

ライフサイエンスデータベース統合推進事業
統合データ解析トライアル
研究開発課題
「機械学習を用いたタンパク質ーリガンド結合部位
予測ツールの自動生成パイプラインの開発」

研究開発終了報告書

研究開発期間：平成25年9月～平成26年1月

研究代表者：番野 雅城

(東京大学大学院 農学生命科学研究科、
博士課程)

§1 研究開発のねらい

生命システムの解明には、タンパク質のリガンド結合部位の解析が重要である。しかしながら、一般に生化学実験や構造決定などの実験的手法は、複雑な手順・高額な装置・研究者の熟練等が必要である。ゆえに、バイオインフォマティクス的手法を用いたリガンド結合部位予測・解析システムが期待されている。しかし、現在のところ、様々なリガンドに対して多数の異なる予測システムが開発されてきたが、その使用法は統一的でなく、十分な精度が得られていないものも多いのが現状である。また、タンパク質―リガンド結合部位に関するデータは日々蓄積されているが、それらも予測ツールに有効に活用されていない。本研究開発では、統一的なフレームワークで、利用者の目的にあった精度の高いリガンド結合部位予測ツールを、最新のデータを統合することにより、自動的に生成するシステムを構築する。生成される予測ツールは、立体構造が決定していないタンパク質に対する予測も可能であり、その計算コストは小さい。これによって、特定のリガンドに結合するタンパク質をデータベースから網羅的に探索することも可能である。

具体的には、利用者が指定するリガンドに対して、そのリガンドが結合するタンパク質の部位(リガンド結合部位)を予測するツールを自動生成するパイプラインを開発する。生成される予測ツールは、タンパク質のアミノ酸配列を入力とし、機械学習の手法を用いてリガンドの結合部位を予測するプログラムである。開発されるパイプラインは予測ツール生成と結合部位予測の2つの作業フローからなる。予測ツール生成のフローでは、予測ツールを生成するパイプラインのシステムワークフローを示す。利用者がリガンド(種)を入力すると、そのリガンド(種)の結合部位情報をデータベースから抽出して、それをもとに学習用データセットを作成する。また、結合部位の特徴抽出を行い、新規の予測ツールを生成する。結合部位予測フローでは、利用者は生成された予測ツールに対して、結合部位を予測したいタンパク質のアミノ酸配列を入力する。予測ツールはそのアミノ酸配列をもとにリガンド結合残基を予測する。加えて、本システムの開発・運用で新たに生成される予測ツールおよびデータベースは次世代のデータ相互参照(セマンティックウェブ)技術に準拠した再利用・再目的化に適した様式で整備する。

上記目的を達成するためのデータ資源として、平成 23、24 年度の統合化推進プログラムで PDBj によって開発された wwPDB/RDF、アミノ酸配列データベースである UniProt、及び PDB のタンパク質鎖と UniProt の相互参照情報を記述した EBI SIFTS を利用した。

§2 研究成果

本研究では、予測ツール生成を行うためのプログラム、及びそのバックエンドとして、データセット自動生成で使うリガンド結合残基データベース(PLBSPResidue)を開発した。

予測ツール生成を行うためのプログラムとして、データセット生成パイプラインと予測ツール生成パイプラインの2つを開発した。

データセット生成パイプラインは、リガンド結合タンパク質取得、配列冗長性除去、リガンド結合残基取得の3つの操作からなる。リガンド結合タンパク質取得の操作ではユーザーが指定したリガンドと結合するタンパク質を本研究で開発した基盤データベースから検索し、対応するアミノ酸配列を取得する。配列冗長性除去の操作では、アミノ酸及び DNA 配列の冗長性除去ツールである CD-HIT を用いて配列類似性 50%、カバー率 50%の条件で配列冗長性の除去を行う。リガンド結合タンパク質取得の操作では、リガンド種とタンパク質を指定して、それに対応するリガンド結合部位の一覧を取得する。

予測ツール生成パイプラインではデータセット生成パイプラインで取得されたデータセットを基に機械学習アルゴリズムの一種であるサポートベクターマシン(SVM)を用いてリガンド結合部位予測ツールを生成する。SVM の学習には予測対象を特徴ベクトルに変換すること、正解を表す正例データセットと不正解を表す負例データセットを用意すること、及び学習パラメーターであるコストパラメーターCとガウシアンカーネルのパラメーター γ を最適化することが必要である。特徴ベクトルには PSI-BLAST から生成された位置特異的スコア行列(PSSM)を基に生成した。PSSM のカラムを配列中で連続した w 残基分結合し $20 \times w$ 次元の特徴ベクトルに変換し、中心の残基がリガ

