平成２１年度　成果報告書

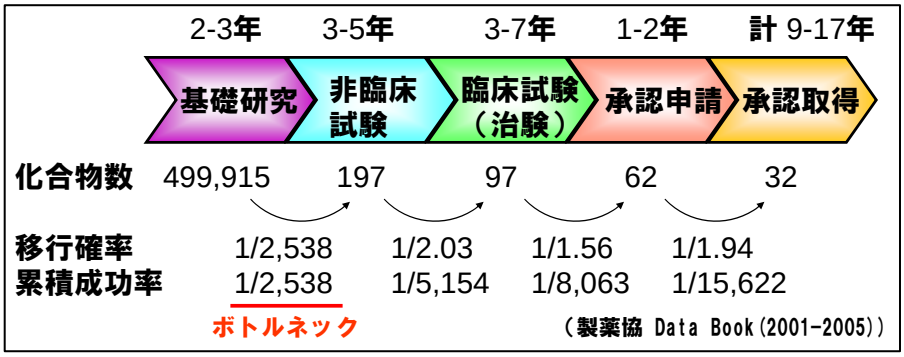
「パターン認識アルゴリズムに基づく高精度な創薬シード・リード化合物探索手法のシステム開発」

目　次

1　研究開発課題の背景

（1）医薬品開発の現状と問題点

医薬品開発には、膨大な時間と費用を要する。日本製薬工業協会のまとめ（2001～2005 年）では、実に開発期間 9～17 年、開発費用 1 新薬あたり約 500 億円、医薬品開発の成功確率 15,622 分の 1 と見積もられている。特に、化合物ライブラリーから非臨床段階までの成功確率が 0.04%と非常に低い値を示していることからも、膨大な種類の化合物ライブラリーからヒット化合物を見つけ出す工程は、医薬品開発の効率化のボトルネックになっている（図１）。

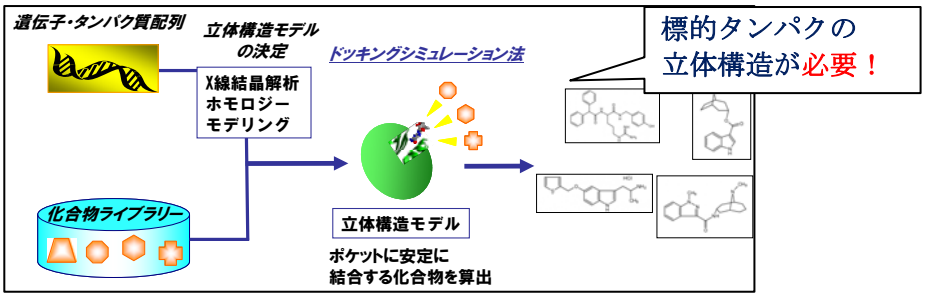


（図１．創薬プロセスの移行確率・累積性効率）

上記の工程を加速化する最も有力な方法として、計算機の中で化合物スクリーニングを行うインシリコスクリーニングが実践されている。また化合物探索の世界的プロジェクトとして、近年、「ケミカルゲノミクス・ケミカルバイオロジー」と称し各国が国策として取り組んでいる。

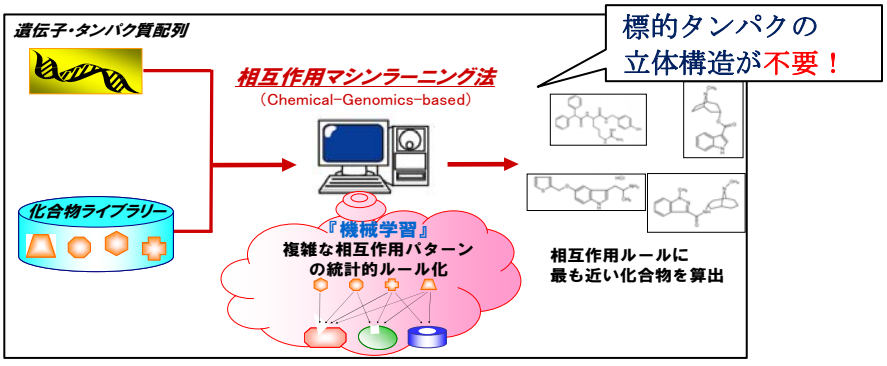
（2）従来法の問題点と新手法による改善点

一般に、広く用いられているインシリコスクリーニング手法は、薬物とタンパク質の複合体立体構造モデルを科学的にシミュレートするドッキングシミュレーション法である（図２－１）。しかしながら、市販されている医薬品のターゲットタンパクの大半は、立体構造情報を得るのが難しいとされる膜タンパクであり、タンパク質の立体構造情報に依存しない新しい予測法の開発が現実問題としては重要な課題とされていた。



（図２－１．従来法：ドッキングシミュレーション法）

このような背景を鑑み、当社は、従来法であるドッキングシミュレーションとは全く概念の異なる独自の方法論（相互作用マシンラーニング法）の開発を行い、実用化に成功している(図２－２)。



（図２－２．新規手法：相互作用マシンラーニング法）

この計算手法は、従来のインシリコスクリーニング技術であるドッキング計算とは異なり、立体構造情報を用いずに、ケミカルゲノミクス情報（タンパク質－化合物の網羅的相互作用情報）のパターン認識に基づく機械学習アルゴリズム(サポートベクターマシーン)を用いて化合物予測を行う世界に類を見ない方法論であり、すでに、共同研究先である京都大学薬学研究科奥野研究室において、GPCR ファミリーについてその予測率と新規骨格発見能力の高さが証明されたところである。

当社は、これらの新規手法をもとにして平成 20 年 3 月 31 日に起業された企業であり、受託解析事業として事業を開始し、すでに製薬企業より数件の受注を行っている。しかしながら、これらは相応のスキルを有する研究者が、手動ベースで解析を行っているものであり、民間ユーザーが手軽に利用できるソフトウエアとしての提供が待たれている。また、普及のためには GPCR ファミリー以外の主要創薬ターゲットであるイオンチャネル・キナーゼファミリーへの手法適応の研究開発及び各々の学習モデルの精度向上が必要とされる。

２　研究開発の全体計画

２－１　研究開発課題の概要

① 創薬シード・リード化合物探索システム(相互作用マシンラーニングモジュール)のパッケージ化：

京都大学の特許技術を基本にした予測プログラムは、それぞれの計算ステップが断片化されているため専門研究者のマニュアル操作によって実行している。これらステップごとに断片化したプログラムをパッケージとして１本化する。

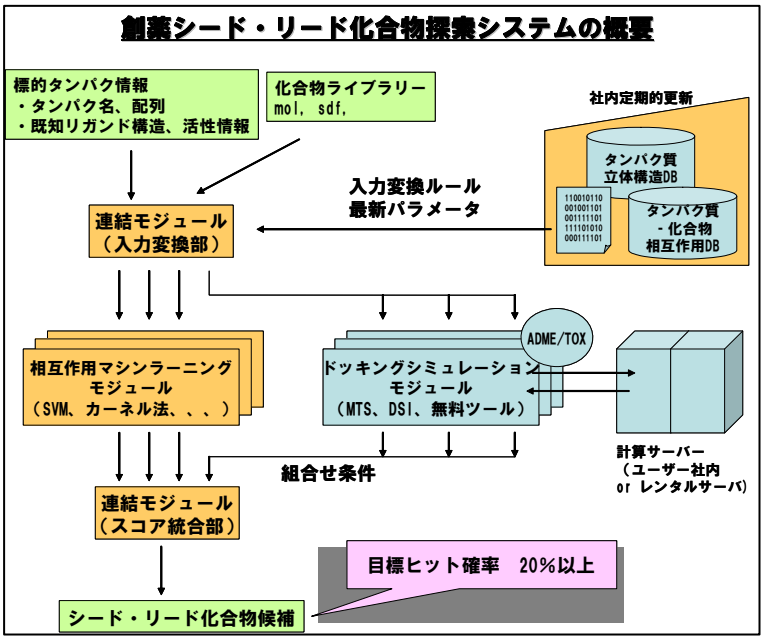


図３．創薬シード・リード化合物探索システムの概要

上図３に示すとおり、製品は、連結モジュール（入力変換部）、連結モジュール（スコア統合部）、相互作用マシンラーニング予測モジュールとドッキングシミュレーション予測モジュールから構成される。製品は、標的タンパク情報（タンパク名、配列、既知リガンド構造、活性情報）と予測対象の化合物ライブラリーの化学構造情報（mol, sdf 形式）をユーザ入力情報とし、標的タンパク質をターゲットとするシード・リード化合物候補を予測結果として出力する。

