番地１
(74)代理人　100115200
弁理士　山口　修之
(74)代理人　100158366
弁理士　井戸　篤史
(72)発明者　奥野　恭史
京都府京都市左京区吉田下阿達町４６－２９　京都大学大学院薬学研究科内
(72)発明者　種石　慶
京都府京都市左京区吉田下阿達町４６－２９　京都大学大学院薬学研究科内

最終頁に続く

(54)【発明の名称】ケミカルゲノム情報に基づく、タンパク質－化合物相互作用の予測と化合物ライブラリーの合理的設計

(57)【特許請求の範囲】
【請求項１】
（Ａ）
第１の化学物質は化合物であり、
第２の化学物質は核酸、タンパク質又はそれらの複合体であり、
第１の特徴量は、前記第１の化学物質のconstitutional descriptors、topological descriptors、walk and path counts、connectivity indices、information indices、2D autocorrelations、edge adjacency indices、burden eigen value descriptors、topological charge indices、eigenvalue-based indices、functional group counts、atom-centered fragments、及びmolecular propertiesから選ばれる複数を組み合わせた記述子がベクトルとして表現され、
第２の特徴量は、前記第２の化学物質のアミノ酸配列からスペクトラム法により計算された記述子のみがベクトルとして表現され、
互いに相互作用することが既知である前記第１の化学物質と前記第２の化学物質とについて、
該第１の化学物質の第１の特徴量と該第２の化学物質の第２の特徴量とを組み合わせた特徴ベクトルから、請求項１に記載のデータ処理方法によってを、サポートベクターマシンによって写像変換することにより、特徴空間に学習モデルを構築する工程と、
（Ｂ）

相互作用の予測対象となる第１の化学物質の第１の特徴量と
相互作用の予測対象となる第２の化学物質の第２の特徴量とを組み合わせた特徴ベクトルを
写像変換することにより前記特徴空間にマッピングする工程と、
（Ｃ）
前記特徴ベクトルの前記特徴空間内の位置によって、
前記相互作用の予測対象となる第１の化学物質と前記相互作用の予測対象となる第２化学物質とが
相互作用するか否かが判定される工程と、
からなるデータ処理方法。

【請求項２】
請求項１に記載のデータ処理方法を計算機に実行させるための
データ処理プログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、２種類の化学物質のデータベースを用いて効率よくスクリーニングを行い、合理的設計を行うためのデータ処理方法に関する。詳細には、化合物の化学特性などの化学物質情報と遺伝子の配列情報など生物学的情報との両情報を用いて、化合物－タンパク質相互作用予測するデータ処理方法に関する。

【背景技術】
【０００２】
ヒトゲノムが公開されている現在、ゲノム情報を利用した医薬応用が注目されている。ゲノム情報に基づく医薬品開発には、タンパク質（遺伝子）と化合物の相互作用を解明することが不可欠であるが、これら相互作用を実験的に解明するには莫大な労力が必要である。タンパク質と化合物との相互作用を予測する従来技術としては、一般に、立体構造モデルを用いた予測システムが知られている（非特許文献１～４）。このシステムは、生体高分子の立体構造情報に基づき、リガンドとの安定な複合体構造およびその結合の強さを推定する方法であり、ドッキングスタディと呼ばれている。

【０００３】
特表２００２－５３０７２７号公表公報(特許文献１）は、ケミカル空間の高活性領域を特定し、ライブラリ構築を行うことを記載している。しかし、この方法は、ケミカル情報（構造活性相関情報、ファルマコフォア情報など）のみを用いてケミカル空間を定義しているに過ぎない。

【０００４】
また発明者自身の従来法（非特許文献５）は、化合物群とタンパク質群とについて、別々にクラスター解析などの統計処理をした後、化合物処理データとタンパク質処理データといを融合し、２次元マップに表示することにより、タンパク質と化合物の相互作用ペアを推測するものである。

【０００５】
２００４年に米国ＮＩＨは国家プロジェクトとしてケミカルゲノミクスプロジェクトを開始した。以来、米国や欧州を中心にゲノム情報の化学分野への応用が世界中で取り組まれている。従って、少なくとも米国など先進国において、効率的な予測方法に対する需要が存在する。

【０００６】
【特許文献１】特表２００２－５３０７２７号公表公報
【非特許文献１】Yoshifumi Fukunishi、 Yoshiaki Mikami、 and Haruki Nakamura “Thefillingpotential method:A methodforestimating the free energy surfaceforprotein-liganddocking” J.Phys. Chem. B.(2003) 107、13201-13210.
【非特許文献２】Shoichet、 B.K.、D.L. Bodian、and I.D. Kuntz、 “Molecular docki

ng usingshapedescriptors.”J.Comp. Chem.、 1992. 13(3)、 380-397.
【非特許文献３】Jones G、 WillettP、 Glen RC、 Leach AR、Taylor R.“Development and validation of a geneticalgorithm for flexible docking.” J Mol Biol.1997.267(3):727-748.
【非特許文献４】Rarey M、 KramerB、Lengauer T.“Time-efficient docking of flexibleligands intoactivesites ofproteins.” Proc Int Conf IntellSyst Mol Biol.1995;3:300-308.
【非特許文献５】Okuno、 Y.、Yang、J.、Taneishi、 K.、 Yabuuchi、H.、 Tsujimoto、G、 “GLIDA:GPCR-LiganddatabaseforChemical Genomic Drug Discovery” Nucleic AcidsResearch、34、D673-677、 2006

【発明の開示】
【発明が解決しようとする課題】
【０００７】
タンパク質の立体構造モデルを用いた予測手法は、Ｘ線結晶解析などで信頼できる立体構造を得ている場合でなければ科学的根拠を有する予測が成り立たないこと、さらに立体構造を用いた計算は負荷が高く計算時間が膨大にかかるため、無限のバリエーションを有する化合物群やタンパク質群の組み合わせを網羅的に計算することが不可能であることなどの問題点を有していた。
【０００８】
また、化合物の化学物質情報のみに基づく手法では、予測性能が低く効率化が求められていた。本発明は化合物のスクリーニングを合理的にかつ効率的に行うことを目的とする。
【課題を解決するための手段】
【０００９】
本発明のデータ処理方法は、第１の化学物質は化合物であり、第２の化学物質は核酸、タンパク質又はそれらの複合体であり、第１の特徴量は、前記第１の化学物質の１種類以上の化学物質情報からなるベクトルとして表現され、第２の特徴量は、前記第２の化学物質の１種類以上の生物学的情報からなるベクトルとして表現されるものであって、第１の特徴量を表す第１空間と第２の特徴量を表す第２空間との相関が高くなるように、多変量解析手法又は機械学習法によって、第１の特徴量および前記第２の特徴量を写像変換するデータ処理方法である。
【００１０】
また、本発明の他のデータ処理方法は、第１の化学物質と第２の化学物質とが相互作用するか否かが判定されるデータ処理方法であり、以下の工程（Ａ）～（Ｃ）からなる。（Ａ）互いに相互作用することが既知である第１の化学物質と第２の化学物質とについて、該第１の化学物質の第１の特徴量および該第２の化学物質の第２の特徴量を、第１の特徴量を表す第１空間と第２の特徴量を表す第２空間との相関が高くなるように、多変量解析手法又は機械学習法によって、第１の特徴量および第２の特徴量を写像変換する工程、（Ｂ）
相互作用の予測対象となる第１の化学物質の第１の特徴量を写像変換することによって該第１の化学物質を、（Ａ）の写像変換後の第１の特徴量を表す第１空間にマッピングし、相互作用の予測対象となる第２の化学物質の第２の特徴量を写像変換することによって該第２の化学物質を、（Ａ）の写像変換後の第２の特徴量を表す第２空間にマッピングする工程、（Ｃ）相互作用の予測対象となる第１の化学物質の、（Ｂ）の写像変換後の第１の特徴量の座標位置と相互作用の予測対象となる第２の化学物質の、（Ｂ）の写像変換後の第２の特徴量の座標位置とによって、該第１の化学物質と該第２の化学物質とが相互作用するか否かが判定される工程。
【００１１】
また、本発明の他のデータ処理方法は、クエリ側の予測対象となる化学物質と相互作用すると予測される化学物質を選出するものであって、以下の工程（Ａ）、（Ｂ）、（Ｃ１）、（Ｄ１）（Ｅ１）又は（Ａ）、（Ｂ）、（Ｃ２）、（Ｄ２）、（Ｅ２）からなる。（Ａ

）互いに相互作用することが既知である第１の化学物質と第２の化学物質とについて、該第１の化学物質の第１の特徴量および該第２の化学物質の第２の特徴量を、第１の特徴量を表す第１空間と第２の特徴量を表す第２空間との相関が高くなるように、多変量解析手法又は機械学習法によって、第１の特徴量および第２の特徴量を写像変換する工程、（Ｂ）相互作用の予測対象となる第１の化学物質の第１の特徴量を写像変換することによって該第１の化学物質を、（Ａ）の写像変換後の第１の特徴量を表す第１空間にマッピングし、相互作用の予測対象となる第２の化学物質の第２の特徴量を写像変換することによって該第２の化学物質を、（Ａ）の写像変換後の第２の特徴量を表す第２空間にマッピングする工程、（Ｃ１）相互作用の予測対象のうち、クエリ側が第１の化学物質である場合、第１空間において、相互作用の予測対象となる第１の化学物質における第１の特徴量が占有する領域を目的領域として算出する工程と、（Ｄ１）第２空間において、第１空間の目的領域に対応する標的領域を算出する工程と、（Ｅ１）第２空間にマッピングされた第２の化学物質のうち、該標的領域内に存在する第２の化学物質が選出される工程と、又は、（Ｃ２）相互作用の予測対象のうち、クエリ側が第２の化学物質である場合、第２空間において、相互作用の予測対象となる第２の化学物質における第２の特徴量が占有する領域を標的領域として算出する工程と、（Ｄ２）第１空間において、第２空間の標的領域に対応する目的領域を算出する工程と、（Ｅ２）第１空間にマッピングされた第１の化学物質のうち、該目的領域内に存在する第１の化学物質が選出される工程。

【００１２】

また、本発明のデータ処理方法は、第１の化学物質は化合物であり、第２の化学物質は核酸、タンパク質又はそれらの複合体であり、第１の特徴量は、前記第１の化学物質の１種類以上の化学物質情報からなるベクトルとして表現され、第２の特徴量は、前記第２の化学物質の１種類以上の生物学的情報からなるベクトルとして表現され、第１の特徴量と第２の特徴量とを、多変量解析手法又は機械学習法によって写像変換することにより、特徴空間を構築するデータ処理方法である。写像変換される特徴量は、第１の特徴量と第２の特徴量を連結した特徴量であってもよい。

【００１３】

本発明の他のデータ処理方法は、上記データ処理方法で構築した特徴空間を用い、相互作用の予測対象となる第１の化学物質の第１の特徴量と相互作用の予測対象となる第２の化学物質の第２の特徴量とを多変量解析手法又は機械学習法によって写像変換することにより特徴空間にマッピングし、第１の特徴量の特徴空間内の位置と第２の特徴量の特徴空間内の位置とによって相互作用の予測対象となる第１の化学物質と相互作用の予測対象となる第２化学物質とが相互作用するか否かが判定されるデータ処理方法である。また、特徴空間にマッピングされる特徴量は、相互作用の予測対象となる第１の化学物質の第１の特徴量と相互作用の予測対象となる第２の化学物質の第２の特徴量とを連結した特徴量であってもよい。

【００１４】

また、本発明のライブラリは、上記データ処理方法によって得られた判定結果や選出された化学物質から設計される。また、本発明のデータ処理プログラムは、上記データ処理方法を計算機に実行させるためのプログラムである。また、本発明の化学物質の製造方法は、上記データ処理方法を用いて相互作用の予測のクエリ側の化学物質と相互作用すると予測された計算対象側の化学物質を合成するものである。さらに、本発明の化合物は、上記データ処理方法を用いることによって生産された化合物である。

【発明の効果】

【００１５】

本発明により、化合物の化学特性などの化学物質情報とゲノム情報などの生物学的情報との両情報を用いて、化合物－タンパク質の相互作用を予測する計算手法を開発することができた。この手法では、従来型の化合物の化学特性情報のみを用いた予測とは異なり、遺伝子の配列情報を加えることによって予測の精度向上をはかることに成功した。

【００１６】

従って、この計算手法は具体的に次の２つに適用することができる。
１）化合物ライブラリの生物活性に基づく合理的設計が実現できる。
２）従来法よりも性能の良いリード化合物探索ができる。
【００１７】
以下に、本発明の好ましい実施形態を示すが、当業者は本発明の説明および当該分野における周知慣用技術からその実施形態などを適宜実施することができ、本発明が奏する作用および効果を容易に理解することが認識されるべきである。従って、本発明のこれらおよび他の利点は、必要に応じて添付の図面等を参照して、以下の詳細な説明を読みかつ理解すれば、当業者には明白になることが理解される。
【発明を実施するための最良の形態】
【００１８】
以下、本発明を説明する。本明細書の全体にわたり、単数形の表現は、特に言及しない限り、その複数形の概念をも含むことが理解されるべきである。従って、単数形の冠詞または形容詞（例えば、英語の場合は「ａ」、「ａｎ」、「ｔｈｅ」など）は、特に言及しない限り、その複数形の概念をも含むことが理解されるべきである。また、本明細書において使用される用語は、特に言及しない限り、当該分野で通常用いられる意味で用いられることが理解されるべきである。したがって、他に定義されない限り、本明細書中で使用される全ての専門用語および科学技術用語は、本発明の属する分野の当業者によって一般的に理解されるのと同じ意味を有する。矛盾する場合、本明細書（定義を含めて）が優先する。
【００１９】
本明細書において「化学物質」とは、物質という一般用語の中で、とくに化学的な立場で物質を取り扱う場合の用語であり、任意の一定の分子構造をもつ物質をいう。
【００２０】
本明細書において「空間座標」とは、対象物の空間内の位置を特定するための指標である。「空間」とは、「スペース」と交換可能に使用され、集合の別名であり、なんらかの位相や幾何学的構造を想定するものをいう。空間は、例えば、化学物質情報や生物学的情報により定義される。空間としては、化合物記述子や化学特性などの化学物質情報で定義されるケミカル空間や、発現情報、パスウェイ情報、機能情報、および生物活性などの生物学的情報で定義されるバイオ空間（Ｃｕｒｒｅｎｔ　Ｏｐｉｎｉｏｎ　ｉｎ　Ｃｈｅｍｉｃａｌ　Ｂｉｏｌｏｇｙ　２００５、　９：２９６－３０３）を例示することができる。
【００２１】
本明細書において「特性」とは、ある化学物質が有する特別の性質をいう。特性としては、例えば、融点、沸点、比重などの物理特性、反応性、酸性、アルカリ性などの化学特性、タンパク質構造、酵素活性、レセプターとの結合能、サイトカイン能、細胞との相互作用力などの生物学的特性を挙げることができるがそれらに限定されない。
【００２２】
本明細書において、第１特性と第２特性との間に「単純な関連が見られない」とは、両者の間に直感的な相関が見られないことをいう。第１特性を示すベクトルをａ、ｂとし、第２特性を示すベクトルをｘ、ｙとし、ａとｘ、ａとｙ、ｂとｘ、ｂとｙのペアには相関が直感的な相関が見られないとする。しかし、線形結合ベクトルｍ＊ａ＋ｎ＊ｂとＭ＊ｘ＋Ｎ＊ｙ（ｍ、ｎ、Ｍ、　Ｎは０で無い係数）との間に相関が見られる場合がある。つまり、第１特性と第２特性との間に単純な関連が見られないものであっても、適当な変換を加えたＦ（ａ、ｂ、…）とＦ’（ｘ、ｙ、…）の間に相関が見られる場合がある。
【００２３】
本明細書において「化合物」とは、化学変化によって２種またはそれ以上の元素の単体に分けることができる純粋物質をいう。2種以上の元素の原子の化学結合によって生じた純粋物質といってもよい。通常、各元素の組成比は一般に定比例の法則に従って一定であるが、不定比化合物のように組成比がある範囲で連続的に変化しても安定な結晶をつくる