ンド結合残基であるかどうかを予測した。今回は正例としてリガンド結合残基を与え、負例にはデータセットに利用したタンパク質中でリガンド結合残基から 5 残基から 25 残基離れた全ての残基を与えた。パラメーター最適化には、プログラム側で自動的に最適なパラメーターを決定できるようにするため遺伝的アルゴリズムを用いた。

汎用パイプラインシステムである Galaxy を用いて、これらパイプラインを組み合わせてリガンド結合部位予測ツール生成ワークフローを構築し公開した。

本ワークフローを用いて鉄イオン結合タンパク質、脂質結合タンパク質、プリンヌクレオチド結合タンパク質についてそれぞれリガンド結合部位予測ツールを生成し、交差検定による性能評価を行った。その結果を表 1 に示した。この結果からデータベースの情報を基に機械的に高精度な予測ツールを生成することが示せた。

PLBSPResidue は、リガンド結合残基情報、UniProt—PDB 残基番号対応表、PDB ファイル中のリガンド間共有結合情報からなるグラフデータベースである。ここに登録されている情報は、wwPDB/RDF と相互参照可能な RDF ファイルの形式で保存されている。これによって、PDB から UniProt、ChEBI など RDF やオントロジーを公開しているデータベースに対しても相互参照可能になる。

リガンド結合残基情報は、PDB ファイル中のどの残基がどのリガンドと 5Å 以内に接近したかを記述したデータである。研究代表が所属する研究室では PDB の座標データを八分木で表現し、タンパク質—リガンド間の距離を高速に計算するアルゴリズムを開発しており、PDB 中に存在する全リガンドとタンパク質の結合部位の情報を網羅的に収集し、このデータを RDF 化しグラフデータベースに実装した PLBSP (Protein Ligand Binding Site Pair) を保有している。このデータベースはリガンド結合残基の全原子情報を保持しているため、データセット生成の度にこのデータベースに問い合わせるには応答速度が遅くなってしまった。そのため、残基とリガンドの情報のみに限定した軽量で小規模な RDF ファイルを作成した。

UniProt—PDB 間残基番号対応表は、PDB ファイル中の残基番号が UniProt に登録されているアミノ酸配列上のどこに対応するかを記述したデータである。PDB では、同じタンパク質に対して複数回条件を変えて実験したものが登録されている場合や、同じタンパク質を部分ごとに構造を決定する場合がある。それらの結合残基情報を一つのアミノ酸配列上に対応づけるために、UniProt、及び PDB と UniProt の相互参照情報を記述した EBI SIFTS の情報を元に RDF ファイルを作成した。

PDB ファイル中のリガンド間共有結合情報は、PDB ファイル中でリガンド間の共有結合情報を記述したデータである。PDB ファイル中ではリガンドに対して HETATM コードと呼ばれる 3 桁の ID が振られているが、リガンド同士が共有結合することによって全体で別の分子を形成していることがある。そのため、HETATM コードのみを用いてリガンドを特定することは不十分であったため、PDB ファイル中のどのリガンドとどのリガンドが共有結合しており、全体でどのような分子を形成しているかを記述した RDF ファイルを作成した。

表1 生成されたリガンド結合部位予測ツールの予測精度。

ツール名	感度 [%]	特異度 [%]	AUC [%]
脂質結合部位予測ツール	59.2	75.6	74.2
鉄原子結合部位予測ツール	33.2	99.1	87.0
ブリン結合部位予測ツール	31.3	98.0	82.3

(1) ツール名: **UTProt Galaxy**

概要: 汎用的なパイプラインシステムである Galaxy 上に、本研究で開発した予測ツール生成パイプラインを開発した

アクセス方法: プロジェクトのトップページ (<http://utprot.net/>) から “TOP > UTProt Galaxy > 上段 Shared Data > Published Workflows”。

構築したワークフローについて下記の 4 つとなる。

(a) ワークフロー名: **Predict Ligand Binding Residue**

予測対象となるタンパク質のアミノ酸配列の FASTA ファイルと、対象となるリガンドの HETATM コードを入力すると、SVM を用いてリガンド結合部位予測ツールを生成し、各残基のリガンド結合残機らしさを示したスコアを返す。

(b) ワークフロー名: **Create Ligand Binding Residue Predictor**

対象となるリガンドの HETATM コードを入力すると、SVM を用いてリガンド結合部位予測ツールを生成する。

(c) ワークフロー名: **Get Non Redundant Ligand Binding Protein With Ligand Binding Residue**