［各モジュールの作製について］

Ａ．連結モジュール・入力変換部の作製

連結モジュール・入力変換部は、ユーザ入力インターフェースとしての機能を持ち、標的タンパク質名などユーザ入力情報を各予測モジュールの入力フォーマット（例、相互作用パター

ンの教師セット、立体構造）へと変換するとともに、各予測モジュールへの並列ジョブ入力を実践する。ここで、変換処理を迅速に行うために、独自の変換ルールや、予測精度の精密化を目指した標的タンパクごとの最適パラメータの作成も適宜行う。

B．相互作用マシンラーニング予測モジュールの作製

現在、開発代表者が実行している相互作用マシンラーニング予測プログラムは、それぞれの計算ステップが断片化されているため専門研究者のマニュアル操作によって実行している。これらステップごとに断片化されたプログラムをパッケージとして１本化する。

C．連結モジュール・スコア統合部の作製

連結モジュール・スコア統合部は、予測モジュールの各ソフトで予測された予測スコアを組み合わせた総合スコアを算出し、総合スコアの高い順にシード・リード化合物候補を出力する。

② 外部ソフトとの連携のための連結モジュールの付加による統合システムの開発

当社の主要技術である相互作用マシンラーニング法は、既存手法に置き換わるというよりも、既存手法(ドッキングシミュレーション法等)と組合わせることで、より高い予測精度を期待できる。本開発システムでは、これらの高い予測精度を示す複数の方法を組み合せることにより、さらなる予測性能の向上を目指す。

パラメータ条件を変えながら計算を行うことにより、計算結果の正解傾向とパラメータとの因果関係やソフト間での予測傾向などを詳細に分析、性能分析を行った個々の予測ソフトの組合せを図ることにより、予測精度を向上させる最適な組合せ条件とパラメータを決定する。

これらによる精度向上方策を確立した上で、既存手法の予測ソフト(無償、市販製品)との組み合わせ予測を実現する連結モジュールを開発する。連結モジュールは、入力変換部とスコア統合部とに分ける。

２－２　研究開発の最終目標（平成 22 年 9 月末）

① 創薬シード・リード化合物探索システム(相互作用マシンラーニングモジュール)のパッケージ化

・ 各モジュール(連結モジュール・入力変換部の作製、相互作用マシンラーニング予測モジュールの作製)の完成。
・ 目標ヒット率はμMオーダーで 10%。

② 外部ソフトとの連携のための連結モジュールの付加による統合システムの開発

・ 代表的なドッキングシミュレーションソフトであるＭＯＥとの連結化の完成。
・ 目標ヒット率はμMオーダーで 20%。

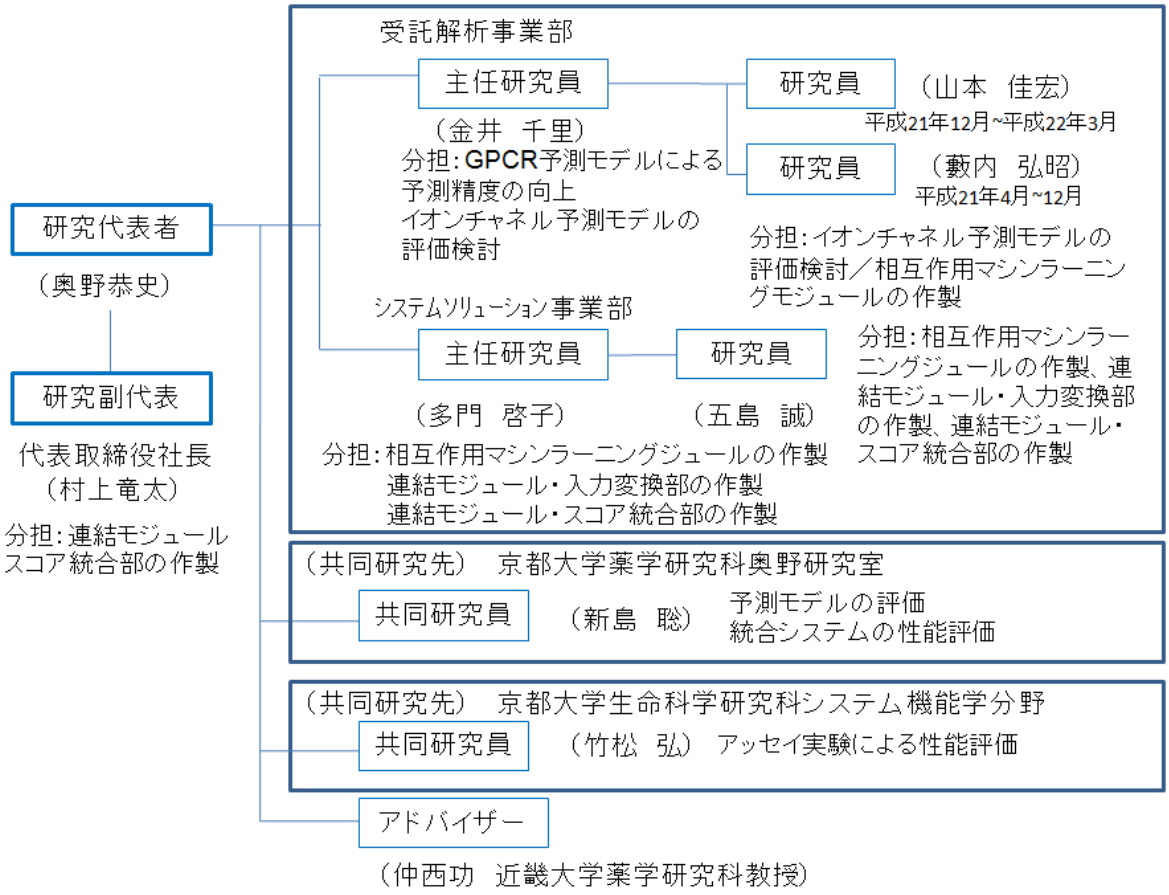
２－３　研究開発の年度別計画

金額は非公表

研究開発項目	２０年度	２１年度	２２年度	計	備考
パターン認識アルゴリズムに基づく高精度な創薬シード・リード化合物探索手法のシステム開発					
①創薬シード・リード化合物探索システム(相互作用マシンラーニングモジュール)のパッケージ化	－	－	－	－	
②外部ソフトとの連携のための連結モジュール（スコア結合部）の製作		－	－	－	
間接経費額（税込み）	－	－	－	－	
合　計	－	－	－	－	

3　研究開発体制

3－1　研究開発実施体制



４　研究開発実施状況

４－１　GPCR 予測モデルによる予測精度の向上

４－１－１　GPCR 予測モデルによる予測精度の向上

昨年度の研究成果である GPCR 予測モデルの構築スキームでもって、β２アドレナリン受容体（b2AR）について、予測モデルの構築、外部ライブラリーを用いた活性予測、アッセイ実験による精度検証を行った。