ものもまた、本明細書において化合物の範疇に入れる。

【００２４】

本明細書において「生体物質」とは、生物に関連する任意の物質を言う。生体物質もまた、化学物質の一種として捕らえることができる。本明細書において「生体」とは、生物学的な有機体をいい、動物、植物、菌類、ウイルスなどを含むがそれらに限定されない。生体物質には、タンパク質、ポリペプチド、オリゴペプチド、ペプチド、ポリヌクレオチド、オリゴヌクレオチド、ヌクレオチド、核酸（例えば、ｃＤＮＡ、ゲノムＤＮＡのようなＤＮＡ、ｍＲＮＡのようなＲＮＡを含む）、ポリサッカライド、オリゴサッカライド、脂質、これらの複合分子（糖脂質、糖タンパク質、リポタンパク質など）などが包含されるがそれらに限定されない。通常、生体物質は、核酸、タンパク質、脂質、糖、プロテオリピッド、リポプロテイン、糖タンパク質およびプロテオグリカンなどであり得る。好ましくは、生体物質は、核酸（ＤＮＡまたはＲＮＡ）またはタンパク質である。

【００２５】

本明細書において「レセプター」とは、細胞上または核内などに存在し、外界からの因子または細胞内の因子に対する結合能を有し、その結合によりシグナルが伝達される分子をいう。レセプターは通常タンパク質の形態をとる。レセプターの結合パートナーは、通常リガンドという。

【００２６】

本明細書において「アゴニスト」とは、あるリガンドのレセプターに結合し、その物質のもつ作用と同等又は類似の作用を現わす因子をいう。また、「アンタゴニスト」とは、あるリガンドのレセプターへの結合に拮抗的に働き、それ自身はそのレセプターを介した生理作用を現わさない因子をいう。拮抗薬、遮断剤（ブロッカー）、阻害剤（インヒビター）などもこのアンタゴニストに包含される。

【００２７】

本明細書において「相関が最大になるように」「座標を変換」するとは、ある空間における各要素と別の空間における各要素との関係が、全体としてみたときに最大限に相関している状態を言う。このような定義は種々の計算手法によって達成することができる。

【００２８】

本明細書において「相関」とは、数理統計学や生物統計学において、一般に２つまたはそれ以上の変量のあいだの関連性をいい、2つの確率変数の間の直線的な共変関係をいう。数量分類学において、対象とする二つの操作的分類単位（ＯＴＵ）間の類似係数（類似の程度）によって表現することができる。従って、相関分析とは、対のデータに基づいて相関の有無を検証し、あるいは相関の大きさを推定したりする統計方法のことである。

【００２９】

本明細書において「相関係数」とは、２つの確率変数Ｘ、Ｙの間の関連を示す、一次変換で不変な量をいう。ＸとＹの共分散をＸおよびＹのおのおのの分散の積の平方根で割った値である。「相関関数」には、空間の相関関数と時間の相関関数がある。２点の物理量Ａ（ｒ）とＡ（ｒ’）の積の平均値Ａ（ｒ）Ａ（ｒ’）を空間の相関関数という。これが２点間の距離に対して指数関数的に変化するとき、これを

【００３０】

【数１０１】

$$\overline{A(r)A(r')} \propto \exp\left(-\frac{|r-r'|}{\xi}\right)$$

【００３１】

とおいて、 相関距離ξが定義される。時間の相関関数に対しても同様に相関時間τが定義される。ξやτは臨界点で発散する。

【００３２】

本明細書において「相関距離」とは、空間の各点に存在する確率変数Ａ（ｒ）の相関関数〈Ａ（ｒ）Ａ（ｒ’）〉が２点間距離ｒ＝｜ｒ－ｒ’｜の増大とともに絶対値として減

少していく場合、その減衰の目安となる距離をいう。相関関数がｅｘｐ（－ｒ／ξ）の形である場合、ξが相関距離である。減衰が指数関数的でない場合に一義的に定義することはできないが、多くの場合、適当な長さの尺度として定まる。

【００３３】

本明細書において相関を最大にする方法として、正準相関分析（ＣＣＡ）、カーネル正準相関分析（ｋｅｒｎｅｌ　ＣＣＡ）、サポートベクターマシン（ＳＶＭ）法等の多変量解析手法及び機械学習法がある。

【００３４】

本明細書において「化学物質情報」とは、化学物質の化合物としての情報をいう。より詳細に定義すると、化学構造自体や化学構造から計算処理によって算出された各種記述子（文献　Ｈａｎｄｂｏｏｋ　ｏｆ　Ｃｈｅｍｏｉｎｆｏｒｍａｔｉｃｓ：Ｆｒｏｍ　Ｄａｔａ　ｔｏ　Ｋｎｏｗｌｅｄｇｅ　Ｇａｓｔｅｉｇｅｒ、Ｊｏｈａｎｎ（ＥＤＴ）／Ｐｕｂｌｉｓｈｅｒ：Ｊｏｈｎ　Ｗｉｌｅｙ　Ｐｕｂｌｉｓｈｅｄ　２００３／１０）、化学構造や記述子より計算推定される化学特性、さらには化合物を計測して得られる化学特性が含まれる。通常、各々の情報は数値または数値列（ベクトル）として表現される。各々の化合物は、適宜選択した各種化学物質情報の数値列を連結した数値列（ベクトル）として表現される。

【００３５】

通常、化学物質情報は、化合物記述子によって定義される。このような記述子としては、例えば、一次元記述子、二次元記述子および三次元記述子からなる群より選択されるものを挙げることができる。「記述子」とは、ディスクリプタともいい、ある情報を記述するための方法およびそれにより表された表現物をいう。記述子は、電波、磁波、音、光、色、画像、数字、文字など又はそれらの組み合わせによって表現することができる。分子構造を特徴づける記述子群の特定は、数多くの化合物を解析するプロセスで重要な工程である。数多くの記述子が提案されているが、分子構造へのアプローチに応じてこれらを分類することができる（Ｍ．Ｈａｓｓａｎ　ｅｔ　ａｌ．、Ｍｏｌｅｃｕｌａｒ　Ｄｉｖｅｒｓｉｔｙ、１９９６、２、６４；Ｍ．Ｊ．ＭｃＧｒｅｇｏｒ　ｅｔ　ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．、１９９９、３９、５６９；Ｒ．Ｄ．Ｂｒｏｗｎ、Ｐｅｒｓｐｅｃｔｉｖｅｓ　ｉｎ　Ｄｒｕｇ　Ｄｉｓｃｏｖｅｒｙ　ａｎｄ　Ｄｅｓｉｇｎ、１９９７、７／８、３１参照。以上は先に本明細書に参考文献として援用する。Ｒ．Ｄ．Ｂｒｏｗｎ　ｅｔ　ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．１９９６、３６、５７２；Ｒ．Ｄ．Ｂｒｏｗｎ　ｅｔ　ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．１９９６、３７、１；Ｄ．Ｅ．Ｐａｔｔｅｒｓｏｎ　ｅｔ　ａｌ．、Ｊ．Ｍｅｄ．Ｃｈｅｍ．１９９６、３９、３０４９；Ｓ．Ｋ．Ｋｅａｒｓｌｅｙ　ｅｔ　ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．１９９６、３６、１１８参照。以上を本明細書に参考文献として援用する）。１次元（１Ｄ）特性は、分子量やｃｌｏｇＰ等の全体的な分子特性を表す。２次元特性（２Ｄ）には、分子の機能性や結合性が含まれる。２Ｄ記述子の例としては、ＭＤＬサブストラクチャーキー（ＭＤＬ　Ｉｎｆｏｒｍａｔｉｏｎ　ＳｙｓｔｅｍｓＩｎｃ．、１４６００　Ｃａｔａｌｉｎａ　Ｓｔ．、Ｓａｎ　Ｌｅａｎｄｒｏ、ＣＡ　９４５７７）（Ｍ．Ｊ．ＭｃＧｒｅｇｏｒ　ｅｔ　ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．、１９９７、３７、４４３参照。これを本明細書に参考文献として援用する）やＭＳＩ５０記述子（Ｍｏｌｅｃｕｌａｒ　Ｓｉｍｕｌａｔｉｏｎｓ　Ｉｎｃ．、９６８５　Ｓｃｒａｎｔｏｎ　Ｒｏａｄ、Ｓａｎ　Ｄｉｅｇｏ、ＣＡ　９２１２１－３７５２）等が挙げられる。例えば、薬剤化合物に対する要件を特定する際に有用な、周知の５つの法則（ｒｕｌｅ　ｏｆ　ｆｉｖｅ）は、１次元記述子及び２次元記述子から導かれる（Ｃ．Ａ．Ｌｉｐｉｎｓｋｉ　ｅｔ　ａｌ．、Ａｄｖａｎｃｅｄ　Ｄｒｕｇ　Ｄｅｌｉｖｅｒｙ　Ｒｅｖｉｅｗｓ、１９９７、２３、３参照。これを本明細書に参考文献として援用する）。

【００３６】

３次元記述子（３Ｄ）の算出には、適度なエネルギーを有する少なくとも１つの３次元

構造体が必要である。更に、複数のコンホメーション（立体配座）からの寄与を考慮にいれて、３次元記述子を算出してもよい。また、リガンド結合において重要な特徴に基づいて、あるいは、その他の重要な所望の特徴に応じて、記述子を選択するようにしてもよい。あるいは、数多くの化合物群の解析に多数の記述子を用いる場合には、主成分分析（ＰＣＡ）や部分最小２乗法（ＰＬＳ）等の統計手法により最少数の重要な記述子群を求めればよい。

【００３７】

本明細書において「生物学的情報」とは、遺伝子やタンパク質などの生体物質である化学物質の生物学的特性にかかわる情報をいう。生物学的情報とは、例えば、遺伝子の塩基配列情報、一次構造情報、二次構造情報、三次構造情報、四次構造情報、立体構造情報、発現情報、パスウェイ情報、機能情報、および生物活性情報などが挙げられる。本発明においては、この生物学的情報は、計算処理や計測によって数値化した数値を各要素として持つ数値列（ベクトル）として表現される。

【００３８】

生体物質の特性は、その独自の記述子で生物学的特性を指標に表現することができる。従って、本明細書において、ある状態に関する「指標」とは、その状態を表すための目印となる関数をいう。本明細書では、例えば、生物または細胞であれば、その生物または細胞内の種々の物理的指標（電位、生体内温度、移動速度・距離、局在化率、扁平率、伸長率、回転速度など）、化学的指標（ゲノム量、特定の遺伝子の転写産物（例えば、ｍＲＮＡ、翻訳タンパク質、翻訳後修飾されたタンパク質、イオン濃度、ｐＨなどの量、代謝産物の量、イオン量など）、生物学的指標（個体差、進化速度、薬物応答など）など、あるいはその生物または細胞の環境、例えば、温度、湿度（例えば、絶対湿度、相対湿度など）、ｐＨ、塩濃度（例えば、塩全体の濃度または特定の塩の濃度）、栄養（例えば、ビタミン量、脂質量、タンパク質量、炭水化物量、金属イオン濃度など）、金属（例えば、金属全体の量または特定の金属（例えば、重金属、軽金属など）の濃度など）、ガス（例えば、ガス全体の量または特定のガス（例えば、酸素、二酸化炭素、水素など）の量）、有機溶媒（例えば、有機溶媒全体の量または特定の有機溶媒（例えば、エタノール、DMSO、メタノールなど）の量）、圧力（例えば、局所圧または全体の圧（気圧、水圧）など、粘性、流速（例えば、培地中に生物が存在する場合のその培地の流速、膜流動など）、光度（ある特定波長の光量など）、光波長（例えば、可視光のほか紫外線、赤外線なども含み得る）、電磁波、放射線、重力、張力、音波、対象となる生物とは異なる他の生物（例えば、寄生虫、病原菌、細菌、ウイルスなど）、化学薬品（例えば、医薬品、食品添加物、農薬、肥料、環境ホルモンなど）、抗生物質、天然物、精神的ストレス、物理的ストレスなどのような指標に対する反応性または耐性を、そのような状態に関する「指標」として使用することができる。

【００３９】

本明細書において「遺伝子」とは、遺伝形質を規定する因子をいう。通常染色体上に一定の順序に配列している。タンパク質の一次構造を規定する遺伝子を構造遺伝子といい、その発現を左右する遺伝子を調節遺伝子という。本明細書では、特定の状況において、「遺伝子」は、「ポリヌクレオチド」、「オリゴヌクレオチド」および「核酸」ならびに／あるいは「タンパク質」、「ポリペプチド」、「オリゴペプチド」および「ペプチド」をさすことがある。

【００４０】

本明細書に置いて「一次構造」とは、特定のペプチドのアミノ酸配列をいう。「二次構造」とは、ポリペプチド内の局所的に配置された三次元構造をいう。これらの構造はドメインとして一般に公知である。ドメインは、ポリペプチドの緻密単位を形成し、主に５０～３５０アミノ酸長であるポリペプチドの部分である。代表的なドメインは、β－シート及びα－ヘリックスである。「三次構造」とは、ポリペプチドモノマーの完全な三次元構造をいう。「四次構造」とは、独立した三次単位の非共有的会合により形成される三次元構造をいう。