対象となるリガンドの HETATM コードを入力することで、リガンド結合残基データベースから入力リガンドと結合するタンパク質のリガンド結合残基と配列冗長性が除去されたアミノ酸配列を取得する。

(d) ワークフロー名: **Get Non Redundant Ligand Binding Protein**

対象となるリガンドの HETATM コードを入力することで、リガンド結合残基データベースから指定リガンドと結合するタンパク質の配列冗長性が除去されたアミノ酸配列を取得する。

(2) ツール名: **リガンド結合残基データベース**

概要: PDB 中にある 5Å 以内に接近した残基とリガンドの組を網羅したグラフデータベース。

グラフデータベースである AllegroGraph 上で構築されており、SPARQL などの問い合わせ言語を用いてデータ取得が行える。

アクセス方法: プロジェクトのトップページ (<http://utprot.net/>) から “TOP > PLBSP Residue” へリンクをたどる。

(3) データベース名: **データセット生成、予測ツール生成パイプライン**

概要: CUI で動作するデータセット生成パイプライン及び予測ツール生成パイプライン。

アクセス方法: <http://github.com/masakibb2/xbind>

(4) データベース名: **UTProt CKAN**

概要: 汎用的なオープンデータポータル管理システムである CKAN を、予測ツール管理バックエンドとして公開した。

アクセス方法: プロジェクトのトップページ (<http://utprot.net/>) から “TOP > UTProt

CKAN” ヘルリンクをたどる。

§3 研究開発計画および計画に対する達成状況

(1) 達成状況

研究開発実施計画にあったツール生成ワークフロー、およびデータセット生成ワークフローは CUI アプリケーションの形で開発し、GitHub 上で公開した。システム公開については、上記プログラムを用いて、汎用パイプラインシステム Galaxy 上に、Web アプリケーションの形でリガンド結合部位予測ツール生成ワークフローを公開した。予測ツール管理のバックエンドについては、CKAN を用いて公開した。

当初計画にはなかったが、研究開発を進めていく上でリガンド結合残基データベース PLBSPResidue を開発し公開した。

研究開発実施計画の中にあるリガンド検索インターフェースについては、独自の実装をするよりも、PDBeChem や LigandExpo のような既に公開されているシステムを用いた方が望ましいと考えたので実装しなかった。

(2) ツールの将来性への展望

タンパク質リガンド結合部位予測以外の予測問題についての予測ツール生成パイプライン開発が挙げられる。データベースよりデータセットを収集し機械学習を用いた相互作用を予測するというデータ駆動型の研究は、リガンド結合部位予測に限らず、リガンド結合タンパク質予測やタンパク質間相互作用部位予測など様々な形で行われている。特にリガンド結合部位そのものを予測する手法は、使用するデータセットについて共通する部分が多く本システムを基にすることで容易にパイプライン化できると考えられる。

本研究開発をモデルケースに様々な問題に対して予測ツール生成パイプライン化が進むことで、実験研究者の要望に対して今以上に柔軟な対応ができると期待される。

次に、タンパク質-リガンド相互作用情報について、PDB 以外のデータベースも活用することである。機械学習を用いた手法はデータセットが十分存在する上で初めて有効である。つまり、立体構造情報に限らず生化学的に決定された相互作用情報も取り入れることで今以上に様々なリガンドの予測ツールの生成が可能になると考えられる。特に統合データベースの一つである糖鎖データベース JCGGDB は RDF 化が進められており、本研究で開発したリガンド結合残基データベースと容易に相互参照が行えると考えられる。例として、レクチンフロンティアデータベースがあげられる。このデータベースは生化学実験によって決定されたレクチンと糖鎖の相互作用情報を集めたデータベースであり、JCGGDB に含まれる。糖鎖は結晶化することが難しいものもあることから、PDB 中に登録されている構造の数も少ない。PDB に含まれない相互作用情報を補完するリソースとして期待できる。