また、β２アドレナリン受容体（b2AR）について、既存の計算手法２種（LBDD、SBDD）との比較を行い、その精度検証を行った。

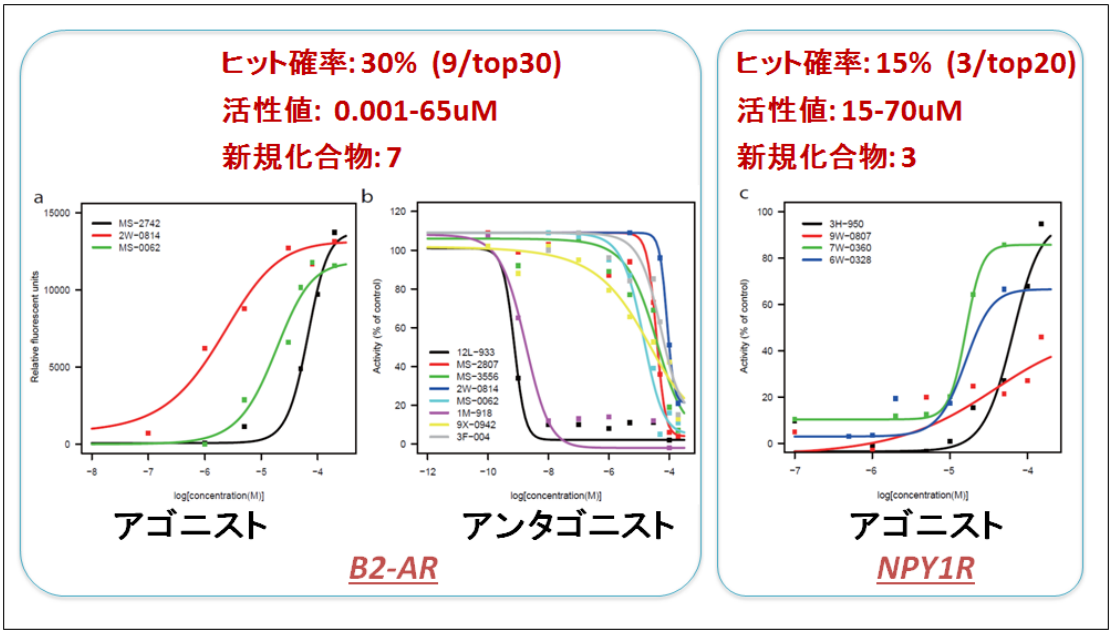
４－１－２　b2AR 及び NPY1R に対する活性予測と実証実験

相互作用マシンラーニング法を GPCR のβ２アドレナリン受容体（b2AR）を標的としたインシリコスクリーニングに適用し、活性予測と実証実験を行った。

まず､317 種類の GPCR のタンパク質配列とそれらに作用する 866 種の化合物の化学構造からなる 5207 通りの既知の相互作用ペアを計算機に機械学習させ､GPCR と化合物との既知相互作用に内在する統計的保存パターンを抽出し、それらに基づく予測モデルを構築した。そして、この予測モデルを用いて、Bionet 社化合物ライブラリーの 11,739 化合物から、b2AR と相互作用する可能性の高い化合物の絞込みを行った。11,739 化合物のうち、活性スコアに基づいて上位 30 個の化合物について、アゴニストおよびアンタゴニスト活性測定を行った。

その結果、図 4-1-1 左に示すように、上位 30 化合物中、9 個の化合物が活性を示し（ナノモルオーダーの非常に高活性な化合物を含む）、実に 30%という驚異的なヒット確率を示した。また、7 個の化合物は新規構造であり、新薬開発の特許戦略上、最も重要である新規化合物の探索にも本手法は成功している。

図 4-1-1. 予測化合物の検証実験結果：GPCR 作用化合物



また、上記の b2AR 以外に、ペプチド系 GPCR の一つであるニューロペプチドＹ１受容体（NPY1R）を標的としたインシリコスクリーニング計算に適用し、Bionet 社化合物ライブラリーの 11,739 化合物の活性予測を行い、活性スコアに基づいて上位 20 個の化合物について In vitro 活性実験を行った。その結果、上位 20 化合物中、3 個の化合物が活性を示し（15%のヒット確率）、いずれも非ペプチド系の新規骨格を有する化合物であった（図 4-1-1 右）。

４－１－３　CGBDD と既存手法との比較・精度検証

１、ケミカルゲノミクス情報の収集・整備

CGBDD および LBDD による ADRB2 リガンド予測を行うため、GLIDA データベースから、G タンパク質共役型受容体（GPCR）とそのリガンドに関する相互作用の情報 5206 件を収集した。この情報には、ヒト ADRB2 の相互作用 40 件が含まれている（図 4-1-2）。

図 4-1-2. ヒト ADRB２　タンパク配列情報

```
DYKDDDAMGQPGNGSAFLLAPNRSHAPDHDVTQQRDEVWVVGMGIVMSLIVLAIVFGNVLVITAIAKFERLQTVTNYFITSLACADLVMGLAVVPFGAAHILMKM
WTFGNFWCEFWTSIDVLCVTASIETLCVIAVDRYFAITSPFKYQSLLTKNKARVIILMVWIVSGLTSFLPIQMHWYRATHQEAINCYAEETCCDFFTNQAYAIAS
SIVSFYVPLVIMVFVYSRVFQEAKRQLNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDAAVRGIL
RNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYKFCLKEHKALKTLGIIMGTFTLC
WLPFFIVNIVHVIQDNLIRKEVYILLNWIGYVNSGFNPLIYCRSPDFRIAFQELLCLRRSSLKAYGNGYSSNGNTGEQSG
```

２、化合物、タンパク質の記述子計算

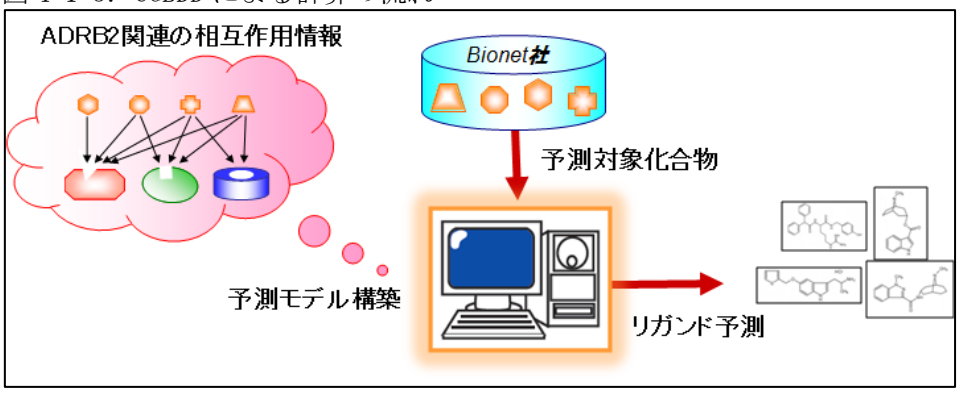
収集した相互作用をベクトルとして表現するために、各化合物の化学構造、各タンパク質のアミノ酸配列について、その属性（記述子と呼ぶ）を計算した。なお、化合物記述子には、物性や構造的特徴を数値化する商用ソフト DragonX を用い、タンパク質記述子にはアミノ酸組成頻度を数え上げるプログラムを用いた。

また、Bionet 社の化合物ライブラリについて、同様にして記述子計算を行った。このとき、ライブラリ化合物間で構造類似性検索を行い、著しく化学構造の類似したものを取り除いた。その結果、11500 手類の化合物を予測対象として用いることにした。

３、CGBDD による ADRB2 リガンド予測

代表的な学習アルゴリズムであるサポートベクターマシン（SVM）を用いて、化合物-タンパク質相互作用の有／無を判別した。具体的には、正例（相互作用する化合物 - タンパク質ペア）および負例（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した。続いて、この学習モデルを用い、Bionet 化合物-ADRB2 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予測を行った（図 4-1-3）。