【００４１】
本明細書において用いられる数理処理は、例えば、生命システム解析のための数学、コロナ社、清水和幸（１９９９）などにおいて記載される周知技術を適用することができる。
【００４２】
本明細書において「ライブラリ」とは、スクリーニングをするための化合物または生体物質などの化学物質の一定の集合をいう。ライブラリは、同様の性質を有する化合物の集合であっても、ランダムな化合物の集合であってもよい。好ましくは、同様の性質を有すると予測される化合物の集合が使用されるが、それに限定されない。
【００４３】
本明細書において「相互作用」とは、２以上の分子が存在する場合、ある分子と別の分子との間の作用をいう。そのような相互作用としては、水素結合、ファンデルワールス力、イオン性相互作用、非イオン性相互作用、受容体リガンド相互作用、静電的相互作用およびホストーゲスト相互作用が挙げられるがそれらに限定されない。分子間の相互作用を定量化する手法の１つとしては、分子間相互作用の熱力学的あるいは速度論的な定量評価法が挙げられるが、これらに限定されるものではない。分子間相互作用の熱力学的あるいは速度論的な定量評価法としては、例えば、熱測定（ｃａｌｏｒｉｍｅｔｏｒｙ）、表面プラズモン共鳴法、超遠心分析法などが挙げられるがこれらに限定されない。分子間相互作用は、複合体を形成した状態と解離した状態の熱力学量の変化を示す指標により表現可能であり、例えば、結合定数、解離定数、結合／解離に伴う標準化学ポテンシャル変化、エンタルピー変化、イオン結合数変化などにより表現可能である。
【００４４】
本明細書において、「相互作用情報」は、例えば、分子間における結合の有無、結合活性、薬理活性などにより表現されるが、これらに限定されるものではない。「薬理活性」は、一般に、薬物の効力を示す指標である。例えば、生理活性の５０％阻害効果を示す濃度であるＩＣ５０や生理活性の５０％亢進効果を示す濃度のＥＣ５０として与えられ、また、「結合の有無」および「結合活性」についても、例えば、解離定数Ｋｄとして与えられる。
【００４５】
本明細書において「情報処理」または「処理」とは、情報をより取り扱いやすくするため、一つの形式から他の形式へ変換または統合することをいう。また、「データ処理」とは、データを一つの形式から他の形式へと変換または統合する過程をいう。
【００４６】
本明細書において「パターン認識」とは、自然情報処理の１つである。例えば、形態、図形、物体、画像、音声、生理的現象のような単純な数量として与えられない情報を識別し認定することをいう。このような諸情報を、「パターン情報」又は、単に「パターン」という。パターン認識を行う識別器としては、例えば、ＳＶＭ（サポートベクターマシン）、ベイズ分類、ニューラルネットワークなど、機械学習により大量のデータから識別パラメータを構成する手法を用いることが可能である。本明細書において、「相互作用パターン」は、上述の識別器により統計モデルとして定義づけられる情報をいう。
【００４７】
本明細書において「モデル」とは、ある特定の条件に従う数学的、物理的な系をいい、物理、生物などの自然科学における系を理解するために用いられる。特に、統計的な解析の対象となる場合、そのモデルを「統計モデル」という。本明細書において使用される場合、ある現象の「モデル化」とは、データの背後にある現象の解明と予測、制御や新たな知識発見のために「モデル」を導入することをいう。
【００４８】
本明細書においては、集合Xの要素が任意に与えられたとき、関数ｆによって集合Yの要素がひとつ対応づけられていることを「集合Xから集合Yへの写像」と表現する。また、関数ｆによって集合Xから集合Yへと移行することを「写像変換」という。本発明においては、

集合は化学物質の集合を意味する。

【００４９】

また、本明細書における「マッピング」とは、トレーニングデータをCCAやPCAやカーネル化などの計算によって写像変換することによって、算出される重み係数行列やカーネル関数をテストデータにかけることでなされる。

【００５０】

本明細書におけるライブラリは、例えば、コンビナトリアル・ケミストリ技術、醗酵方法、植物および細胞抽出手順などの手法により、作製又は入手することができる。コンビナトリアル・ライブラリを作成する方法は、当該技術分野で周知である。例えば、Ｅ．Ｒ．Ｆｅｌｄｅｒ、Ｃｈｉｍｉａ １９９４、４８、５１２－５４１；Ｇａｌｌｏｐら、Ｊ．Ｍｅｄ．Ｃｈｅｍ．１９９４、３７、１２３３－１２５１；Ｒ．Ａ．Ｈｏｕｇｈｔｅｎ、Ｔｒｅｎｄｓ Ｇｅｎｅｔ．１９９３、９、２３５－２３９；Ｈｏｕｇｈｔｅｎら、Ｎａｔｕｒｅ １９９１、３５４、８４－８６；Ｌａｍら、Ｎａｔｕｒｅ １９９１、３５４、８２－８４；Ｃａｒｅｌｌら、Ｃｈｅｍ．Ｂｉｏｌ．１９９５、３、１７１－１８３；Ｍａｄｄｅｎら、Ｐｅｒｓｐｅｃｔｉｖｅｓ ｉｎ Ｄｒｕｇ Ｄｉｓｃｏｖｅｒｙ ａｎｄ Ｄｅｓｉｇｎ２、２６９－２８２；Ｃｗｉｒｌａら、Ｂｉｏｃｈｅｍｉｓｔｒｙ １９９０、８７、６３７８－６３８２；Ｂｒｅｎｎｅｒら、Ｐｒｏｃ．Ｎａｔｌ．Ａｃａｄ．Ｓｃｉ．ＵＳＡ １９９２、８９、５３８１－５３８３；Ｇｏｒｄｏｎら、Ｊ．Ｍｅｄ．Ｃｈｅｍ．１９９４、３７、１３８５－１４０１；Ｌｅｂｌら、Ｂｉｏｐｏｌｙｍｅｒｓ １９９５、３７ １７７－１９８；およびそれらで引用された参考文献を参照のこと。これらの参考文献は、その全体を、本明細書中で参考として援用する。

【００５１】

コンビナトリアル・ケミストリとハイスループット・スクリーニングの最近の発展に伴い、数多くの化合物に対する実験的アプローチが可能になった。例えば、Ｄ．Ｋ．Ａｇｒａｆｉｏｔｉｓ ｅｔ ａｌ．、Ｍｏｌｅｃｕｌａｒ Ｄｉｖｅｒｓｉｔｙ、１９９９、４、１；Ｕ．Ｅｉｃｈｌｅｒｅｔ ａｌ．、Ｄｒｕｇｓ ｏｆｔｈｅ Ｆｕｔｕｒｅ、１９９９、２４、１７７；Ａ．Ｋ．Ｇｈｏｓｅ ｅｔ ａｌ．、Ｊ．Ｃｏｍｂ．Ｃｈｅｍ．、１、１９９９、５５；Ｅ．Ｊ．Ｍａｒｔｉｎ ｅｔ ａｌ．、Ｊ．Ｃｏｍｂ．Ｃｈｅｍ．、１９９９、１、３２；Ｐ．Ｒ．Ｍｅｎａｒｄ ｅｔ ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．、１９９８、３８、１２０４；Ｒ．Ａ．Ｌｅｗｉｓ ｅｔ ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．、１９９７、３７、５９９；Ｍ．Ｈａｓｓａｎ ｅｔ ａｌ．、Ｍｏｌｅｃｕｌａｒ Ｄｉｖｅｒｓｉｔｙ、１９９６、２、６４；Ｍ．Ｊ．ＭｃＧｒｅｇｏｒ ｅｔ ａｌ．、Ｊ．Ｃｈｅｍ．Ｉｎｆ．Ｃｏｍｐｕｔ．Ｓｃｉ．、１９９９、３９、５６９；Ｒ．Ｄ．Ｂｒｏｗｎ、Ｐｅｒｓｐｅｃｔｉｖｅｓ ｉｎ Ｄｒｕｇ Ｄｉｓｃｏｖｅｒｙ ａｎｄ Ｄｅｓｉｇｎ、１９９７、７／８、３１を参照のこと。以上を本明細書に参考文献として援用する。このため、数多くの化合物に関する演算特性を解析する技術が、薬剤開発において、ますます重要になってきている。特定ライブラリ、すなわち、標的ライブラリの構築並びにプライマリ・ライブラリの構築という２つの適用例では、数多くの化合物に関する演算特性の解析により、薬剤設計にとって特に重要な情報を提供することができる。

【００５２】

以下に本発明を実施するための実施形態の説明を記載するが、この実施形態は本発明を実施するための単なる例示であり、本発明の範囲はそのような好ましい実施形態に限定されないことが理解されるべきである。

【００５３】

まず、本発明の実施形態について、図を用いて説明する。本発明者らの手法は、図１上段の化学物質情報のみを用いていた従来手法に、図１下段のバイオインフォマティクス技術を融合させたものであり、これはゲノム情報等に代表される生物学的情報を考慮に入れた新しい相互作用予測方法である。

【００５４】

リード化合物の探索やライブラリ設計を計算で行う際には、個々の化合物の類似度を反映するため、相対的な位置を示す座標空間が必要である。例えば、図１上段のケミカル空間上の丸印はそれぞれ異なる化合物を表しており、特性の似ている化合物は相対的に近い位置関係になるように配置されている。これら化合物の位置から構成される座標空間をケミカル空間という。

【００５５】

同様に、タンパク質又はその遺伝子について、類似関係を相対的な位置として表現したものが図１下段のバイオ空間である。図１下段のバイオ空間内の四角印がタンパク質又は遺伝子を表す。さらに、個々の化合物とタンパク質の結合をリンク（図１、中央矢印）することによって、ケミカル空間とバイオ空間を融合したモデルを作ることができる。

【００５６】

図２は、本発明の概念図であり、本発明の実施態様の一例を示したものである。ケミカル空間とバイオ空間との融合モデルから、クエリとなる化合物についての標的タンパクを予測することができる。まず、１）クエリの化合物（星印）の化学構造からその化合物がケミカル空間座標にマッピングされる（図２上段）。２）ケミカル空間にマッピングされた該化合物の近隣にある化合物からバイオ空間へのリンク情報をたどること（図２中央・矢印）により、その未知化合物が関係するバイオ空間内のエリア（図２下段・円内）を指定することができる。３）このエリア内のタンパク質群が、この活性未知の化合物が相互作用する可能性のあるタンパク群と推定される。

【００５７】

図３は、本発明の概念図であり、本発明の別の実施態様を示したものである。ケミカル空間とバイオ空間との融合モデルがあれば、図３中央の矢印で示すように、クエリのタンパク質から、該タンパク質に相互作用するリード化合物群を予測することも可能となる。

【００５８】

さらに、ケミカル空間とバイオ空間とを融合したモデルは化合物ライブラリの合理的設計にも適用できる。図４に示すように、広大なケミカル空間のうち、限定されたバイオ空間に対応するエリア（biologically relevant chemical space）内の化合物群が、バイオ空間を形成するタンパク群と相互作用する可能性が高いと考えられる。そのため、生物活性を有する化合物ライブラリを合理的に設計することができる。また、バイオ空間のタンパク質を、例えばGPCRファミリーなどに限定するとFocused libraryを設計することができる。

【００５９】

ケミカル空間とバイオ空間とを融合したモデルの構築方法について、本手法の特徴を示す。図５を参照しながら説明する。従来法（図５上段）は、化学物質情報のみを用い、化合物の化学特性ができる限り多様になるように、ケミカル空間座標を定義していた。しかし、化合物の多様性と生物活性との直接の因果関係がないという点が問題であった。そこで、本手法は、化学物質情報と生物学的情報との両情報を用いて、ケミカル空間とバイオ空間との両空間の相関が高くなるように互いの空間座標を定義することにした。これはバイオ空間との関連を考慮してケミカル空間座標を定義すると、化合物の生物活性を反映した空間座標を構築できるためである。

【００６０】

次に、本発明の実施形態をより詳細に説明する。第１の化学物質は化合物であり、第２の化学物質は、核酸、タンパク質又はそれらの複合体等の生体物質である。また、第１の空間は、ケミカル空間と呼ばれ、第２の空間は、バイオ空間と呼ばれる。化合物としては、任意の化合物のライブラリを用いることができる。例えば、コンビナトリアル・ライブラリ、ある企業が有する任意の化合物ライブラリ、公的機関が運営するデータベース、コンソーシアムによって運営されている化合物ライブラリ・データベースなどを挙げることができるがそれらに限定されない。

【００６１】

第1の特徴量は化学物質情報からなるベクトルとして表現され、第2の特徴量は生物学的

情報からなるベクトルとして表現される。これら第１の特徴量と第２の特徴量とは互いに単純な関連が見られないものであり得る。

【００６２】

本発明の空間座標の作成にはどのような情報を用いてもよい。１つの実施形態では、本発明における化合物の空間座標は化学物質情報から定義される。ここで、化学物質情報は、化合物記述子によって定義される。化合物記述子は、一次元記述子、二次元記述子および三次元記述子からなる群より選択される。一次元記述子は、化学組成を記述することを特徴とし、二次元記述子は、化学トポロジーを記述することを特徴とし、三次元記述子は、三次元形状および官能性からなる群より選択される特徴を記述することを特徴とする。

【００６３】

また、化合物記述子はファルマコフォアであってもよい。ファルマコフォアは、少なくとも３つの空間的に離れたファルマコフォア中心を含む。各ファルマコフォア中心は、（ｉ）空間位置と、（ｉｉ）ある化学特性を特定する所定のファルマコフォア型と、を含む。基本セットのファルマコフォア型には、少なくとも、水素結合受容体、水素結合供与体、負電荷中心、正電荷中心、疎水性中心、芳香族中心、ならびに他のいずれのファルマコフォアの型にも入らないデフォルトカテゴリが含まれる。また、空間位置を、隣接するファルマコフォア中心間の隔絶距離あるいは隔絶距離範囲として与えることによって、ファルマコフォアをより詳細に記述することができる。

【００６４】

化学物質は、化合物記述子、化学構造や記述子より計算推定される化学特性、さらには化合物を計測して得られる化学特性の数値を各要素として持つ数値列（ベクトル）として表現される。よって、各々の化学物質は第１空間座標上のベクトルとして位置が特定される。