加えて、実験者も使いやすいデータセットの取捨選択を支援するインターフェースの開発を行う。本研究ではデータセットを機械的に生成するため、専門家から見ても不適当なデータセットが含まれてしまう可能性がある。またユーザーが予測したい問題によっては、特定の代謝経路に関わるタンパク質やリガンドのみに絞りたいなど、どのデータセットの選択についても個別の要望が出てくると考えられる。こうした問題に対応するため、ユーザーにシステムが作り上げたデータセットの妥当性について検証してもらうことで、予測の精度の向上が期待できる。その実現のために、本システムで用いるデータセットをセマンティックウェブの技術に準拠している利点を活用し、そのデータセットに関連する情報が複数のデータベースから横断的に取得し、一つの画面に整理し表示するようなユーザーインターフェースが考えられる。ユーザーはデータセットに含まれるタンパク質やリガンドに関する情報を基に、現在作成しようとしている予測ツールのデータセットとしてふさわしいか判断する。特にリガンドの情報に関しては、統合データベースの一つである LinkDB の情報が充実している。LinkDB は様々なゲノムネット化合物データベースの統合化を行い、各レコード間の関係性を記述した RDF などを提供している。これにより、データセットに含まれる化合物が関わる反応経路や薬効など様々な情報を取得できると考えられる。

§4 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○番野 雅城	東京大学大学院 農学生命科学研究科	博士課程	システムの設計開発	H25.10-H26.1

§5 成果発表等

(1)原著論文発表 (国内(和文)誌 1件、国際(欧文)誌 0件)

1. 込山悠介、**番野雅城**、鏈水優行、加藤文彦、大向一輝、武田英明、清水謙多郎、生命科学の複数 LOD の統合による目的別タンパク質間分子間相互作用 Linked Open Data の構築、人工知能学会誌、掲載誌 29 巻、4 号、平成 26 年 7 月 (査読中、minor revision)

(2)その他の著作物(総説、書籍など)

(3)国際学会発表及び主要な国内学会発表

① 招待講演 (国内会議 0件、国際会議 0件)

② 口頭発表 (国内会議 1件、国際会議 0件)

1. ○込山悠介(東京大学大学院農学生命科学研究科)、**番野雅城**(同上)、Gul Saad(同上)、鏈水優行(同上)、植木快(東京大学農学)、清水謙多郎(東京大学大学院農学生命科学研究科)、UTProt: Database Integration and Tool Development for Interactomics Utilizing Biosemantics(UTProt:バイオセマンティクスを用いた網羅的な分子間相互作用学のためのデータベース統合とツール開発)、2013 Annual Convention of Japanese Society for Bioinformatics、東京、平成 25 年 10 月

③ ポスター発表 (国内会議 3件、国際会議 1件)

1. ○**番野雅城**(東京大学大学院農学生命科学研究科)、込山悠介(同上)、中村周吾(同上)、清水謙多郎(同上)、Development of system that predict sugar binding protein and sugar binding site based on protein sequence using support vector machine (サポートベクターマシンを用いたタンパク質配列に基づく糖結合部位を予測するシステムの開発)、2013 Annual Convention of Japanese Society for Bioinformatics、東京、平成 25 年 10 月

2. ○込山悠介(東京大学農学生命科学研究科)、**番野雅城**(同上)、Gul Saad(同上)、鏈水優行(同上)、植木快(東京大学農学)、清水謙多郎(東京大学大学院農学生命科学研究科)、UTProt: Database Integration and Tool Development for Interactomics Utilizing Biosemantics(UTProt:バイオセマンティクスを用いた網羅的な分子間相互作用学のためのデータベース統合とツール開発)、2013 Annual Convention of Japanese Society for Bioinformatics、東京、平成 25 年 10 月

3. ○植木快(東京大学農学生命科学研究科)、込山悠介(同上)、**番野雅城**(同上)、中村周吾(同上)、清水謙多郎(同上)、Sequence derived predictor of lipid-binding proteins with support vector machine(サポートベクターマシンを用いた脂質結合タンパク質の配列由来の予測器)、2013 Annual Convention of Japanese Society for Bioinformatics、東京、平成 25 年 10 月

(4)知財出願
なし

(5)受賞・報道等

① 受賞

Linked Open Data Challenge Japan 2013 において入賞

② マスコミ(新聞・TV等)報道

なし

③ その他

なし

§6 自己評価

研究開発の目標である「利用者が指定するリガンドに対して、そのリガンドが結合するタンパク質の部位(リガンド結合部位)を予測するツールを自動生成するパイプライン開発については、Webアプリケーションとして予測ツール生成プログラムを 公開し、高精度な予測ツールが生成されることを確認する評価試験まで行うことができた。これは当初の研究開発のねらいを達成できたといえる。しかしながら出来上がった予測ツールをユーザー間で共有する枠組みの整備や予測結果を可視化する方法の検討など課題を残す。可能な限りユーザーからのユースケースを集めより改良し、完成度の高い Web アプリケーションにしていきたい。

以上

ライフサイエンスデータベース統合推進事業
統合データ解析トライアル
研究開発課題
「機械学習を用いたタンパク質－リガンド結合部位
予測ツールの自動生成パイプラインの開発」