図 4-1-3. CGBDD による計算の流れ

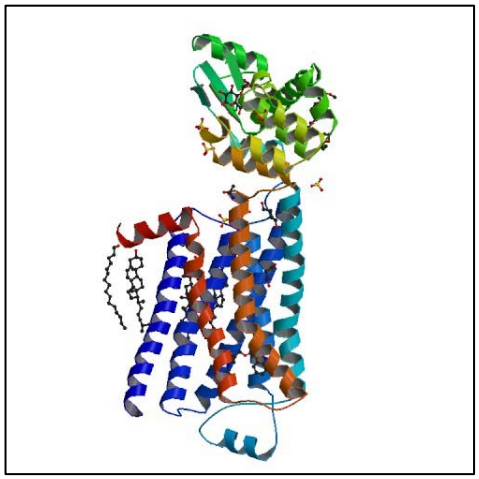


４、LBDD による ADRB2 リガンド予測

化学記述子によって定義される空間にて近傍に位置する化合物は、類似する生物活性を示すことが知られている。我々は、化学記述子から計算される主成分空間上での既知リガンドからの距離を尺度として、予測する Bionet 化合物-ADRB2 間のスコアを算出することにした。具体的には、各予測化合物について、すべての ADRB2 既知リガンド 40 種類に対して、主成分ベクトルのピアソン相関係数を計算し、その最大値を LBDD スコアとした。なお、主成分軸には、累積寄与度が 80%に達するまでの 30 軸を用いた。

５、SBDD による ADRB2 リガンド予測

図 4-1-4. ADRB2 の結晶構造



近年同定された高解像度の ADRB2 の結晶構造（PDB 番号：2RH1：図 4-1-4）を利用して、ADRB2-各化合物間の結合時の自由エネルギーを計算した。まず、タンパク質構造の前処理として、ソフトウェア MOE を用い、重原子を固定した状態で水素原子の座標を熱力学的に安定な位置に配置した。そして、ドッキングシミュレーションソフト GOLD を用いて、結晶構造上でリガンド（carazolol）結合部位にて各化合物の座標を繰り返し配置させ、フィット関数の小さいものから 10 個選んだ。なお、ADRB2 の残基 Asp113 および Asn312 については、先行研究によってリガンド-ADRB2 間に水素結合を形成すると判明しているため、ドッキングの拘束条件に追加して計算を行った。そして、これら 10 種類の配置についてスコア関数 Chemscore を計算し、その最小値を自由エネルギーとしてスコア化した。

６、ADRB2 リガンド予測に対するカルシウムアッセイ（Calcium mobilization assay）実験

上記 CGBDD、LBDD、SBDD の予測スコア上位 20 化合物をそれぞれ購入し、in vitro で活性確認の実験を行った。リガンドで刺激された ADRB2 は、Gq タンパク質を介して、細胞内のカルシウム濃度を上昇させることが知られている。そこで、試験化合物の ADRB2 活性を確認するため、まず、CHO-K1 細胞から、ADRB2 を強制的に発現した安定細胞を作製した。そして、FLIPR Calcium Assay 4 Kit (Molecular Devices, Sunnyvale, CA)を用いて、各化合物投与による蛍光(Ex 485 nm, Ex 525 nm)の変化を測定し、ADRB2 活性を確認した（使用機器：図 4-1-5)。ADRB2 アンタゴニスト効果の測定時には、あらかじめ、既知 ADRB2 アゴニストであるイソプロテレノール（50nM）で刺激し、10 分後に試験化合物を投与してその阻害効果を確認した。これらの実験を同一条件下で 4 回測定し、統計的に有意（$P < 0.05$、対応のない T 検定、$n = 4$）なものについては、活性の強弱を確認するため、複数の濃度点（1.0-200uM）で測定し、50%効果（阻害）濃度を求めた。

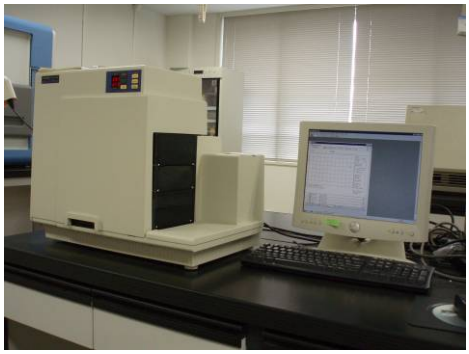


図 4-1-5. 測定装置「FLEX station」

使途：細胞内のカルシウム濃度を蛍光強度として測定する

（３）実証実験の結果

スコア上位化合物について、カルシウムアッセイにて各々の手法で ADRB2 活性が確認できたものを表 4-1-1 に示す。

表 4-1-１　スコア順位と ADRB2 活性

ID	CGBDD 順位	LBDD 順位	SBDD 順位	EC_{50} (uM)	IC_{50} (uM)
NS-00895987	1	309	707	-	0.0007
NS-00898737	14	1616	1096	-	0.0017
NS-00902139	4354	8163	26	< 1uM	30
NS-00902920	10	322	-	1.9	83
NS-00901956	1130	4468	18	-	3
NS-03595378	1706	20	-	-	3
NS-00916999	4565	9726	15	-	3
NS-00926838	7510	7613	12	-	3

NS-00925291	**13**	3945	5620	17	13
NS-00923997	26	2271	6840	-	13
NS-00909249	9241	1790	**10**	-	25
NS-00927450	**6**	**3**	5616	-	33
NS-00198632	**4**	238	2248	-	37
NS-00903394	27	124	-	-	48
NS-00106740	53	**14**	6044	-	60
NS-00926872	**8**	6825	1836	65	-
NS-03596918	1471	3886	**19**	100	-

CGBDD スコア上位 20 化合物のうち、ADRB2 活性が確認できたものは 7 個（うち 3 つはアゴニスト活性）であった。これらのうち、NS-0895987 および NS-00898737 は非常に強力なアンタゴニスト活性（IC50=0.7nM, 1.7nM）を示した。また、他の化合物については、活性自体は弱いものの、既知の ADRB2 リガンド骨格を持たないため、新規骨格化合物といえる。

一方、LBDD スコア上位 20 化合物のうち、ADRB2 活性が確認できたものは 3 個（アゴニスト活性なし）であった。

また、SBDD スコア上位 20 化合物のうち、ADRB2 活性が確認できたものは 6 個（うち 2 つはアゴニスト活性）であった。このうち一つは、強いアゴニスト活性（EC50<1uM）を示した。

（4）結果の考察及び実証実験の達成度

以上の結果より CGBDD は、化合物ライブラリーを対象としたリガンド予測計算においても、従来法を超えるリガンド発見能力を有することが示された。特に、低計算コストで新規骨格化合物を発見に繋がったことから、CGBDD は新薬開発におけるニーズにぴったり適合した手法といえる。特に注目すべきは、それぞれのスコア順位を比較したときに、その相関性が弱く、いずれの手法も互いに相補的な結果となっている点である。これは、CGBDD が、従来法では発見できなかったリガンドを射程圏内に捉えていることを意味している。すなわち、これら 3 種のスクリーニング手法の使い分けにより、より合理的な新薬開発が可能となると考えられる。

４－１－４　まとめ

以上をもって、GPCR の予測モデルのモデル構築方法の検討、精度評価については、実施計画上の目標を到達した（μMオーダーで 10%以上）。よって、GPCR に関する予測精度の向上に関する開発目標については今年度をもって完了とする。

４－２　イオンチャネル・キナーゼ予測モデルの評価検討

４－２－１イオンチャネル・キナーゼ予測モデルの評価検討

イオンチャネルについては、昨年度の研究成果により、予測モデルの構築方法について確認できたことから、21 年度は大規模な相互作用情報を用いた予測モデルの構築、活性予測、実証実験を行った。

キナーゼについては、昨年度の実証実験の結果を精査し、さらに各キナーゼターゲットに対する活性予測を行い、選択性の評価を行った。

４－２－２　（イオンチャネル）・データベースの評価

イオンチャネルは、生体膜間にイオンを透過させる機能をもつタンパクの総称であり、GPCR と同様な膜貫通タンパク質の一種である。細胞の内外における各種イオンの濃度あるいは膜電位の維持、神経細胞など電気的興奮性細胞での活動電位の発生、シグナル伝達などに関与する。イオンチャネルの開閉の制御様式には幾つかあるが、主に電位依存性とリガンド依存性の 2 種類がある。電位依存性とは、膜電位の変化に応じてチャネルが開閉し、イオンの濃度勾配からなる駆動力により特定のイオンを選択的に透過させるものである。一方、リガンド依存性とは、受容体とイオンチャネルが共役している構造で、リガンドが受容体に結合することでチャネルが開いてイオンを透過させるものである