【００６５】

本発明における生体物質の空間座標は、生物学的情報によって定義される。ここで、生物学的情報は、配列情報、二次構造、三次構造、四次構造、立体構造情報、発現情報、パスウェイ情報、および機能情報からなる群より選択される少なくとも１種類の情報を用いることができる。生体物質は、生物学的情報を計算処理や計測によって数値化した数値を各要素として持つ数値列（ベクトル）として表現される。よって、各々の生体物質は第２空間座標上のベクトルとして位置が特定される。

【００６６】

好ましい実施形態では、第１空間と第２空間と間の相関が高くなるように、より好ましくは相関が最大になるように、第１空間の特徴量及び第２空間の特徴量が写像変換される。ここで、第１空間と第２空間との相関が高くなるようにするには、正準相関（ＣＣＡ）、カーネル正準相関（kernel　CCA）、サポートベクターマシン（ＳＶＭ）法などの多変量解析や機械学習法またはそれらの等価方法により行うことができる。

【００６７】

正準相関分析（ＣＣＡ）とは、２種類の異質なデータセット、例えば、化合物とタンパク質が与えられたときに、そのデータセット間の相関関係を解析する多変量解析手法の一種である。ＣＣＡは、両データセット間の相関を最もよく表すように写像変換し、それによって２つのデータ間の相関を解析するものである。具体的手順を以下に示す。

【００６８】

正準相関分析の具体的手順は、以下のような物を例示することができる（Ｔ．　Ｗ．Ａｎｄｅｒｓｏｎ．　Ａｎ　Ｉｎｔｒｏｄｕｃｔｉｏｎ　ｔｏ　Ｍｕｌｔｉｖａｒｉａｔｅ　Ｓｔａｔｉｓｔｉｃａｌ　Ａｎａｌｙｓｉｓ．　Ｗｉｌｅｙ　＆　Ｓｏｎｓ、１９８４．、Ｈ．　Ｈｏｔｅｌｌｉｎｇ．　Ｒｅｌａｔｉｏｎｓ　ｂｅｔｗｅｅｎ　ｔｗｏ　ｓｅｔｓ　ｏｆ　ｖａｒｉａｔｅｓ．　Ｂｉｏｍｅｔｒｉｋａ、２８：３２１－３７７、１９３６．）。

【００６９】

行列Ｘ、Ｙの行には化学物質のエントリが、列には化学物質情報が並ぶ２種の異質なデ

ータ（例えば化合物とタンパク質）を行列Ｘ、Ｙ（第１空間が行列Ｘ、第２空間が行列Ｙ）と表現したとき、

【００７０】

【数１１１】

$$X = n \times p\text{行列} \qquad Y = n \times q\text{行列} \qquad n \geq p \geq q \geq 1$$

$$f = Xa \qquad g = Yb$$

$$\sigma_f^2 = \frac{1}{n-1} f'f \qquad \sigma_{fg} = \frac{1}{n-1} f'g \qquad \sigma_g^2 = \frac{1}{n-1} g'g$$

第１空間と第２空間の相関を最大にするために、

相関係数

【００７１】

【数１１２】

$$r_{fg} = \frac{\sigma_{fg}}{\sigma_f \sigma_g}$$

【００７２】

を最大にする係数ベクトルａ、ｂの組を探し、

ここで

【００７３】

【数１１３】

$$\sigma_f^2 = \sigma_g^2 = 1$$

【００７４】

、の条件付のとき、

【００７５】

【数１１４】

$$r_{fg} = \sigma_{fg}$$

【００７６】

を最大にするとき、

【００７７】

【数１１５】

$$r_{fg}$$

【００７８】

を正準相関、

【００７９】

【数１１６】

$$f, g$$

【００８０】

を正準変量と呼ぶ。より具体的には、正準相関解析において、XとYの特異値分解を行い、

【００８１】

【数１１７】

$$X = U_X D_X V'_X \qquad Y = U_Y D_Y V'_Y$$

$$U'_X U_Y = UDV'$$

Ｕ、Ｄ、Ｖを算出し、そのＵ、Ｄ、Ｖを用い、
【００８２】
【数１１８】

$$A=\sqrt{n-1}V_X D_X^{-1}U \qquad B=\sqrt{n-1}V_Y D_Y^{-1}V$$

$$F=\sqrt{n-1}U_X U \qquad G=\sqrt{n-1}U_Y V$$

を求め、
ただし、A、 B、F、 Gは、
【００８３】
【数１１９】

$$f_i=Xa_i \qquad g_i=Yb_i \qquad i=1,\ldots,q$$

$$A=\lfloor a_1,\ldots,a_q \rfloor \quad B=\lfloor b_1,\ldots,b_q \rfloor$$

$$F=\lfloor f_1,\ldots,f_q \rfloor \quad G=\lfloor g_1,\ldots,g_q \rfloor$$

$$F=XA \quad G=YB$$

であり、ここで、i=1から順番に相関の高いもの
【００８４】
【数１２０】

$$r_{fg}$$

【００８５】
を得ることができる。
【００８６】
カーネル正準相関分析（ｋｅｒｎｅｌ ＣＣＡ）とは、通常の正準相関分析にカーネル法を導入した手法であり、線形モデルに基づく正準相関分析に対して、非線形モデルに基づく相関分析がカーネル正準相関分析では可能である。上記正準相関分析で対象にした第１空間のＸと第２空間のＹとをそれぞれヒルベルト空間に写像したＸ’とＹ’について、正準相関分析を行う方法である（Ｓ．Ａｋａｈｏ、 Ａ ｋｅｒｎｅｌ ｍｅｔｈｏｄ ｆｏｒ ｃａｎｏｎｉｃａｌ ｃｏｒｒｅｌａｔｉｏｎ ａｎａｌｙｓｉｓ、Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｍｅｅｔｉｎｇ ｏｆ Ｐｈｙｃｈｏｍｅｔｒｉｃ Ｓｏｃｉｅｔｙ （ＩＭＰＳ）、 ２００１）。
【００８７】
サポートベクターマシン（ＳＶＭ）法とは、教師付き識別問題を解くための機械学習アルゴリズムである（文献Ｂ．Ｅ．Ｂｏｓｅｒ、Ｉ．Ｍ．Ｇｕｙｏｎ、ａｎｄ Ｖ．Ｎ．Ｖａｐｎｉｋ．Ａ ｔｒａｉｎｉｎｇ ａｌｇｏｒｉｔｈｍ ｆｏｒ ｏｐｔｉｍａｌ ｍａｒｇｉｎ ｃｌａｓｓｉｆｉｅｒｓ．Ｉｎ Ｄ．Ｈａｕｓｓｌｅｒ、ｅｄｉｔｏｒ、５ｔｈ Ａｎｎｕａｌ ＡＣＭ Ｗｏｒｋｓｈｏｐ ｏｎ ＣＯＬＴ、ｐａｇｅｓ １４４－１５２）。ＳＶＭでは、データを２種類に分類するために各データ点との距離が最大となる分離平面（超平面）を求めるマージン最大化という考え方を用いる特徴を有する。さらに、カーネル関数を用いてパターンを有限もしくは無限次元の特徴空間へ写像し、特徴空間上で線形分離を行う方法を取ることによって非線形分離問題にも優れた性能を示すという特徴も有する。ここで、教師付き分類問題を、タンパク質と化合物の結合予測に適用すると、タンパク質と化合物が結合するというクラスと結合しないというクラスを分類する識別器を作ることになる。この場合、文献や実験などで得られる既知のタンパク質と化合物の結合データを教師データとして用いることができる。
【００８８】

多変量解析手法とは、複数の変数（項目、属性、次元数）を持つデータ（多変量データ）を利用し、その変数間の相互の関係性をとらえるために使われる統計的手法の総称である。重回帰分析や判別分析、正準相関分析、主成分・因子分析、クラスター分析、多次元尺度法、フェース分析、数量化分析、コンジョイント分析などの手法がある。複雑なデータが持つ傾向や特徴を「要約」したり、結果に影響する相関関係を明らかにして「原因発見」や「推定・予測」を行ったり、また因果関係のモデル化などに有効である。ＣＣＡ、カーネルＣＣＡもまた、この多変量解析に該当する。

【００８９】

機械学習法とは、人工知能における研究課題の一つで、人間が自然に行っている学習能力と同様の機能をコンピュータで実現させるための技術・手法のことである。ある程度の数のサンプルデータ集合を対象に解析を行い、そのデータから有用な規則、ルール、知識表現、判断基準などを抽出する。

【００９０】

１つの実施形態では、本発明は、ある化学物質と相互作用すると予測される化学物質を選出するデータ処理方法を提供する。

【００９１】

この方法において用いられている「目的領域」とは、ケミカル空間内に算出される領域を指す。「標的領域」とはバイオ空間内に算出される領域を指す。

【００９２】

相互作用の予測対象のうちクエリ側がタンパク質である場合、すなわち標的タンパク質に作用する化合物を予測する場合について詳細に説明する。まず、タンパク質について任意の活性を選択する。標的タンパク質の特性を選択すると、それに対応する空間内の任意の標的領域を算出することができる。

【００９３】

標的タンパク質又は標的タンパク質群が与えられている場合、その標的タンパク質と配列や構造などが相同なタンパク質群を選出し、それらが占有する空間領域として標的領域を算出することができる。また、タンパク質に定義されている機能（例えば、遺伝子オントロジー（ｇｅｎｅ　ｏｎｔｏｌｏｇｙ）など）に基づいて標的領域を算出する場合は、標的タンパク質又は標的タンパク質群と同等の機能が定義されているタンパク質群を選出し、それらが占有する空間領域として標的領域を算出することができる。

【００９４】

さらには、タンパク質に定義されている遺伝子発現パターン、パスウェイ位置情報、生物活性情報（例えば、マイクロアレイデータ、反応経路、薬理活性など）に基づいて標的領域を算出する場合は、標的タンパク質又は標的タンパク質群と同等の遺伝子発現パターン、パスウェイ位置情報、生物活性情報が定義されているタンパク質群を選出し、それらが占有する空間領域として標的領域を算出することができる。

【００９５】

算出された標的領域から所定の距離以下に存在するケミカル空間の目的領域を算出する工程を説明する。まず標的領域内の各々のタンパク質に対し結合し得る化合物群が特定される。特定された化合物各々について、ケミカル空間内で所定の距離以下に存在する目的領域を算出する。ここでの距離とは、ユークリッド距離、マンハッタン距離などの距離の公理を満たすものから、相関係数やカーネルなどの類似度指標も含む。

【００９６】

ケミカル空間の目的領域に存在する化合物が選出される工程では、算出された目的領域に対応する化合物が選出される。これは、目的領域が算出されていれば、自動的な計算によって選出することも可能である。

【００９７】

また、他の実施形態では、相互作用の予測対象のうちクエリ側が化合物である場合、すなわち目的の化合物に作用するタンパク質を予測する方法を提供する。

【００９８】

１つの実施形態では、本発明は、サンプルデータを用いて第１空間と第２空間とを相関させるようトレーニングすること工程をさらに包含する。本明細書において「トレーニング」とは、装置の使用のための訓練に要する計算機操作で、取付け操作、操作卓操作、変換操作、印刷操作のような活動や、必要なデモンストレーションを行なうのに使われる操作をいう。本明細書において「トレーニングデータ」とは、操作の始めにロボットのコンピュータへ入力される練習データをいう。

【００９９】

１つの実施形態において、トレーニングは、直交行列Ａ＝$C_{xx}^{-1/2}$ＵとＢ＝$C_{yy}^{-1/2}$Ｖを生成する（ここで、ｄｅｔ（Ａ）＝ｄｅｔ（Ｂ）＝１かつ数１３１で示される数式の通りである）。

【０１００】

【数１３１】

$$C_{xx}=E[(X-m_x)(X-m_x)^T],\ C_{yy}=E[(Y-m_y)(Y-m_y)^T],\ C_{xy}=E[(X-m_x)(Y-m_y)^T],\ K=C_{xx}{}^{-1/2}\cdot C_{xy}\cdot C_{yy}{}^{-1/2}=U\cdot S\cdot V^T$$

【０１０１】

第１モダリティの第１空間を表すＡＸと第２モダリティの第２空間を表すＢＹとの間の相関は最大となり、これにより、該第１モダリティから該第２モダリティへの特徴の移転が可能となることを特徴としてもよい。

【０１０２】

別の実施形態では、前記トレーニングにより、行列Ａと行列Ｂを生成する。第１モダリティの第１空間を表わすＸＡと第２モダリティの第１空間を表わすＹＢとの間の相関は最大となり、これにより、第１モダリティから第２モダリティへの特徴の移転が可能となる。特徴の移転は、行列X、Yの行には化学物質のエントリが、列には化学物質情報が並ぶ2種の異質なデータ（例えば化合物とタンパク質）を行列Ｘ、Ｙ（第１空間が行列X、第２空間が行列Y）と表現し、正準相関解析を行って両空間の相関を最大にすることができる。

【０１０３】

本明細書において「モダリティ」とは、特徴的属性をいう。１つの実施形態において、第１空間を表すＡＸのクエリは、第２空間を表すＢＹの前記クエリの結果のみが与えられると、ＢＹはＡＸと最大の相関を有することから特定可能である。

【０１０４】

また、本発明は、第１の化学物質の第１の特徴量と第２の化学物質の第２の特徴量とを多変量解析手法又は機械学習法によって１つの特徴空間へ写像変換するデータ処理方法を提供する。機械学習法としてはＳＶＭ法を用いることができる。

【０１０５】

また、具体的な実施形態では、機械学習法において、１）相互作用予測の対象となる第１化学物質の特徴量と第２化学物質の特徴量とを、第１の化学物質の第１の特徴量の空間座標のデータベースおよび第２の化学物質の第２の特徴量の空間座標のデータベースによって構築された特徴空間にマッピングする工程；２）該問い合わせペアが、空間エリア内に存在する場合、第１化学物質と第２化学物質とが結合すると判定し、空間エリア内に存在しない場合、第１化学物質と第２化学物質とが結合しないと判定する工程を包含する。

【０１０６】

特徴空間にマッピングする特徴量は、第１の化学物質の特徴量と第２の化学物質の特徴量とを連結した特徴量であってもよい。また、第１の化学物質の特徴量及び第２の化学物質の特徴量はベクトルで表現されるが、連結された特徴量はベクトル同士の内積であるカーネルを含む。