研究開発終了報告書 添付資料

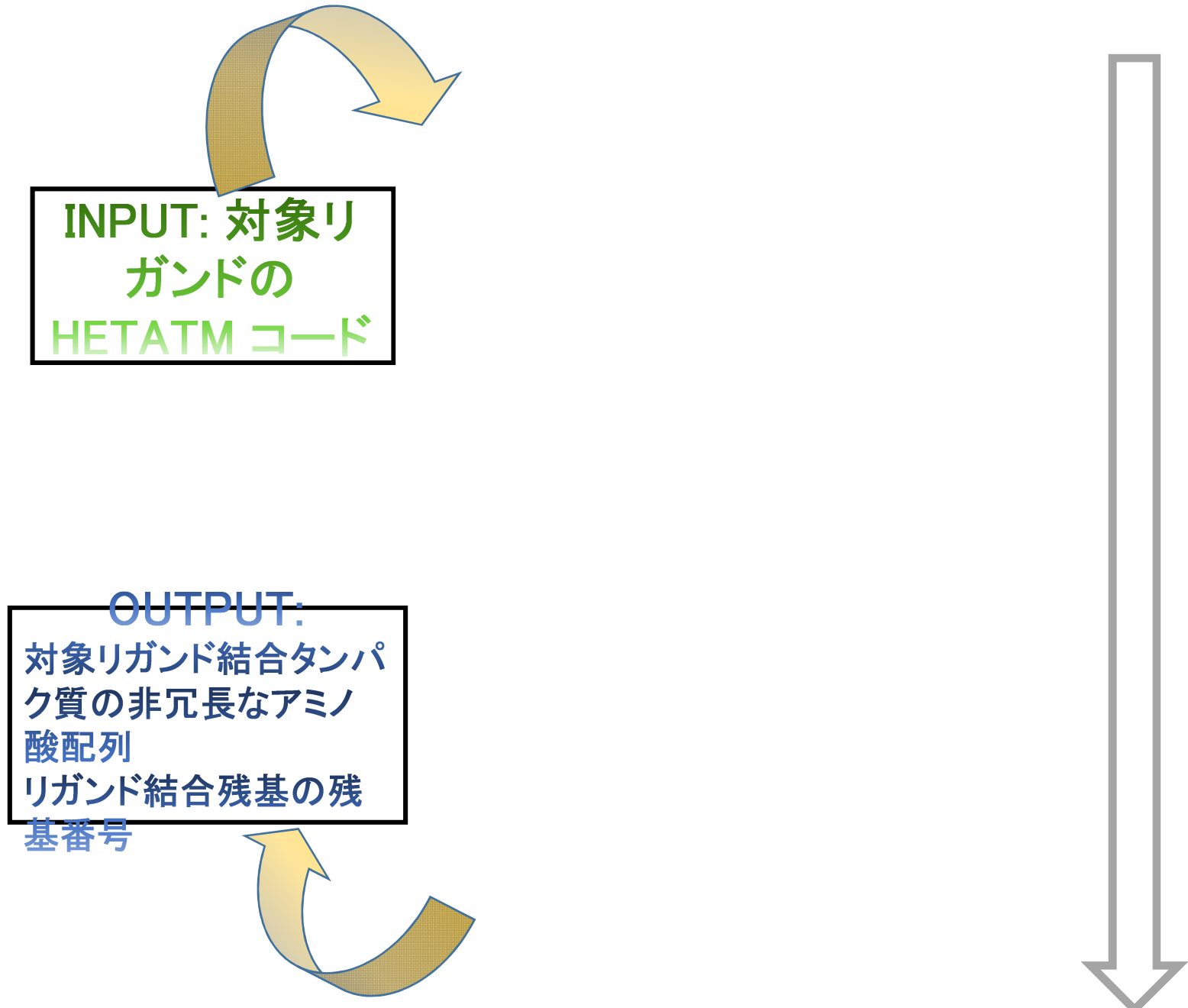
研究開発期間： 平成25年9月～平成26年1月

研究代表者： 番野 雅城
東京大学 大学院 農学生命科学研究科
博士課程

目次

- データセット生成パイプライン、予測ツール生成パイプライン
... 3～ 6
- リガンド結合部位予測ツール生成ワークフロー
... 7 ～17
- リガンド結合残基データベース
... 18 ～ 28

データセット生成パイプラインの流れ



予測ツールの生成パイプラインの流れ

1

.....YYSVSAFGKNTSAI.....
.....YWDISGPGAGLENI.....
.....CPDVSSTDIEYVFL.....