購入したデータベースと当社データベースを統合・整理し、データベースの評価を行った。以下の表にイオンチャネルのそれぞれのグループにおける相互作用数（リガンド数）を示す。昨年度のデータベースの状況と比べて大幅に情報量が上がっており、総リガンド数は 15792 個（昨年度）から 48429 個（今年度）へと改善されている。

４－２－３　（イオンチャネル）・予測モデルの構築とその検証

イオンチャネルリガンドデータベースからデータサンプリングを行い、相互作用マシンラーニング法により学習モデルを作成した。サンプリングは 1 タンパクあたり最大 1800 個までのリガンドを採用する方法を用いた。サンプリングした相互作用数は約 6 万であった。

サンプリングしたリガンド―タンパクの相互作用情報から相互作用マシンラーニング（SVM）モデルを構築する際に用いるリガンドとタンパクのベクトル化を行う必要がある。それらの手法として昨年度はリガンドとタンパクのそれぞれのベクトル化手法を組み合わせて検証したが、その結果から、リガンド側は MOE2D、CATS、MACCS の 3 手法を、タンパク質側は mism30 を用いると予測性能が良好である事を見出しているので、今回の大規模データに対してこれらのベクトル化手法を適用して検証した。しかしながら、化合物側の MACCS の手法については機械学習計算が不安定になりモデルを構築することが出来なかった。

機械学習に用いていないリガンド情報を各イオンチャネルのターゲットについて予測計算を行うことによって、モデルの予測性能を ROC 曲線図にて確認した。ナトリウムチャネルの Nav1.8 以外については良好な予測性能を示している。特に TRPV1、NMDE、Cav2.2(N 型 Ca チャネル)に 3 つについては非常に高い性能を示している。逆にナトリウムチャネルについては学習データに偏りがあった為、低い精度だったと思われる。この点に関しては、さらなる相互作用データの充実を図ることにより解消できる。

４－２－４　（イオンチャネル）・予測モデルの構築と実証実験

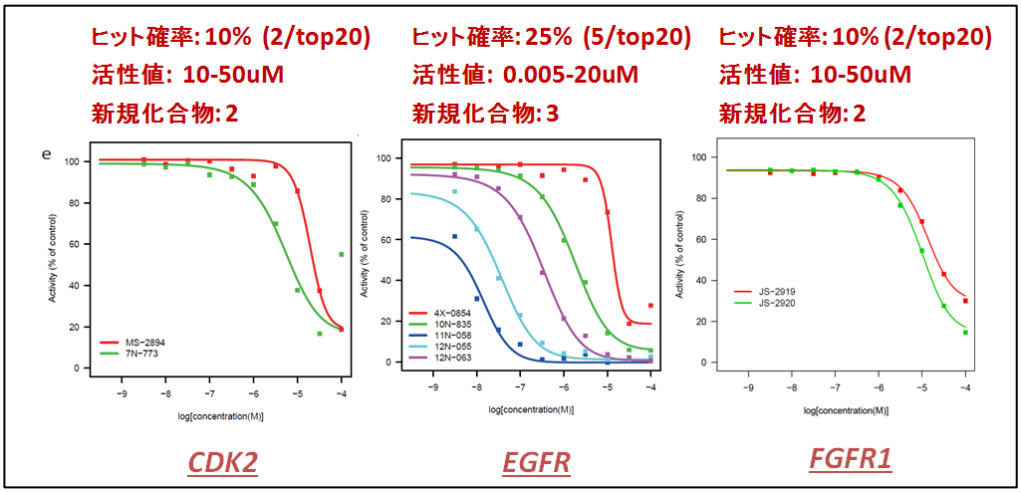
構築した予測モデルのうち、化合物記述子に MOE2D を用いたものを用いて、計算による活性予測を行い、アッセイ実験による精度評価を行った。上位 50 化合物について活性評価を行ったところ、13 化合物のヒットが確認された（＜5μM 以下）。ヒット率としてはおよそ 26%（13/50）を達成できている。

４－２－５　（キナーゼ）・実証実験結果の精査

GPCR、イオンチャネルと並ぶ重要な創薬標的タンパク質であるキナーゼファミリーについて、20 年度の実証実験の結果を精査し、ヒット率及び新規化合物の有無について検証を行った。

この予測では、5616 種の既知の相互作用ペアを計算機に機械学習させて予測モデルを構築し、Bionet 社化合物ライブラリーの 11,739 化合物の活性予測を行っている。CDK2、EGFR、FGFR1 の三種類のキナーゼを予測対象とし、各キナーゼに対して活性スコアの上位 20 化合物の酵素活性阻害実験を行ったところ、図 4-2-2 に示すとおり、各々のキナーゼに対して、2～5 個の化合物に強い阻害活性が見られ、いずれにおいても、10%以上のヒット率と新規骨格の発見能力を示すことが確認された。また、ヒット化合物のうち、新規化合物についてもそれぞれ 2 ないしは 3 個が見出された。

図 4-2-2. 予測化合物の検証実験結果：キナーゼ



４－２－６　（キナーゼ）・選択性についての実証実験

合理的投薬・副作用軽減を目指したときに、目的の標的タンパク質に作用する化合物を設計すること、すなわち「選択性」を制御することが、創薬プロセス上で重要な鍵となる。特に、キナーゼに関しては、タンパク質同士での結合部位の区別化が難しく、特定のキナーゼにのみ作用する阻害剤を設計することが、はなはだ難しいとされている。

そこで、相互作用マシンラーニング法がこの「選択性」予測にも適用可能か否かを明らかにするため、in vitro での検証評価を行った。まず、4-2-5 章で発見した CDK2 ヒット化合物「7N-773」について、同じ学習モデルを用いて、アッセイ可能な 46 種類のキナーゼに対する相互作用予測スコアを算出した。そして、実際にこれらのキナーゼに対して、7N-773 による酵素活性阻害実験を行い、7N-773 10uM における阻害率を測定し、予測結果との照合を行った。

得られた結果として、最も高い予測スコア（0.99 超）を持つキナーゼ 2 種は、いずれもその阻害率が 95%を超えており、また逆に、他のすべてのキナーゼはこの阻害率を大きく下回った。したがって、当モデルは高活性の標的キナーゼを精度良く予測できるといえる。一方、予測スコアの低いキナーゼは、おしなべて阻害活性が低いため、阻害効果の無いキナーゼに対しても、あらかた正しく予測できたといえる。たとえば、スコア 0.8／阻害率 40％を閾値にしてクロス集計表を作成すると、予測スコアの高低と活性の有無の間に有意な関連性（Fisher 正確確率検定、P=0.0089）が確認された。これらにより、我々の相互作用予測モデルは、一般的に困難とされるキナーゼ選択性の予測にも有効であることが示唆された。

４－２－７　まとめ

イオンチャンネルについて、大規模データを用いた予測モデルの構築方法、具体的事例に対する活性予測と実証実験の結果、十分な成果を得ることが出来た。GPCR と同様に精度向上の検証については、21 年度をもって完了とする。

一方、キナーゼについては、高性能なリガンド予測性能を得ることを示すことが出来てきる。さらに選択性に関して、予測スコアの高低と阻害活性の有無との間の有意な関連性を示すことが出来ている。しかしながら、キナーゼについてはその選択性が重要視されることから予測スコアと阻害活性との相関をさらに検証する必要がある。そのため、来年度は、キナーゼに対する更なる精度向上の検証について重点を置いて実施する。