【０１０７】

１つの実施形態では、本発明の方法は、さらに、選出された化学物質をインシリコで生産する工程を包含する。インシリコでの生産方法は、本明細書において別の場所において

記載されており、周知の技術を用いることができる。

【０１０８】

別の実施形態では、本発明の方法は、さらに、ウェットで選出された化学物質を生産する工程を包含する。ウェットでの生産方法は、本明細書において別の場所において記載されており、周知の技術を用いることができる。ウェットでの生産の代表例としては、コンビナトリアル・ケミストリを用いることがあり得る。ウェットでの生産は、遺伝子組み換え技術を用いて達成されてもよい。

【０１０９】

１つの具体的な実施形態では、本発明の方法は、さらに、前記第１空間の化学物質の選出の後、該第１空間の化学物質の前記第２の特徴量を測定して、実際に所望の活性を有する化学物質を選出する工程をさらに包含する。

【０１１０】

本発明において、第１の化学物質と第２の化学物質とが相互作用するか否かが判定される工程には、相互作用する確率を算出する工程が含まれる。また、相互作用する確率はスコアとして算出されてもよい。ここでいうスコアとは、第１空間内の任意の化学物質Ａと第２空間内の任意の化学物質Ｂの結合のしやすさ（結合予測の統計的有意性）を示すものである。

【０１１１】

例えば、化学物質Ａと化学物質Ｂとの結合スコアは次のように定義できる。第２空間内の全ての化学物質数をＮ、そのうち化学物質Ａと結合する個数がＬあったとする。ここで、Ｂに近接する第２空間内の化学物質Ｋ個を考えた場合、そのうち化学物質Ａと結合する個数がＨであったとする。その際の化学物質ＡからＢの結合スコアはｌｏｇ（Ｈ／Ｋ）／（Ｌ／Ｎ））というようなオッズスコアとして定義できる。また逆に、第１空間内の全ての化学物質数をｎ、そのうち化学物質Ｂと結合する個数がｌあったとする。ここで、Ａに近接する第１空間内の化学物質ｋ個を考えた場合、そのうち化学物質Ｂと結合する個数がｈであったとする。その際の化学物質ＢからＡの結合スコアはｌｏｇ（ｈ／ｋ）／（ｌ／ｎ））というようなオッズスコアとして定義できる。従って、化学物質ＡからＢへのスコアとＢからＡへのスコアから、化学物質ＡとＢの総合スコアをｌｏｇ（Ｈ／Ｋ）／（Ｌ／Ｎ））＋ｌｏｇ（ｈ／ｋ）／（ｌ／ｎ））と定義することができる。これらのスコアはＣＣＡやｋｅｒｎｅｌ　ＣＣＡで算出された第１空間と第２空間の相関モデルから算出される。また、ＳＶＭ法を用いた場合は、第１空間内の化学物質群と第２空間内の化学物質群との間の既知の結合ペアと結合しないペアを分離する超平面からの距離に相当するものからスコアを換算することができる。

【０１１２】

「相互作用情報」には、解離定数Ｋｄ、５０％阻害効果濃度ＩＣ５０、５０％亢進効果濃度ＥＣ５０などが挙げられる。医薬品開発の場合、結合の有無、結合活性、薬理活性の基準として、Ｋｄ、ＩＣ５０、ＥＣ５０が、マイクロモルオーダー、ナノモルオーダーであると望ましい。また、相互作用する確率のスコアは、特徴空間内でクラス分類（例えば、結合する化合物とタンパク質ペアのクラスと結合しない化合物とタンパク質ペアのクラスなど）をする境界面（超平面）からの予測対象までの距離で表される。境界面から遠距離にあるほど、相互作用する確率が高くなる。

【０１１３】

本発明は、本発明のデータ処理方法により得られた結果から設計されたライブラリを提供する。また、本発明は、本発明の方法を用いることによって生産された化学物質を提供する。

【０１１４】

また、本発明は、本発明のデータ処理方法をコンピュータに実行させるプログラムを提供する。本発明のプログラムは、コンピュータ読み出し可能な記録媒体に格納される。記録媒体としては、プログラムを記録することができる限り、任意の形態（例えば、フレキシブルディスク、ＭＯ、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＤＶＤ－ＲＯＭのような任意のタイプ

）を使用することができることが理解される。さらに、本発明のデータ処理方法に用いられるデータ構造物は、記録媒体に格納される。具体的には、第１の化学物質の空間座標のデータベースにより定義される第１空間と第２の化学物質の空間座標のデータベースにより定義される第２空間とを備えるデータ構造物や、本発明の方法により構築された特徴空間を備えるデータ構造物などが含まれる。

【０１１５】

本発明は、特定の標的タンパク質に活性を持つ化合物を探索する目的に適用可能のほか、特定の化合物が与えられたときにその化合物に作用される複数のタンパク質の推定にもつながり、薬物の副作用に関する知見を提供する。また、特定の化合物に作用できる人工タンパク質（遺伝子改変タンパク質）の創製などへの応用も可能であると考えられる。

【０１１６】

また、本発明のデータ処理方法を用いて第１の化学物質と第２の化学物質との間の相互作用を予測するデータ処理装置が提供される。また、別の実施形態において、特定の化学物質と相互作用すると予測される化学物質を選出するデータ処理装置が提供される。さらに、別の実施形態において化学物質ライブラリを設計するデータ処理装置が提供される。これらの処理装置は、その処理装置に備え付けられた演算装置上で本発明の方法が実行されることにより実施される。

【０１１７】

本明細書において引用された、科学文献、特許、特許出願などの参考文献は、その全体が、各々具体的に記載されたのと同じ程度に本明細書において参考として援用される。

【０１１８】

以上、本発明を、理解の容易のために好ましい実施形態を示して説明してきた。以下に、実施例に基づいて本発明を説明するが、上述の説明および以下の実施例は、例示の目的のみに提供され、本発明を限定する目的で提供したのではない。従って、本発明の範囲は、本明細書に具体的に記載された実施形態にも実施例にも限定されず、特許請求の範囲によってのみ限定される。

【実施例】

【０１１９】

以下に実施例を示して本発明をさらに詳しく説明するが、この発明は以下の例に限定されるものではない。

【０１２０】

（実施例１：ＣＣＡを用いたバイオ空間およびケミカル空間でのスクリーニング）

従来法（ＰＣＡ）と本手法（ＣＣＡ）の性能評価をするために、それぞれで構築した「ケミカル空間とバイオ空間の融合モデル」を用いてインシリコスクリーニング（Ｉｎ　ｓｉｌｉｃｏ　ｓｃｒｅｅｎｉｎｇ）を行った。

【０１２１】

図６は、予測性能を評価する方法の一つであるＲＯＣ曲線である。このグラフは曲線が上に位置すれば位置するほど、予測性能が良いことを表すものであり、本手法の曲線が従来法の曲線より上に位置することから、本手法の予測性能が、従来法の予測性能に比べて高いことがわかる。以下にその具体的手順を示す。

【０１２２】

（プロトコール）

１、融合モデルの構築に用いる既知のタンパク質と化合物のデータは、ＤｒｕｇＢａｎｋデータベース（ｈｔｔｐ：／／ｒｅｄｐｏｌｌ．ｐｈａｒｍａｃｙ．ｕａｌｂｅｒｔａ．ｃａ／ｄｒｕｇｂａｎｋ／）２００５年８月リリース版から取得した。

２、全ての化合物エントリのｍｏｌ　ｆｉｌｅについて、ＤｒａｇｏｎＸソフトウェアを用いて、９３７個の化合物記述子を算出した。ここで、計算された化合物数は３０７９個である。さらに、ＣＣＡ計算を行うにあたり、属性となる記述子のプロファイルは独立していなければならないため、相関係数０．８以上の相関性を持つ記述子は、情報量の高い記述子３００個に縮約した。

３、全てのタンパク質エントリのｆａｓｔａ　ｆｉｌｅについて、ｍｉｓｍａｔｃｈ　ｓｔｒｉｎｇ　ｋｅｒｎｅｌを生成する手法と同様の手法により、ミスマッチを考慮した連続する２アミノ酸の組成比からなる４００次元（アミノ酸２０種＊２０種）のプロファイルを算出した。ここで、プロファイル化されたタンパク質数は３４７６個である。また、上記２で算出された化合物との結合数は８００６個であった。
４、予測性能の評価には、５分割交差検証法（５　ｆｏｌｄ　ｃｒｏｓｓ　ｖａｌｉｄａｔｉｏｎ）を用いた。すなわち、上記２と３で作成した８００６個の化合物－タンパク質結合データを無作為に４：１に分類し、８０％の結合データをトレーニングデータとしてＣＣＡ計算やＰＣＡ計算をし、バイオ空間とケミカル空間のそれぞれの座標を構築した。残りの２０％の結合データはテストデータとして用いた。ここで、テストデータの負例（結合しないデータ）は、正例である結合データを構成する化合物とタンパク質の組合せで結合しない組合せを発生させ、正例と同数を無作為に選出した。このように作成したテストデータを、トレーニングデータによって構築したバイオ空間とケミカル空間とにそれぞれマッピングした。マッピングとは、トレーニングデータをＣＣＡやＰＣＡ計算することによって、算出される重み係数行列（ＰＣＡの場合は主成分得点係数行列）をテストデータ行列にかけることでなされる。バイオ空間とケミカル空間とにそれぞれマッピングされたタンパク質と化合物について、各々についてスコアを算出した。
５、上記４のようにテストデータを予測したとき、実際に結合するデータを結合すると予測できたものの比率を真陽性率、実際には結合しないデータを結合すると予測してしまったものの比率を偽陽性率と呼ぶ。ここで、特定の予測スコア（閾値）以上の値を持つデータは陽性とみなし、特定スコア以下の値を持つデータは陰性とみなす。上記ＣＣＡとＰＣＡにおいて予測した化合物－タンパク質結合スコアに基づいて、スコアの閾値を動かし、それに伴う偽陽性率と真陽性率を（ｘ、ｙ）としてプロットし、ＲＯＣ曲線を作成した（図６）。

【０１２３】

（実施例２：化合物－ＧＰＣＲ相互作用予測手法）

化合物－ＧＰＣＲ相互作用予測手法の開発を以下の手順で行った。なお、本実施例にあたり用いたデータセットおよび解析方法等に関する詳細を記載した文献等については、本文中に参照番号を付し、本実施例の末尾にその参考文献一覧を添付した。これらの文献は、本明細書中で参考として援用される。

【０１２４】

（＃１　化合物－タンパク質相互作用情報の収集）

ＧＬＩＤＡ（ＧＰＣＲ－ＬＩｇａｎｄ　ＤＡｔａｂａｓｅ）［１］、ＤｒｕｇＢａｎｋ［２］、ＩＵＰＨＡＲ　Ｒｅｃｅｐｔｏｒ　ｄａｔａｂａｓｅ［３］、ＰＤＳＰ　Ｋｉ　ｄａｔａｂａｓｅ［４］から、相互作用する化合物－ＧＰＣＲの組み合わせ５２０７例（化合物：８６６、ＧＰＣＲ：３１７）を収集した。ただし、ここではヒト、マウス、ラットのＧＰＣＲを用い、ＧＰＣＲの定義はＧＰＣＲＤＢ［５］に従った。また、化合物については、続く記述子（ｄｅｓｃｒｉｐｔｏｒ）の計算に構造情報が必要であるため、ｍｏｌ（ｓｄｆ）　形式のファイルが提供されているＧＬＩＤＡおよびＰｕｂＣｈｅｍ　Ｃｏｍｐｏｕｎｄ［６］に登録されている化合物を用いた。

【０１２５】

（＃２　　記述子の計算）

化合物およびタンパク質を特徴ベクトルとして表現するために、以下の方法によりそれぞれの記述子を計算した。

・化学記述子（ｃｈｅｍｉｃａｌ　ｄｅｓｃｒｉｐｔｏｒ）

収集した化合物の構造から、化合物の構造・物性に関する記述子をＤＲＡＧＯＮＸ　ｖｅｒ．１．２［７］により計算した。この研究では、カテゴリー１－１０（ｃｏｎｓｔｉｔｕｔｉｏｎａｌ　ｄｅｓｃｒｉｐｔｏｒｓ、ｔｏｐｏｌｏｇｉｃａｌ　ｄｅｓｃｒｉｐｔｏｒｓ、ｗａｌｋ　ａｎｄ　ｐａｔｈ　ｃｏｕｎｔｓ、ｃｏｎｎｅｃｔｉｖｉｔｙ　ｉｎｄｉｃｅｓ、ｉｎｆｏｒｍａｔｉｏｎ　ｉｎｄｉｃｅｓ、２Ｄ　ａｕｔｏｃｏｒｒｅｌ

ａｔｉｏｎｓ、ｅｄｇｅ ａｄｊａｃｅｎｃｙ ｉｎｄｉｃｅｓ、Ｂｕｒｄｅｎ ｅｉｇｅｎｖａｌｕｅ ｄｅｓｃｒｉｐｔｏｒｓ、ｔｏｐｏｌｏｇｉｃａｌ ｃｈａｒｇｅ ｉｎｄｉｃｅｓおよびｅｉｇｅｎｖａｌｕｅ－ｂａｓｅｄ ｉｎｄｉｃｅｓ）、カテゴリー１７－１８（ｆｕｎｃｔｉｏｎａｌ ｇｒｏｕｐ ｃｏｕｎｔｓおよびａｔｏｍ－ｃｅｎｔｅｒｅｄ ｆｒａｇｍｅｎｔｓ）、カテゴリー２０（ｍｏｌｅｃｕｌａｒ ｐｒｏｐｅｒｔｉｅｓ）の計９２９記述子を計算した。なお、分子の三次元座標に依存する記述子（カテゴリー１１－１４）、官能基や原子タイプの数を数える記述子（カテゴリー１５および１６）、電荷記述子（カテゴリー１９）は、ここでは用いなかった。続いて、これらの記述子のうち、すべての化合物で同一の値として計算出力されるものを取り除き、結果として残った７９７種類の記述子を以下で用いた。

・タンパク質記述子（ｐｒｏｔｅｉｎ ｄｅｓｃｒｉｐｔｏｒ）

ミスマッチを許容したスペクトラム法［８］により計算した。この方法は、タンパク質配列を固定長ｋのアミノ酸配列に分解し、この中に現れる、最大ｍ個のミスマッチまで許容した長さｋのアミノ酸配列パターンの頻度を数えることにより計算される。発明者らは（ｋ、ｍ）を（２、１）に設定した。したがって、計算される記述子は、１アミノ酸のミスマッチを許容した２連アミノ酸２０２種類となる。