PSSMに変換
(PSI-BLAST)

A	-2	-3	-2	-2	-1	3	-2
R	-3	-3	0	-1	-3	-1	-2
D	-1	-4	4	6	-3	0	-2
C	-4	-5	0	1	-3	-1	-4
...	...	...	...	...	...	...	...

N末端側2残基分
のカラム

+

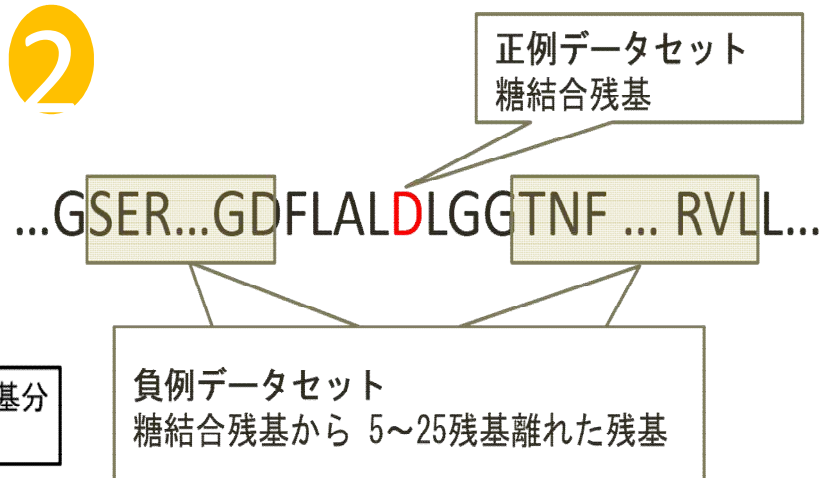
中心残基
のカラム

+

C末端側2残基分
のカラム

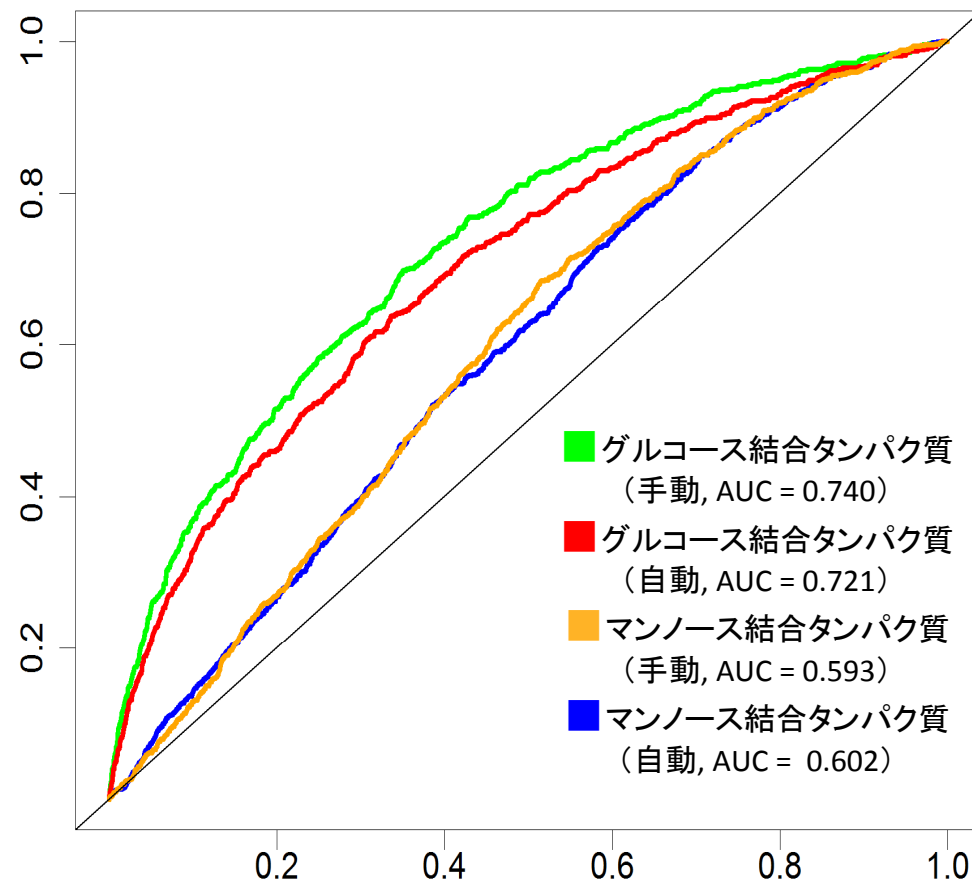
PSI-BLASTを用いてアミノ酸配列を
PSSMに変換。
中心から w 残基分のカラムを特徴ベク
トルとして
使用する。