４－３　(A) 連結モジュール・入力変換部の作製

４－３－１　連結モジュール・入力変換部のプロトタイプシステム作製

連結モジュール・入力変換部は、以下の機能を持つ。

１．ユーザーインターフェース

ユーザーの使用しやすいグラフィカルなインターフェースを設け、学習モデルを作成するために必要なデータ・パラメータを受け取る。

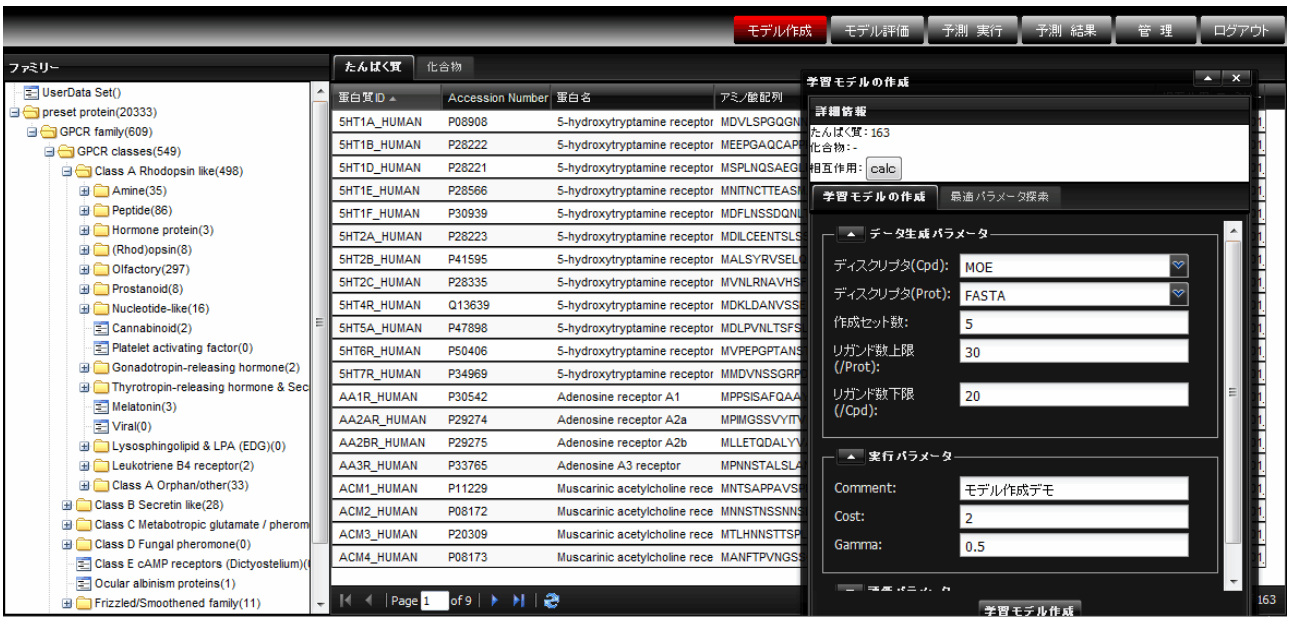


図４－３－１．ユーザーインターフェース例（学習モデル作成）

図４－３－２．パラメータ入力インターフェース（学習モデル作成）

２．学習モデル作成

予測に用いる学習モデルを作成する。

その際、１で選択されたデータ・パラメータを適切な形に変換する。

また、この処理は、一週間程度の長い処理時間を要する場合がある。よって進捗状況を確認できるインターフェースを設ける。

２－１．学習モデルの登録

① システムのメインデータベースはタンパク質、相互作用情報、化合物の各情報を持ち、タンパク質検索することで相互作用情報を通じて化合物の情報を自動的にひき、学習用のデータセットを作成する。

② 検索インターフェースは通常の利用想定としてタンパク質を対象にした検索を行えるものとする。

③ 選択条件に応じて対応する化合物数、相互作用数等を確認できるようにする。

④ Cost や Gamma、ディスクリプタ等のパラメータを入力し、選択されたデータと共に SVM の学習データを作成し、ジョブとして登録する。

パラメーター一覧

項目名	初期値	内容
データ生成パラメータ		
ディスクリプタ（Cpd）	MOE	化合物の SVM 用ベクトル作成方法を選択。 MOE、Dragon、MPE で選択一つ選択。
ディスクリプタ（Prot）	FASTA	タンパク質の SVM 用ベクトル作成方法を選択。 現状、選択肢は FASTA の一つ。
作成セット数	5	負例を自動作成するセット数。
リガンド数上限（/Prot）	30	１タンパク質あたりの相互作用数の上限。
タンパク質数（/Cpd）	20	１化合物あたりの相互作用数の下限。
実行パラメータ		
Comment	モデル作成	学習モデルを認識する名前をつける。
Cost	2	SVM のパラメータ Cost 値。
Gamma	0.5	SVM のパラメータ Gamma 値。
評価パラメータ		
n-foldCV n	5	CrossValidation（モデル性能検証）の検証方法のパラメータ。
n-foldCV Iter	1	CrossValidation の繰り返し回数。

２－２．ジョブ管理

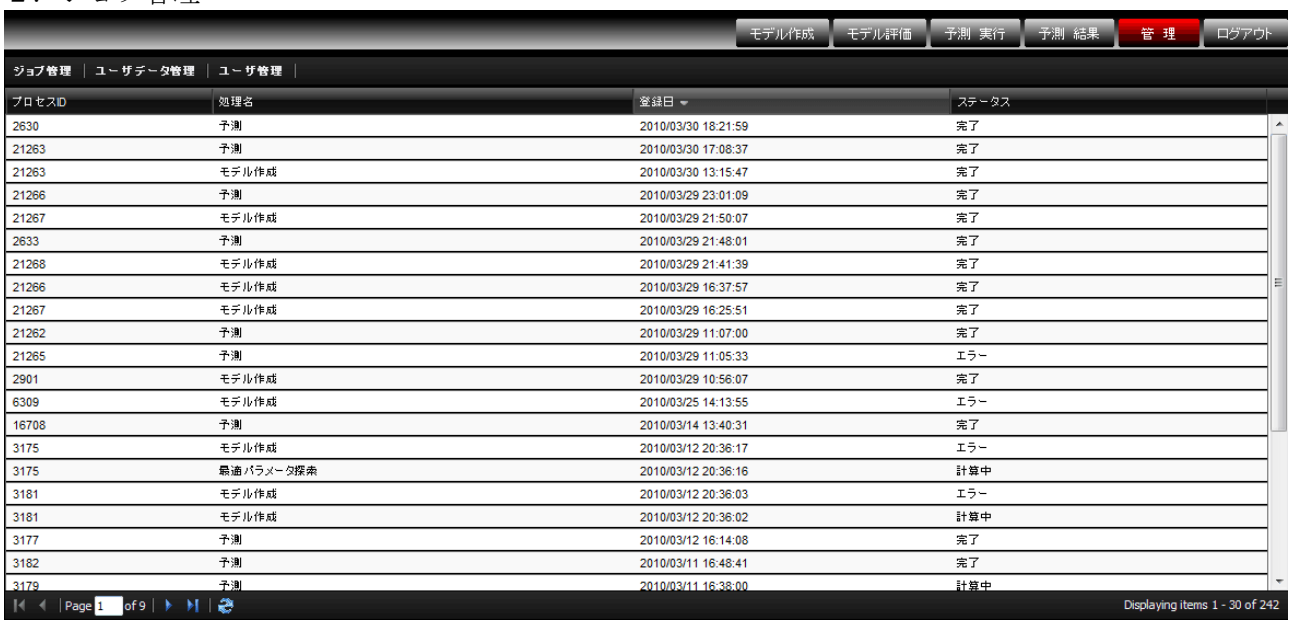


図４－３－３．ジョブ管理インターフェース

① ジョブ管理
処理名、更新日、ステータス（完了・エラー・計算中）が表示される。
② イベントログの表示
ジョブを選択すると、処理のログが表示される。

４－３－２ まとめ

達成状況として、21 年度の目標通りのプロトタイプ版の作製が完了している。22 年度はこれらをもとに、大規模データでの使用に耐えうる最終仕様を決定し、最終製品版の開発を行う。