【０１２６】

（＃３　サポートベクターマシン（ＳＶＭ）による学習モデルの構築）

ＳＶＭは、Ｖａｐｎｉｋら［９］により提案された学習アルゴリズムであり、その高い汎化能力から各方面において多用されている。ＳＶＭは、２つの異なるグループの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。

【０１２７】

発明者らは、化合物－タンパク質相互作用の有無を分離する超平面を求めるために、正例（相互作用するパターン）および負例（相互作用しないパターン）に対応する化学記述子、タンパク質記述子をそれぞれ組み合わせて特徴ベクトルを構築し、ＳＶＭを用いて学習モデルを構築した。ただし、負例については、相互作用しないパターンの情報が得られないため、２つの記述子をランダムに組み合わせて正例と同数を生成した。ここで、ＳＶＭライブラリとして、「Ｓｅｑｕｅｎｔｉａｌ Ｍｉｎｉｍａｌ Ｏｐｔｉｍｉｚａｔｉｏｎ」アルゴリズム［１０、１１］を採用しているｌｉｂｓｖｍプログラム［１２］のコードを用いた。ＳＶＭモデルが得られると、新しいベクトル（化合物－タンパク質ペア）が、相互作用有／相互作用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。さらに、判別だけでなく、サンプルのスコアリングを行う方法も報告されている［１３］。これは、分離面に近いサンプルは、分離面から遠いサンプルよりも、誤って分類される確率が高いであろうという考えに基づいている。発明者らは、化合物－タンパク質間の相互作用予測において、この方法により、その可能性のスコア化、および順位付けを行った。

【０１２８】

（＃４　化合物構造類似性によるリガンド予測）

発明者らは、モデルの比較対象となるリガンド予測方法として、化学記述子から計算される化合物の類似性を用いた。この類似性は、一般的な化合物探索の方法であり、リード化合物を発見する手助けになると言われている［１４］。この研究では、上述の「＃２」で計算した７９７種類の化学記述子を主成分分析し、主成分座標上で既知リガンドと隣接する化合物から順にスコア付けを行った。化合物Ａ－タンパク質Ｂペアのスコアは、主成分空間において化合物Ａから見て最近傍にあるタンパク質Ｂ既知リガンドとの類似度で表現される。主成分は、累積寄与率８０％までのもの（３０主成分）を用いた。また、類似度の尺度として相関係数（Ｐｅａｒｓｏｎ ｃｏｒｒｅｌａｔｉｏｎ ｃｏｅｆｆｉｃｉｅｎｔ）を用いた。

【０１２９】

（＃５交差検証法（ｃｒｏｓｓ ｖａｌｉｄａｔｉｏｎ）によるモデルの評価）

ｎ分割交差検証法（ｎ－ｆｏｌｄ ｃｒｏｓｓ－ｖａｌｉｄａｔｉｏｎ）を用いて学習

モデルの予測性能を評価した。この評価法では、最初に全学習データセットがｎ個の等サイズなサブセットに分割される。続いて、それぞれのサブセットについて、残りのｎ－１サブセットで学習して作られた分類器を用いて予測する。そして、この操作は、すべてのサブセットが一度だけ予測されるように繰り返されて評価される。予測性能の尺度としては、以下の式で計算される正確度（Ａｃｃｕｒａｃｙ）を用いた。正確度は以下の式で表される。

【０１３０】

Ａｃｃｕｒａｃｙ　＝　（　ＴＰ　＋　ＴＮ）／（ＴＰ　＋　ＴＮ　＋ＦＰ＋ＦＮ）

ここで、ＴＰは真陽性、ＴＮは真陰性、ＦＰは偽陽性、ＦＮは偽陰性を表す。

【０１３１】

ランダムな組み合わせで作られる負例によるスコア変動を考慮し、負例を交換しながら１０回の異なるデータセットを生成して５分割交差検証法（５－ｆｏｌｄ　ｃｒｏｓｓ－ｖａｌｉｄａｔｉｏｎ）を繰り返し行い、正確度の平均値により発明者らのモデルを評価した。続いて、計算された相互作用予測スコアからＲＯＣ分析を行った。ここで、各評価において、化合物の構造類似性に基づいたリガンド予測法（＃４）を比較対象とした。

【０１３２】

（＃６　ヒトβ２アドレナリン受容体（β２ＡＲ）のリガンド予測）

発明者らが収集した化合物－ＧＰＣＲ相互作用情報は、今までの研究により「強く結合する」と知られているもののみであり、その他の大部分の化合物－ＧＰＣＲ相互作用は不明である。発明者らの疑問は、リガンド探索において、予想に反して相互作用すると予測された化合物が本当に相互作用しないかどうかということである。そこで、発明者らは、インビトロ（ｉｎ　ｖｉｔｒｏ）結合阻害実験により、相互作用予測スコアと相互作用の有無との関連性を確認した。

【０１３３】

そのために、ヒトβ２ＡＲを標的タンパク質とし、作成した学習モデルを用いてリガンド予測を行った。この受容体は、喘息治療の標的として治療薬の開発が進められている生理学的に重要な遺伝子である。リガンド予測の対象化合物は、ＧＰＣＲとの相互作用が知られている上記８６６化合物（ただし、モデル構築時にヒトβ２ＡＲとの相互作用を学習した化合物は除く）とした。これらの化合物の化学記述子に対してβ２ＡＲのタンパク質記述子（ｐｒｏｔｅｉｎ　ｄｅｓｃｒｉｐｔｏｒ）を組み合わせ、予測用データセットとした。負例組み合わせによるスコア変動を考慮し、負例を交換しながら学習と予測の試行を３０回繰り返し、各リガンドについて、得られたスコアの最大値を最終的な予測スコアとした。

【０１３４】

次に、相互作用予測スコア上位５０（Ｔｏｐ　５０）の化合物について、さらなる調査・実験を行った。まず、文献・特許調査（ＳｃｉＦｉｎｄｅｒ、ＰｕｂＭｅｄ）により、β２ＡＲとの相互作用に関する報告が存在しないか確認した。

【０１３５】

続いて、相互作用情報を確認できなかった化合物のうち、入手可能な化合物について、インビトロ（ｉｎ　ｖｉｔｒｏ）結合阻害実験による検証を行った。この実験では、ヒトβ２ＡＲ強制発現細胞株から膜画分を調製し、放射性β２ＡＲリガンドである［^{125}Ｉ］－シアノピンドロールに対する競合的な阻害効果を確認した。

【０１３６】

ところで、相互作用しないという情報の欠如により、発明者らのモデルでは、ランダムに発生させた化合物－タンパク質ペアを負例（相互作用なし）パターンとして採用している。このため、相互作用予測スコアの低い化合物が本当に相互作用しないか、ということを確認する必要がある。そこで、発明者らは、予測スコア下位５０（Ｂｏｔｔｏｍ　５０）の化合物についても、上位５０（Ｔｏｐ　５０）と同様の文献調査・検証実験を行った。

【０１３７】

（結果）
（交差検証法による新規リガンド予測モデルの評価）
まず、手始めに今回開発した方法と従来法との比較検討を行った。公共データベースから収集したＧＰＣＲ－リガンド相互作用情報を用い、化合物－タンパク質相互作用パターンの特徴ベクトルをＳＶＭ分類器の入力とし、学習モデルを構築した。負例を交換しながら５分割交差検証法を１０回試行した結果、発明者らが開発したモデルの予測性能（ａｃｃｕｒａｃｙ）は９１．３％±０．３％だった。対照として、化合物類似度に基づいた従来法についても同様に５分割検証法を行ったところ、予測性能は８１．９±０．３％だった。また、ＲＯＣ曲線からも、発明者らの開発したモデルの予測性能が高いことが判明した（図７）。

【０１３８】
（ヒトβ２ＡＲリガンド予測）
次に、新規手法をヒトβ２ＡＲの新規リガンド予測に適用し、その有効性を実験により検証した。また、新規手法でのみ予測されるリガンドが従来法では検出できないような新規骨格を持つ化合物を含むかどうか調べた。構築したモデルを用いて、８６６種類のＧＰＣＲリガンドについてβ２ＡＲとの相互作用予測スコアを算出した。

【０１３９】
新規モデルが予測したβ２ＡＲリガンド候補Ｔｏｐ　５０の化合物のうち、文献・特許調査により１４種の化合物がβ２ＡＲとの相互作用に関する報告を確認した（図８　（Ｂ－１）左）。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種類についてインビトロ（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＜ＩＣ５０＜10^{-3}Ｍ）を示した（図８　（Ｂ－１）右）。実験のヒット率は８１％（１７／２１）にのぼり、ここにおいても高い予測的中率が示された。

【０１４０】
一方、下位５０（Ｂｏｔｔｏｍ　５０）の化合物については、β２ＡＲリガンドとして報告されているものは文献および特許調査では確認されなかった（図８　（Ｂ－２）左）。さらに、残りのうち入手可能な９化合物についてインビトロ（ｉｎ　ｖｉｔｒｏ）結合実験を行ったところ、２個の化合物が同程度の強さの相互作用を示したが、残りの７化合物は相互作用を示さなかった（図８　（Ｂ－１）右）。

【０１４１】
これらの予測結果を従来法によるものと比較した図が図９である。実験で相互作用を確認した化合物の半数近くは、化合物の構造類似性に基づく従来の方法ではスコアが低かった。実際に、これらの化合物は、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）および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（図９左）を持っており、化合物の構造類似性に基づく従来の方法では発見できないリガンド群であるといえる。すなわち、相互作用情報に基づく新しいモデルは、多様な構造を持つ化合物が同一タンパク質に作用するという関係を正しく予測することができたといえる。また、これらの化合物の中には、ニューロペプタイド受容体アンタゴニストなど、従来はペプチド受容体に作用する化合物として知られていたものも含まれていたが、遠縁にあたるβ２ＡＲとも相互作用することが実験により確認された。

【０１４２】
（実施例３：化合物ライブラリの設計）
本手法を用いて、化合物ライブラリの設計、すなわち特定の化合物に対する標的遺伝子の予測を行った。図１０に設計手法を示す。

【０１４３】
予測は、米国ＮＣＢＩ／ＰｕｂＣｈｅｍデータベース内の化合物６、３９１、００５件を用いて、それらの化合物が標的とし得るタンパク候補を予測し、化合物ライブラリを構築した。ここで、予測の基準座標となるケミカル空間とバイオ空間との融合モデルの作成には、薬物とその標的タンパク質のデータを蓄積したカナダのＤｒｕｇＢａｎｋデータベースを用いた。

【０１４４】
図１１は、ＰｕｂＣｈｅｍ化合物の生物活性予測の結果である。各列は化合物と標的タンパクの結合可能性の信頼性を表すスコアごとに分かれている。すなわちスコアが高ければ高いほど、その化合物の生物活性の信頼性は高いと考えられる。ここで言うスコアは、第１空間内の任意の化学物質Ａと第２空間内の任意の化学物質Ｂの結合のしやすさ（結合予測の統計的有意性）を示すものと考えられる。
【０１４５】
例えば、化学物質Ａと化学物質Ｂとの結合スコアは次のように定義できる。第２空間内の全ての化学物質数をＮ、そのうち化学物質Ａと結合する個数がＬあったとする。ここで、Ｂに近接する第２空間内の化学物質Ｋ個を考えた場合、そのうち化学物質Ａと結合する個数がＨであったとする。その際の化学物質ＡからＢの結合スコアはｌｏｇ（Ｈ／Ｋ）／（Ｌ／Ｎ））というようなオッズスコアとして定義できる。また逆に、第１空間内の全ての化学物質数をｎ、そのうち化学物質Ｂと結合する個数がｌあったとする。ここで、Ａに近接する第１空間内の化学物質ｋ個を考えた場合、そのうち化学物質Ｂと結合する個数がｈであったとする。その際の化学物質ＢからＡの結合スコアはｌｏｇ（ｈ／ｋ）／（ｌ／ｎ））というようなオッズスコアとして定義できる。図１１では化学物質ＡからＢへのスコアとＢからＡへのスコアとから、化学物質ＡとＢの結合可能性の総合スコアをｌｏｇ（（Ｈ／Ｋ）／（Ｌ／Ｎ））＋ｌｏｇ（（ｈ／ｋ）／（ｌ／ｎ））＋２０と定義した。
【０１４６】
また、各行の項目は標的タンパクの機能（遺伝子オントロジー（ｇｅｎｅ　ｏｎｔｏｌｏｇｙ）に基づく）ごとの分類を表している。表中の数値は、該当する部分に対応する（予測された）化合物の数である。例えば、ｒｅｃｅｐｔｏｒ　ａｃｔｉｖｉｔｙに関するタンパクを標的とし、スコア値２７以上の信頼性を示す化合物は、１９８個予測されたことになる。
【０１４７】
図１２では、図１１と同様に、ＰｕｂＣｈｅｍ化合物の生物活性予測の結果であるが、標的タンパクの機能分類を異なる基準で行ったものである。図１２の見方は図１１と同様である。
【図面の簡単な説明】
【０１４８】
【図１】本発明の概念図である。
【図２】本発明の別の概念図である。
【図３】本発明の別の概念図である。
【図４】本発明の別の概念図である。
【図５】ケミカル空間とバイオ空間とを融合したモデルの構築方法についての概念図である。
【図６】従来法（ＰＣＡ）と本手法との性能評価のため、ＤｒｕｇＢａｎｋデータベースから取得したデータを基に、インシリコスクリーニング（Ｉｎ　ｓｉｌｉｃｏ　ｓｃｒｅｅｎｉｎｇ）を行った結果のＲＯＣ曲線である。
【図７】、従来法（化合物構造類似性による相互作用予測法）と本手法との性能評価のため、収集した相互作用する化合物－ＧＰＣＲのデータを基に、相互作用予測を行った結果のＲＯＣ曲線である。
【図８】β２ＡＲリガンド予測結果の検証の結果を示す。予測スコア上位５０位（Ｔｏｐ　５０）（Ｂ－１）と下位５０（Ｂｏｔｔｏｍ　５０）（Ｂ－２）の化合物に対する調査・実験結果を示す。それぞれ、左側が、文献調査および実験検証で判明した化合物の内訳であり、右側が、［^{125}Ｉ］－シアノピンドロールに対する結合阻害曲線、縦軸が阻害された割合、横軸が各化合物の濃度を示す。
【図９】新規モデルおよび従来法によるβ２ＡＲリガンド予測結果の比較を示す。各点が化合物であり、縦軸は新規モデル、横軸は従来の手法による相互作用予測スコアを示している。

【図１０】本発明の化合物ライブラリ設計の概念図である。
【図１１】ＰｕｂＣｈｅｍ化合物の生物活性予測の結果である。
【図１２】標的タンパクの機能分類を異なる基準で行ったＰｕｂＣｈｅｍ化合物の生物活性予測の結果である。
【図１３】本発明におけるＣＣＡを適用した際の計算フローである。
【図１４】本発明におけるＳＶＭを適用した際の計算フローである。
【表１】

[参考 文献]

[1] Okuno Y, Yang J, Taneishi K, Yabuuchi H & Tsujimoto G. GLIDA: GPCR－ligand database for chemical genomic drug discovery. *Nucleic Acids Res.* 34,D673－7(2006).