2



入力されたリガンド結合部位残基情報をもとに、リガンド結合残基を正例データセット、リガンド結合残基から5~25残基離れた残基を負例データセットとして用いてサポートベクターマシンで学習させる。パラメーター探索には遺伝的アルゴリズムを用いる。

データセット生成パイプラインと 予測ツール生成パイプラインの性能評価



遺伝子 : $c, g, w \in [-14, 14]$

評価関数:

$$\text{Window} = 2 \times (|w| + 1)$$

$$C = 2^c$$

$$\gamma = 2^g + \frac{1}{20 \times \text{Window}}$$

(C, γ, window) を使って
クロスバリデーションし
たときのAUC

変異:

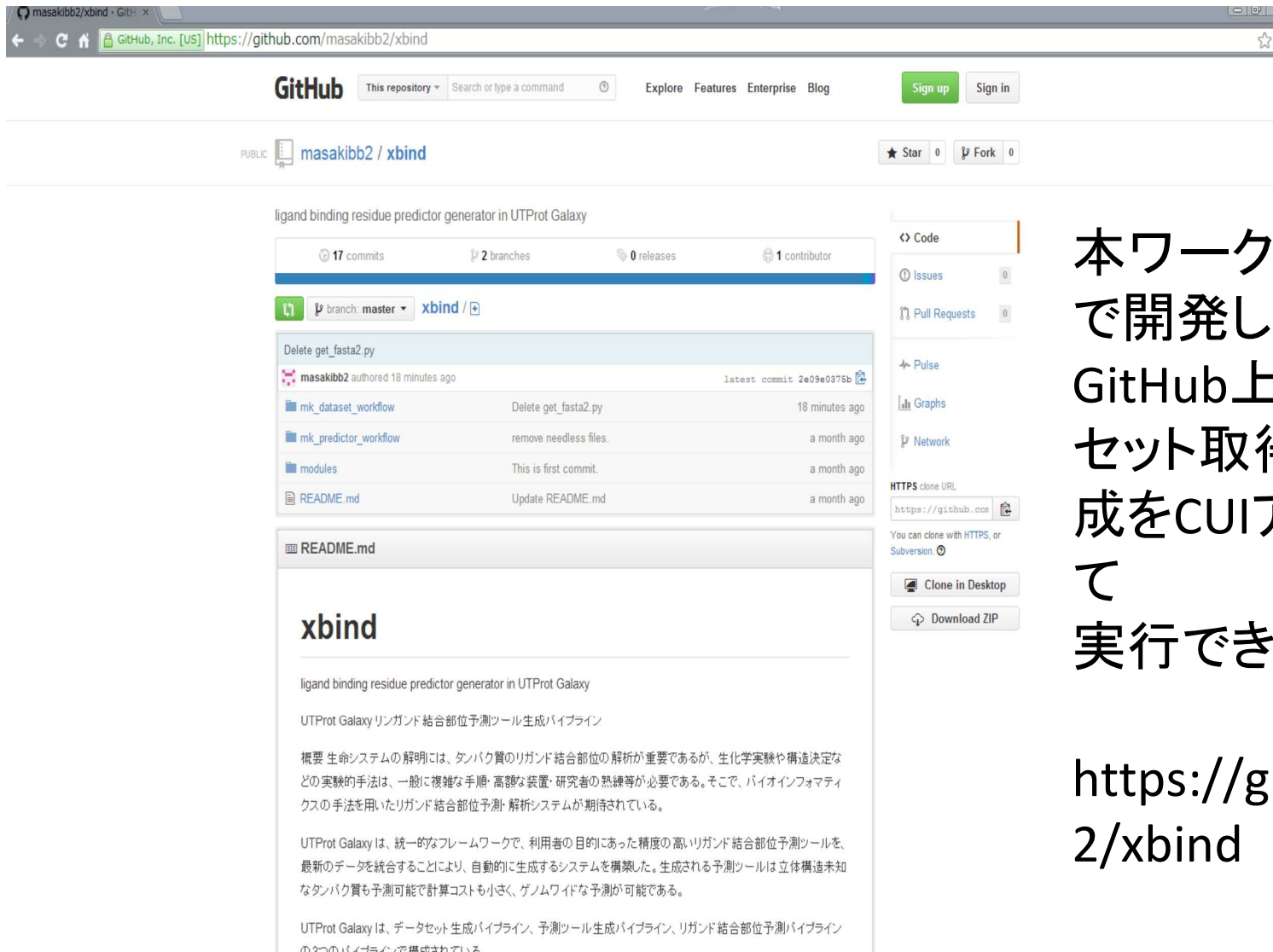
Point mutation 5%

Cross over 80 %

初期集団: 20, 世代数: 3

グルコース結合タンパク質とマンノース結合タンパク質について、
データセット生成パイプラインと予測ツール生成パイプラインを通して実行することで
結合部位予測ツール生成を行った。