４－４　（B）相互作用マシンラーニング予測モジュールの作製

４－４－１　相互作用マシンラーニング予測モジュールのプロトタイプシステム作製

（A）で作成したモデルに対して、予測を実行するためのモジュールを作製する。

相互作用マシンラーニング予測モジュールの作成は、以下の機能を持つ。

１．ユーザーインターフェース

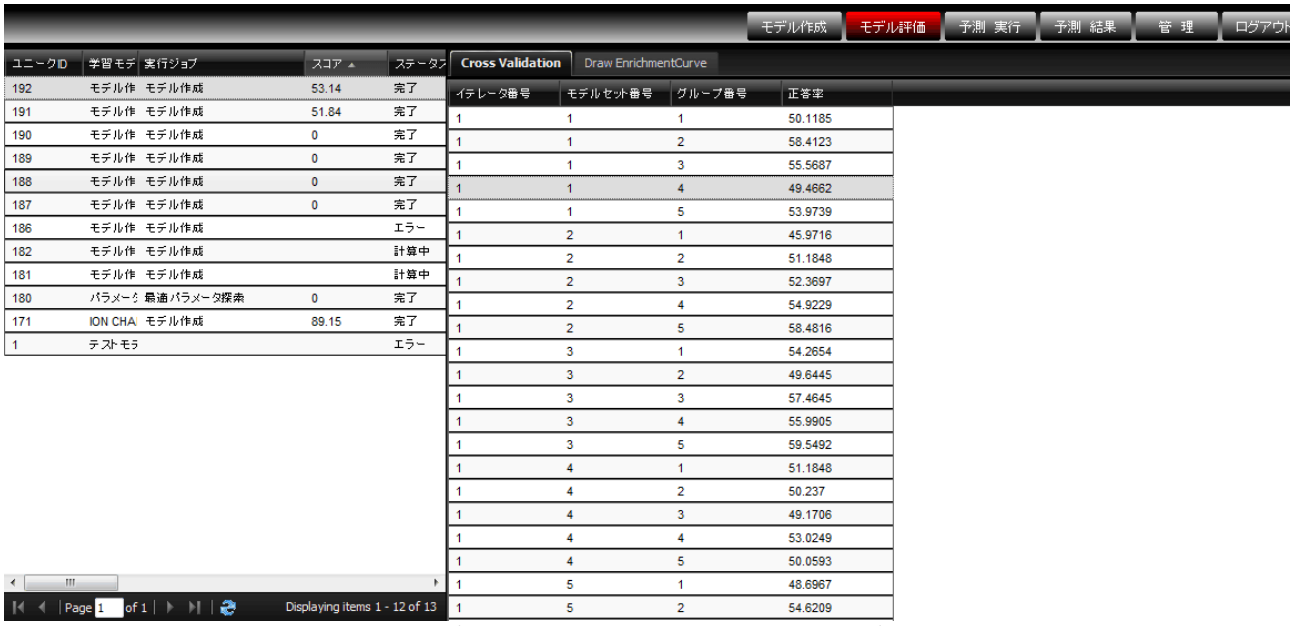


図４－４－１．ユーザーインターフェース例（学習モデル評価）

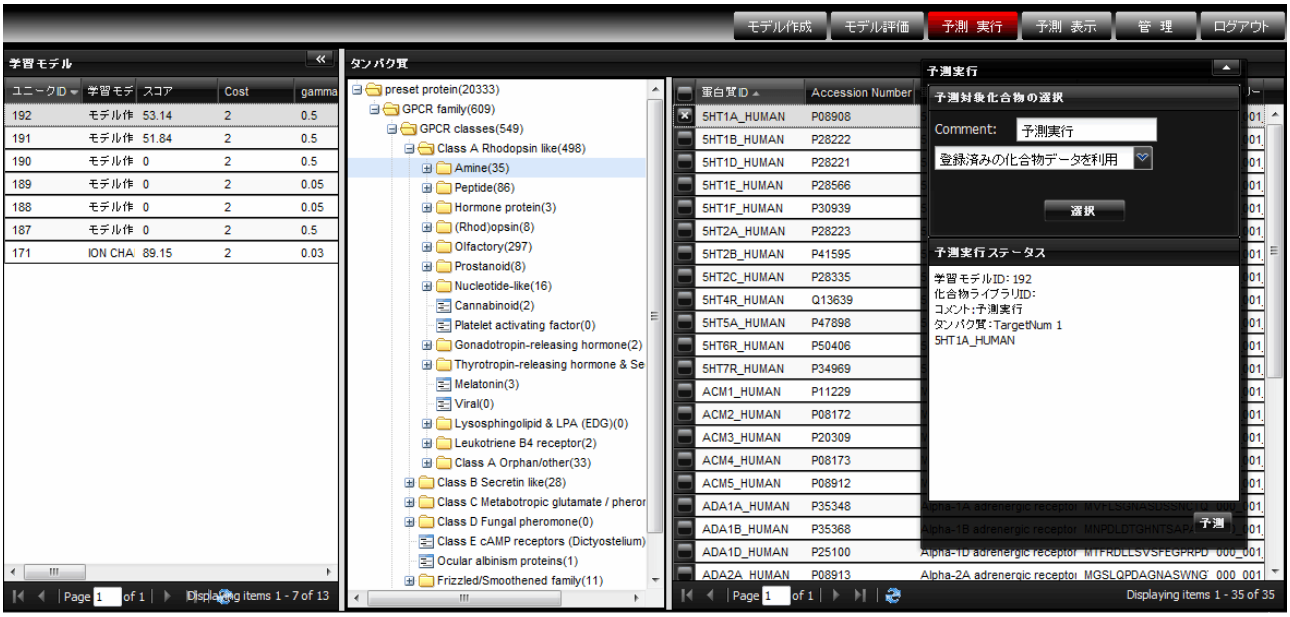


図４－４－２．ユーザーインターフェース例（予測実行）

２．学習モデル評価

学習モデル作成で作成したモデルの性能を評価・確認する。

① 「Cross Validation」タブ

モデル作成計算の詳細を表示する。「イテレータ番号」、「モデルセット番号」、「グループ番号」、「正答率」を表示し、正答率の平均が、モデルの「スコア（性能）」となる。