[2] Fredholm B.B., Fleming W.W., Vanhoutte P.M. & Godfraind T. The role of pharmacolo gy in drug discovery. *Nat. Rev. Drug Discov.* 1,237－8(2002).

[3] Wishart D.S., Knox C, Guo A.C., Shrivastava S, Hassanali M, Stothard P, Chang Z & Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *NucleicAcid s Res.* 34,D668－72(2006).

[4] Roth B.L., Kroeze W.K., Patel S & Lopez E. The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist* 6,252－262(2000).

[5] Horn F, Bettler E, Oliveira L, Camp agne F, Cohen F.E. & Vriend G. GPCRDB information system for G protein －coupled receptors. *Nucleic Acids Res.* 31, 294－7(2003).

[6] Wheeler D.L., Barrett T, Benson D.A., Bryant S.H., Canese K, Chetvernin V, Church D.M., *et. al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 34,D173－80(2006).

[7] DRAGON software is available at http://www.talete.mi.it/main_net.htm

[8] Leslie C.S., Eskin E, Cohen A, Weston J & Noble WS. Mismatch string kernels for discr iminative protein classification. *Bioinformatics* 20,467－76(2004).

[9] Vapnik V.N. The Nature of Statistical Learning Theory Springer: New York, (1995).

[10] Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization: Microsoft Research Technical Report MSRTR －98－14(1998).

[11] Keerthi S.S., Shevade S.K., Bhattacharyya C, & Murthy K.R.K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.* 13,637－649(2001).

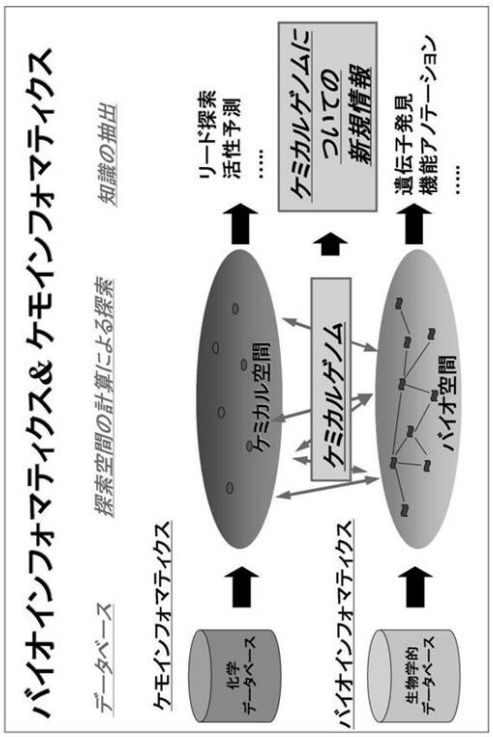
[12] Chang C.C. & Lin C.J. L IBSVM: a library for support vector machines, Software is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[13] Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Advances in Large Margin Classifiers (pp. 61－74).(1999)

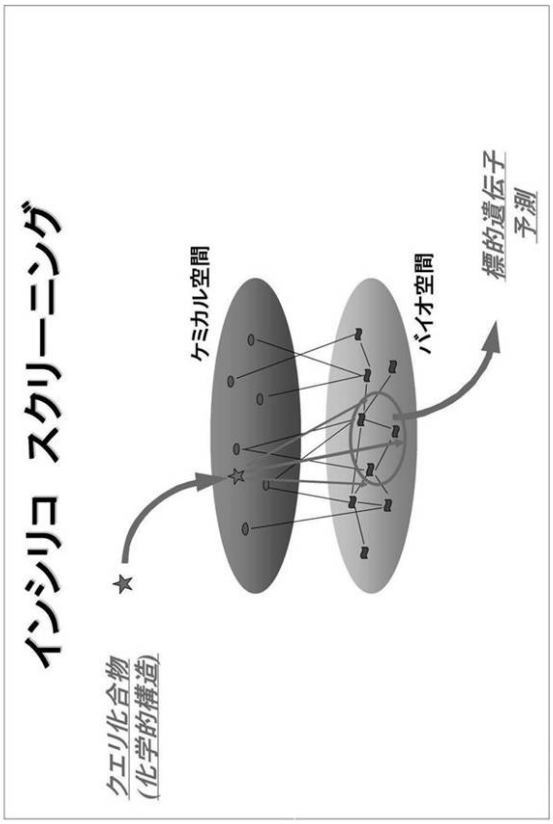
[14] OpreaT .I. Chemical space navigation in lead discovery. *Curr. Opin. Chem .Biol.* 6,384 -9 (2002)

【産業上の利用可能性】
【０１４９】
本技術によって、特に創薬の分野では新薬開発のコストを大幅に下げ、また研究開発サイクルも短縮することができる。これにより、従来よりも短い期間でより良い薬品を市場に送り出すことができる。また、製薬コストに占める研究開発費の割合を下げることで、社会的には医療費負担の低減という貢献が期待できる。

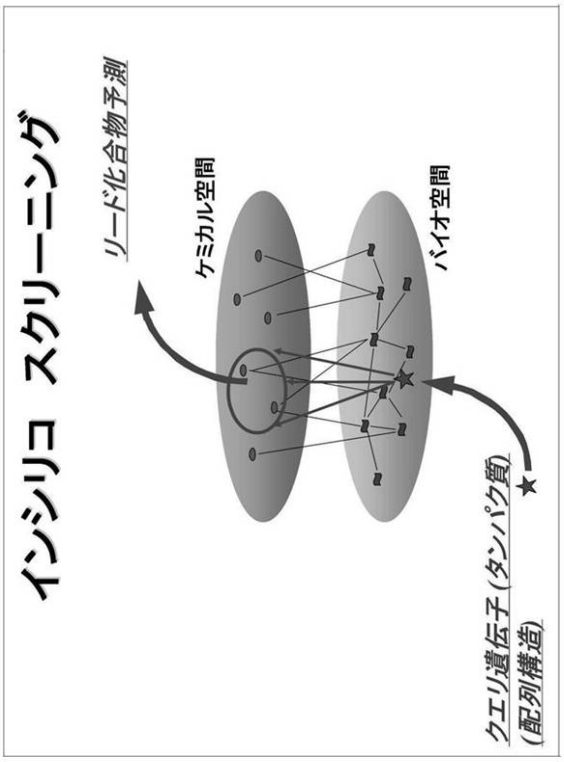
【図１】


バイオインフォマティクス& ケモインフォマティクス
データベース
探索空間の計算による探索
知識の抽出
ケモインフォマティクス
化学データベース
ケミカル空間
リード探索
活性予測
.....
ケミカルゲノム
ケミカルゲノムについての新規情報
バイオインフォマティクス
生物学的データベース
バイオ空間
遺伝子発見
機能アノテーション
.....

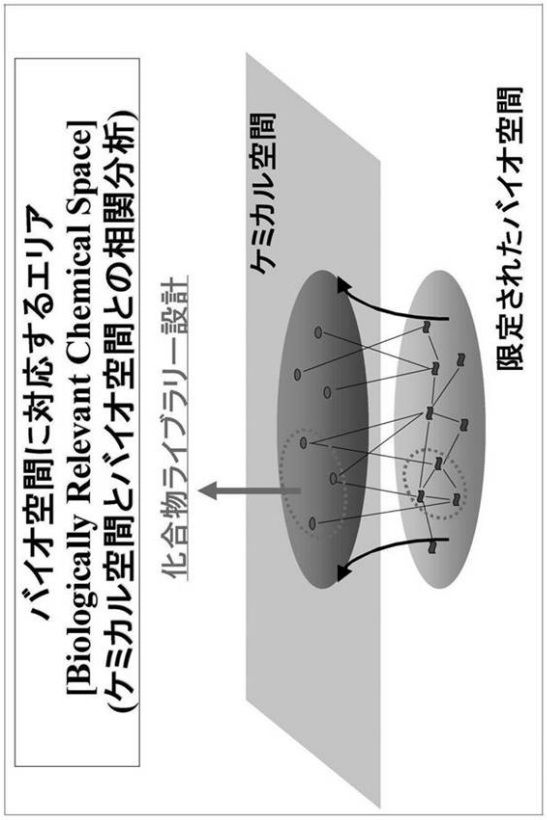
【図２】


インシリコ スクリーニング
クエリ化合物
(化学的構造)
ケミカル空間
バイオ空間
標的遺伝子
予測

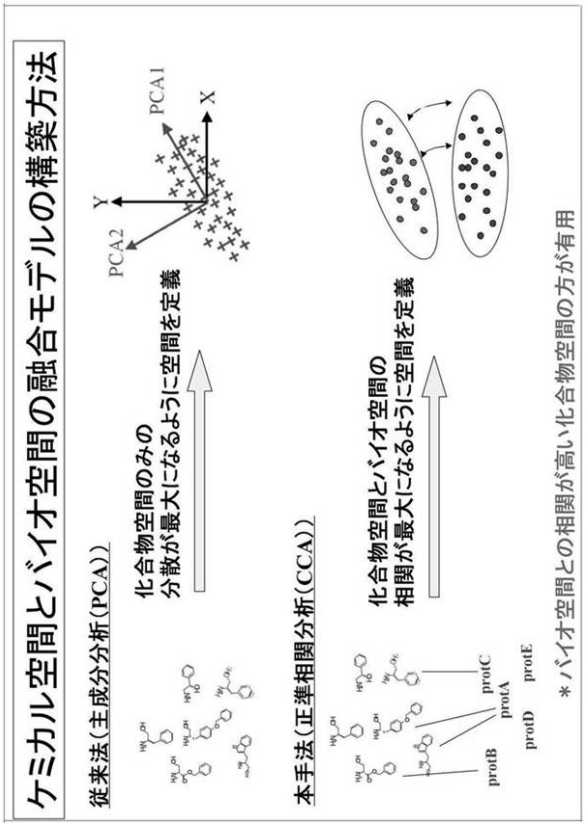
【図３】


インシリコ スクリーニング
リード化合物予測
ケミカル空間
バイオ空間
クエリ遺伝子(タンパク質)
(配列構造)

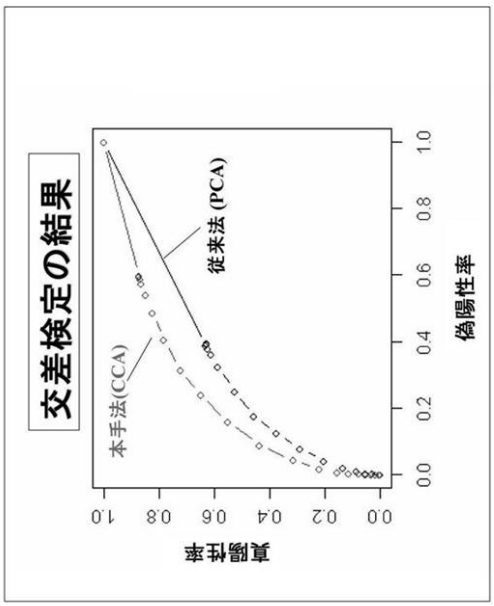
【図４】


バイオ空間に対応するエリア
[Biologically Relevant Chemical Space]
(ケミカル空間とバイオ空間との相関分析)
化合物ライブラリー設計
ケミカル空間
限定されたバイオ空間

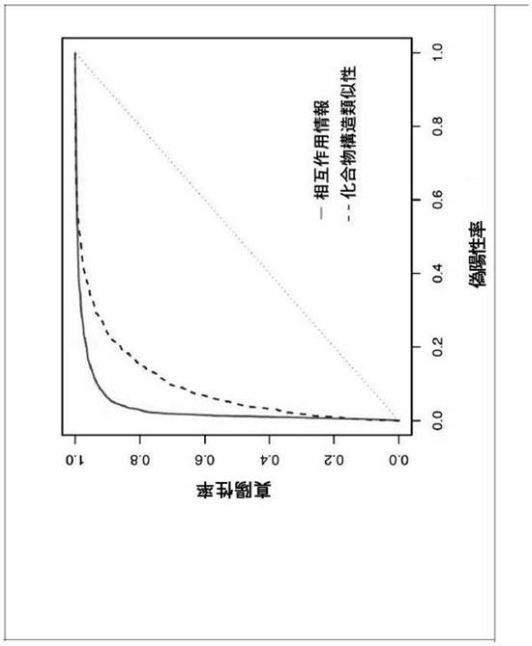
【図５】


ケミカル空間とバイオ空間の融合モデルの構築方法
従来法（主成分分析(PCA)）
化合物空間のみの
分散が最大になるように空間を定義
PCA1
PCA2
X
Y
本手法（正準相関分析(CCA)）
化合物空間とバイオ空間の
相関が最大になるように空間を定義
protB
protC
protA
protD
protE
＊バイオ空間との相関が高い化合物空間の方が有用