② 「Draw EnrichmentCurve」タブ

「タンパク質」、「相互作用数」が表示され、「Draw」ボタンを押下すると、エンリッチメント曲線という、モデルの性能の指標の一つが表示される。

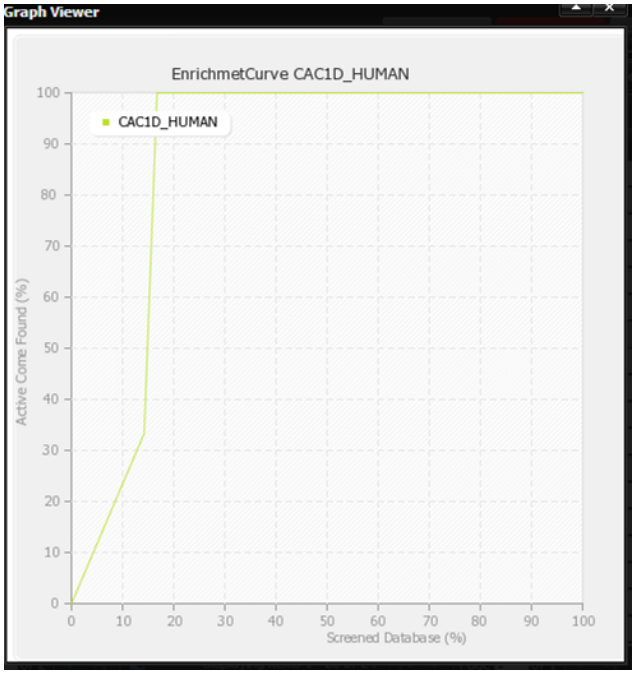


図４－４－３．エンリッチメント曲線

３．予測実行

１で選択されたデータ・パラメータを適切な形に変換する。

また、この処理は、一週間程度の長い処理時間を要する場合がある。よって進捗状況を確認できるインターフェースを設ける。

３－１．予測

① 学習モデルを選択し、予測したいタンパク質を選択する。

② 予測に使用する化合物を、ライブラリー（SDF）か、化合物描画検索によって作成される化合物群か選択する。

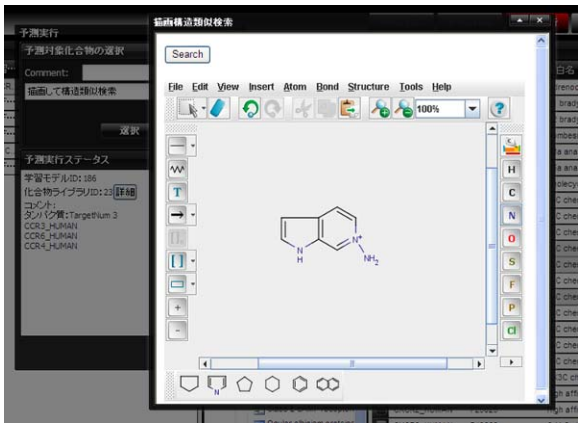


図４－４－４．化合物描画検索

３－２．ジョブ登録

① ジョブ管理

処理名、更新日、ステータス（完了・エラー・計算中）が表示される。

② イベントログの表示

ジョブを選択すると、処理のログが表示される。

４－４－２　まとめ

達成状況として、21 年度の目標通りのプロトタイプ版の作製が完了している。22 年度はこれらをもとに、大規模データでの使用に耐えうる最終仕様を決定し、最終製品版の開発を行う。

４－５　外部ソフトとの連携の為の連結モジュールの付加による統合システムの開発

４－５－１　外部ソフト（MOE）との連結モジュールの作製

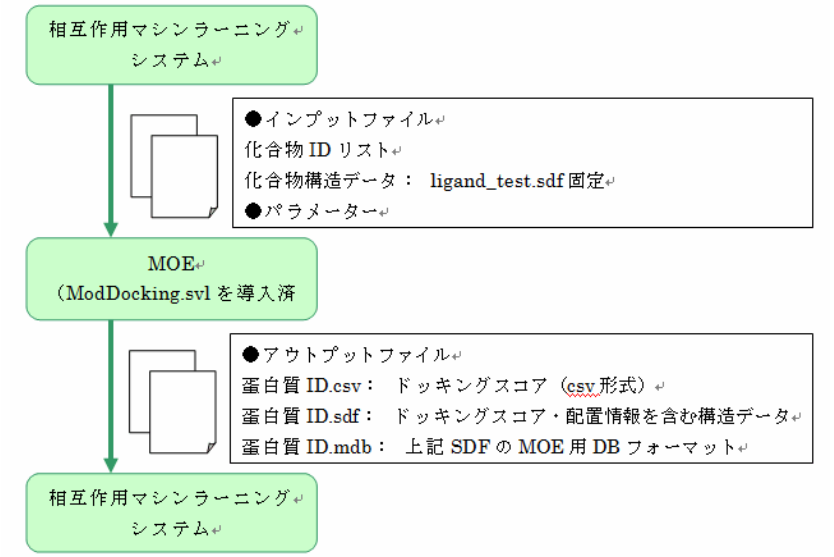
本モジュールは、相互作用マシンラーニングシステムと、タンパク質と化合物の結合部位を探索してドッキングシミュレーションを行う MOE システムの機能を連結させるモジュールである。

相互作用マシンラーニングシステムで予測スコアを計算するタンパク質と化合物に対して、MOE を用いたドッキングスコアも算出・加味することで、結合予測の精度向上に役立てる。

具体的には、タンパク質の ID・化合物の ID リスト・化合物の構造データ（標準の SDF 形式）を与えると、そのタンパク質と各化合物に対して一連のドッキングスコアを計算する。

相互作用マシンラーニングシステムは、その計算結果をインポートし、統合する。

１．　処理の流れ



２．　事前準備

事前に MOE をインストールする。（インストール方法は MOE のマニュアルを参照）

モジュール本体 SVL ファイル（ModDocking.svl）を、MOE インストールルート/lib/svl/user.svl/ディレクトリにコピーして MOE を起動する。

また、デフォルトのタンパク質 DB を MOE システムに導入しておく。

３．　連結パラメータ

① ドッキングスコアを計算したいタンパク質の ID（PDBID）一覧と、化合物 SDF ファイルを渡す。

② 配置計算メソッドを、「London dG」「ASE」「Affinity dG」「Alpha HB」から選択する。デフォルトは「London dG」となる。

４．　ドッキング計算

① 化合物構造データ SDF を、MOE 用データベース形式 MDB に変換する

② ドッキングサイトの探索
タンパク質立体構造に化合物があてはまるサイトを指定する。
本来はマニュアル操作でサイトを探索するのが望ましいが、自動化する場合は、最大サイトのみを使用する。

③ ドッキング実行
３．連結パラメータ②で選択した、配置計算メソッドを元に、ドッキングスコアを計算する。そのメソッドで最もスコアの高かった配置に決定し、SDF に構造情報を出力する。

④ スコアの再計算
で決まった配置に対し、全てのメソッドでスコアを再計算する。

４－５－２　まとめ

達成状況として、21 年度の目標通りのプロトタイプ版の作製が完了している。22 年度はこれらをもとに、大規模データでの使用に耐えうる最終仕様を決定し、最終製品版の開発を行う。

４－６　総括

平成 21 年度は、GPCR 及びイオンチャネルに関して大規模データを用いた予測モデル構築及び活性予測に対する実証実験を行い、十分な成果を得て精度向上を完了することが出来ている。22 年度に関しては、各キナーゼターゲットに対する選択性の予測について、予測モデルの精度向上の検証をおこなうものとする。

また、創薬シード・リード化合物探索システムの開発については、プロトタイプシステムの開発が順調に完了し、21 年度の目標を達成することが出来ている。22 年度については、最終製品仕様を確定し、外部システムとの連携による統合システムの開発を行う予定である。

５　参考資料

５－１　研究発表・講演等一覧

＜著書等＞
１.「次世代創薬テクノロジー　実践：インシリコ創薬の最前線」㈱メディカル・ドゥ，奥野恭史共著

＜一般口頭発表＞
１．日本薬物動態学会第２回ビジョン・シンポジウム　奥野恭史
「ケミカルゲノミクスに基づくインシリコ創薬」
２．構造活性フォーラム 2009 奥野恭史「多重標的創薬のためのインフォマティクス」
３．第５２回日本神経化学会大会　奥野恭史
「創薬シード化合物探索のための画期的な新規計算手法の紹介」
４．独立行政法人　産業技術総合研究所　生命情報科学技術者養成コース「平成 21 年度　創薬インフォマティクスコース　特別講義」奥野恭史「ケミカルゲノミクスに基づく化合物探索」
５．Bio Japan2009 奥野恭史、村上竜太、金井千里、藪内弘昭
「創薬シード化合物探索の為の画期的な新規計算手法」
６．イノベーション・ジャパン 2009 奥野恭史
創薬を加速化する医薬候補化合物の革新的探索計算技術の開発
７．医薬品産業情報研究会（PI フォーラム）奥野恭史
「ケミカルゲノミクス情報を用いたパターン認識による新規スクリーニング手法」
８．第 82 回日本生化学会大会　奥野恭史
「ケミカルゲノミクス情報に基づく創薬インフォマティクス」
９．CAC フォーラム　奥野恭史「ケミカル情報とバイオ情報の統合に基づくインシリコ創薬」
10．IT が拓く未来－創薬を加速するインフォマティクス　奥野恭史
「Chemical Genomics knowledge に基づく画期的 In silico 創薬手法の開発とそのシステム化」
11．第 3 回京都・大阪バイオクラスター連携プロジェクト　関西バイオネットワーク「創薬研究支援ツールの高度化推進」発表交流会　奥野恭史「インシリコ創薬支援ツールで加速するゲノム創薬」

＜その他資料＞
1．第 8 回国際バイオエキスポ・国際バイオフォーラム
村上竜太、金井千里、藪内弘昭　ポスター展示
2．JETRO NY 展示会　奥野恭史
A novel computational approach for drug discovery: Chemical genomics-based in silico ligand screening

５－２　産業財産権
該当なし