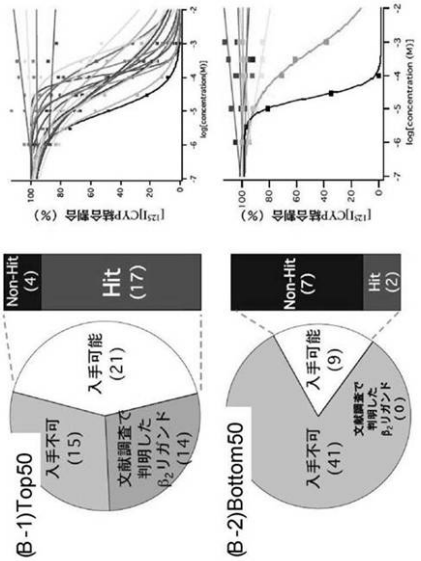
【図６】


交差検定の結果
本手法(CCA)
従来法 (PCA)
真陽性率
偽陽性率
0.0
0.2
0.4
0.6
0.8
1.0

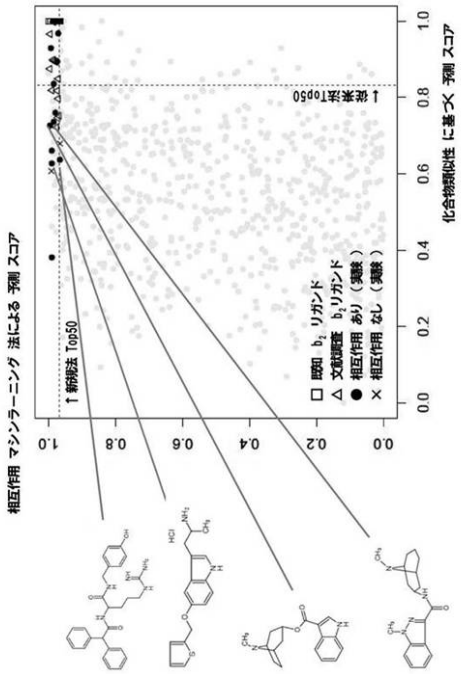
【図７】


真陽性率
偽陽性率
相互作用情報
化合物構造類似性

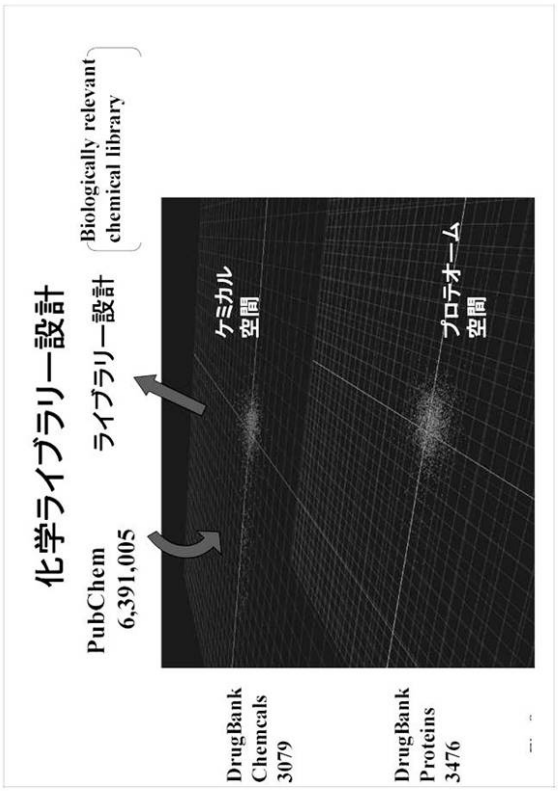
【図８】


(B-1)Top50
入手不可 (15)
文献調査で判明した β2リガンド (14)
入手可能 (21)
Non-Hit (4)
Hit (17)
(B-2)Bottom50
入手不可 (41)
入手可能 (9)
Non-Hit (7)
Hit (2)

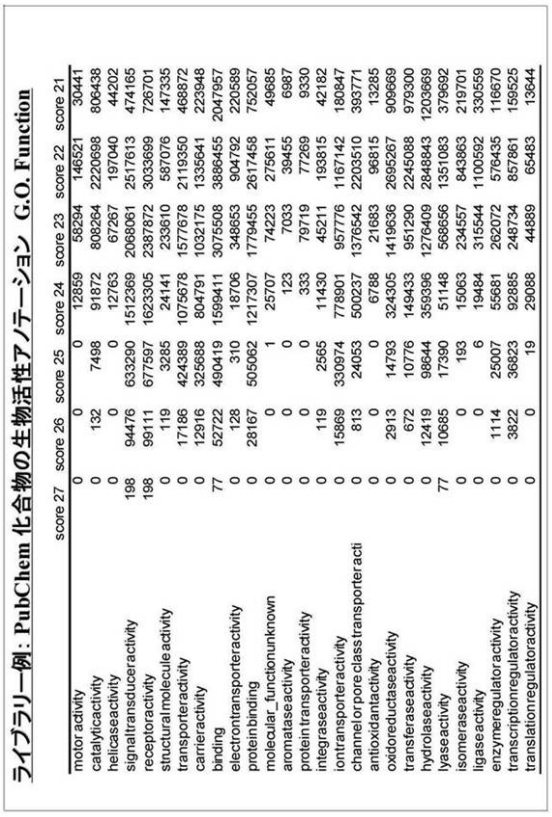
【図９】


相互作用 マシンラーニング 法による 予測 スコア
化合物類似性 に基づく 予測 スコア
↑ 新規法 Top50
↑ 従来法Top50
□ 既知 b_2 リガンド
△ 文献調査 b_1リガンド
● 相互作用 あり（実験）
× 相互作用 なし（実験）

【図１０】


化学ライブラリー設計
PubChem
6,391,005
ライブラリー設計
Biologically relevant
chemical library
ケミカル
空間
プロテオーム
空間
DrugBank
Chemcals
3079
DrugBank
Proteins
3476

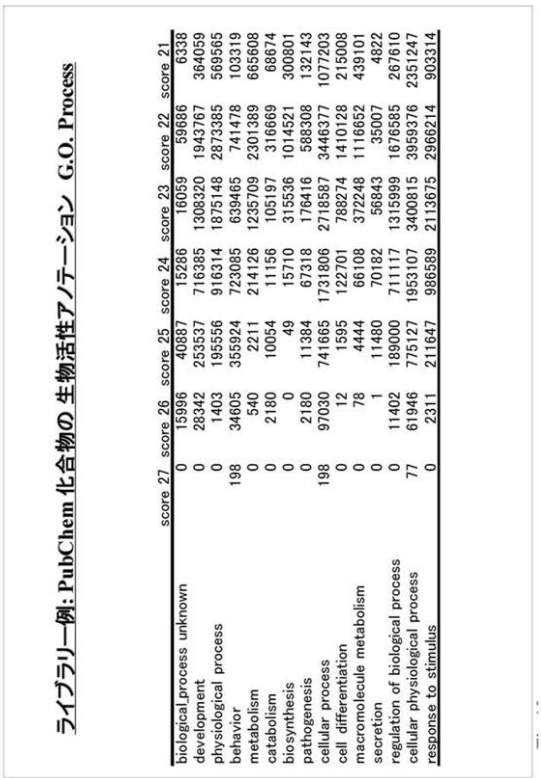
【図１１】



ライブラリー例：PubChem 化合物の生物活性アノテーション G.O. Function

	score 27	score 26	score 25	score 24	score 23	score 22	score 21
motor activity	0	0	0	12859	58294	146521	30441
catalyticactivity	0	132	7498	91872	808264	2220698	806438
helicaseactivity	0	0	0	12763	67267	197040	44202
signaltransducer activity	198	94476	633290	1512369	2068061	2517613	474165
receptoractivity	198	99111	677597	1623305	2387872	3033699	726701
structural molecule activity	0	119	3285	24141	233610	587076	147335
transporteractivity	0	17186	424389	1075678	1577678	2119350	468872
carrieractivity	0	12916	325688	804791	1032175	1335641	223948
binding	77	52722	490419	1599411	3075508	3886455	2047957
electrontransporteractivity	0	128	310	18706	348653	904792	220589
proteinbinding	0	28167	505062	1217307	1779455	2617458	752057
molecular_functionunknown	0	0	1	25707	74223	275611	49685
aromataseactivity	0	0	0	123	7033	39455	6987
protein transporteractivity	0	0	0	333	79719	77269	9330
integraseactivity	0	119	2565	11430	45211	193815	42182
iontransporteractivity	0	15869	330974	778901	957776	1167142	180847
channelor pore class transporteracti	0	813	24053	500237	1376542	2203510	393771
antioxidantactivity	0	0	0	6788	21683	96815	13285
oxidoreductaseactivity	0	2913	14793	324305	1419636	2695267	909669
transferaseactivity	0	672	10776	149433	951290	2245088	979300
hydrolaseactivity	0	12419	98644	359396	1276409	2848843	1203669
lyaseactivity	77	10685	17390	51148	568656	1351083	379692
isomeraseactivity	0	0	193	15063	234557	843863	219701
ligaseactivity	0	0	6	19484	315544	1100592	330559
enzymeregulatoractivity	0	1114	25007	55681	262072	576435	116670
transcriptionregulatoractivity	0	3822	36823	92885	248734	857861	159525
translationregulatoractivity	0	0	19	29088	44889	65483	13644

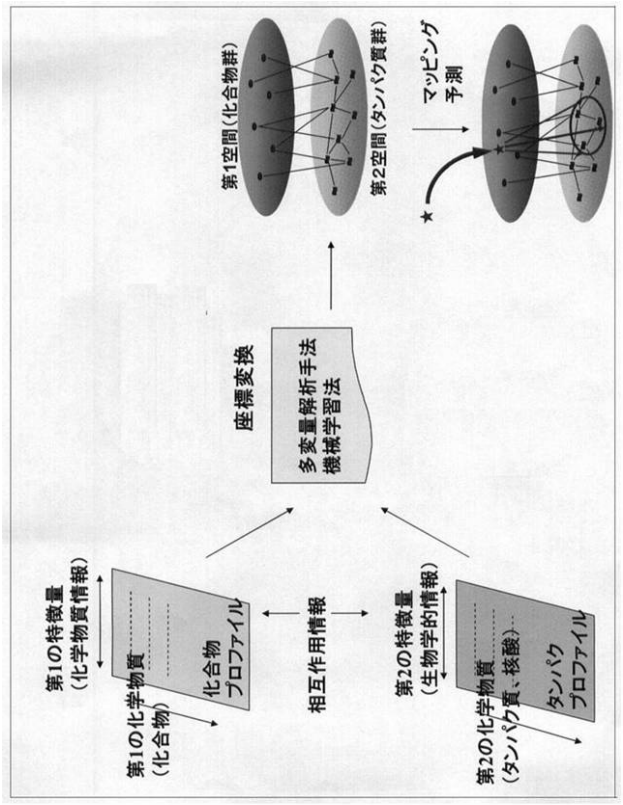
【図１２】



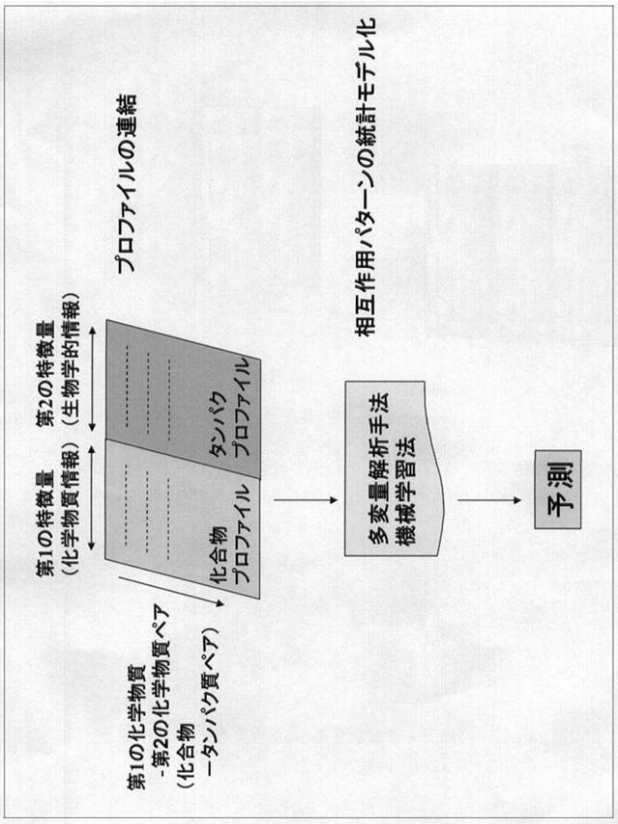
ライブラリー例: PubChem 化合物の 生物活性アノテーション G.O. Process

	score 27	score 26	score 25	score 24	score 23	score 22	score 21
biological process unknown	0	15996	40887	15286	16059	59686	6338
development	0	28342	253537	716385	1308320	1943767	364059
physiological process	0	1403	195556	916314	1875148	2873385	569565
behavior	198	34605	355924	723085	639465	741478	103319
metabolism	0	540	2211	214126	1235709	2301389	665608
catabolism	0	2180	10054	11156	105197	316669	68674
biosynthesis	0	0	49	15710	315536	1014521	300801
pathogenesis	0	2180	11384	67318	176416	588308	132143
cellular process	198	97030	741665	1731806	2718587	3446377	1077203
cell differentiation	0	12	1595	122701	788274	1410128	215008
macromolecule metabolism	0	78	4444	66108	372248	1116652	439101
secretion	0	1	11480	70182	56843	35007	4822
regulation of biological process	0	11402	189000	711117	1315999	1676585	267610
cellular physiological process	77	61946	775127	1953107	3400815	3959376	2351247
response to stimulus	0	2311	21647	986589	2113675	2966214	903314

【図１３】


第1の特徴量
（化学物質情報）
第1の化学物質
（化合物）
化合物
プロファイル
相互作用情報
第2の特徴量
（生物学的情報）
第2の化学物質
（タンパク質、核酸）
タンパク
プロファイル
座標変換
多変量解析手法
機械学習法
第1空間（化合物群）
第2空間（タンパク質群）
マッピング
予測

【図１４】


第1の特徴量
（化学物質情報）
第2の特徴量
（生物学的情報）
プロファイルの連結
第1の化学物質
-第2の化学物質ペア
（化合物
－タンパク質ペア）
化合物
プロファイル
タンパク
プロファイル
多変量解析手法
機械学習法
相互作用パターンの統計モデル化
予測

フロントページの続き

(72)発明者　辻本　豪三
京都府京都市左京区吉田下阿達町４６－２９　京都大学大学院薬学研究科内

審査官　宮久保　博幸

(56)参考文献　国際公開第２００５／０６９１８８（WO，Ａ１）
特開２００６－１２７２４８（ＪＰ，Ａ）
特開２００３－１５９０９５（ＪＰ，Ａ）
特表２００２－５３０７２７（ＪＰ，Ａ）
国際公開第０１／０６９４４０（WO，Ａ１）
山下慶子，能動学習法による創薬スクリーニング－類縁蛋白質のリガンド情報を用いたＧＰＣＲリガンド探索－，第３３回構造活性相関シンポジウム講演要旨集　，２００５年１１月２５日，ｐ．６３－６６
Okuno, Y., GLIDA: GPCR-ligand database for chemical genomic drug discovery, Nucleic Acids Research, ２００６年　1月, Vol.34, D673-D677
Yabuki, Y., GRIFFIN: a system for predicting GPCR?G-protein coupling selectivity using a support vector machine and a hidden Markov model, Nucleic Acids Research, ２００５年　7月, Vol.33, W148-W153

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１９／１０
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩＩ）
ＰｕｂＭｅｄ