データセット生成、予測ツール生成パイプラインの利用



GitHub repository page for `masakibb2 / xbind`. The repository is public and contains a README.md file. The README describes `xbind` as a "ligand binding residue predictor generator in UTProt Galaxy". It mentions that the tool is designed for ligand binding residue prediction in UTProt Galaxy, which is a platform for protein-ligand interaction analysis. The README also notes that the tool is designed to be easy to use and can be used for various purposes, including predicting ligand binding residues and generating prediction pipelines. The repository has 17 commits, 2 branches, 0 releases, and 1 contributor. The latest commit is "Delete get_fasta2.py" by "masakibb2" 18 minutes ago. The repository also has a "mk_dataset_workflow" and a "mk_predictor_workflow".

本ワークフローを構築する上で開発したプログラムをGitHub上で公開した。データセット取得や予測ツールの生成をCUIアプリケーションとして実行できる。

<https://github.com/masakibb2/xbind>

UTProt Galaxy

上段メニューのUsers を押して
Registration でユーザー登録



UTProt Galaxy にユーザー登録することで、
上段メニュー Shared Data > Published Workflows
から今回の試作版ワークフローが利用できる。

UTProt Galaxy の利用(1)

The screenshot displays the UTProt Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'ワークフロー', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area is titled 'Published Workflows' and features a search bar and a table of workflows. The table lists several workflows by masaki, including 'Create Ligand Binding Residue Predictor', 'Get Non Redundant Ligand Binding Protein', 'Get Non Redundant Ligand Binding Protein With Ligand Binding Residue', 'Pipeline Prediction Using Sugar Binding Protein and Residue Predictor', and 'Predict Ligand Binding Residue'. The 'Predict Ligand Binding Residue' workflow is selected, showing its details on the right. The 'Import workflow' button is highlighted with a red circle.

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
Create Ligand Binding Residue Predictor		masaki	★★★★★		
Get Non Redundant Ligand Binding Protein		masaki	★★★★★		
Get Non Redundant Ligand Binding Protein With Ligand Binding Residue		masaki	★★★★★		
Pipeline Prediction Using Sugar Binding Protein and Residue Predictor	work flow of predicting sugar binding residue	masaki	★★★★★		
Predict Ligand Binding Residue		masaki	★★★★★		

Galaxy Workflow 'Predict Ligand Binding Residue'

Annotation:

Step 1: GET ligand binding proteins.
HETATOM code

Step 2: PSSM generator using psi-blast
Source Data: select at runtime
Times of iterator: 2

Step 3: GET fasta using UniProt URI.
Source Data: Output dataset 'output' from step 1

Step 4: Remove Sequence Redundancy using CD-HIT.
Source Data: Output dataset 'output' from step 3

Author: masaki
Related Workflows: All published workflows Published workflows by masaki
Rating: Community (0 ratings, 0.0 average)
Tags: Community: none

メニューの Shared Data を選択すると、本システムが提供しているワークフローの一覧が表示される。ここでは「Predict Ligand Binding Residue」を選択する。

Import workflow を選択することでワークフローメニューに表示される。

UTProt Galaxy の利用(2)



左側メニュー「Get Data」を選択



「Upload File」を選択

続いて、予測対象となるタンパク質を UTProt上にアップロードする。

上段Analyze Data のメニューを選択し、左側メニューの、Get Dataを選択する。

UTProt Galaxy の利用(3)



ファイルアップロードする場合、ファイル選択を選ぶ

ファイルアップロードかコピー&ペーストのいずれかの方法でFASTAファイルがアップロードができる。

コピー&ペーストする場合は、URL/Text: のフォームに貼り付ける。

アップロードが完了すると、中央にメッセージが表示され、左側に、「Pasted Entry」という名前がついて Galaxy 上に保存される。



Galaxy 上に保存されたデータはこちらに表示される。

UTProt Galaxy の利用(4)

Workflow home x

utprot.net:8080/workflow

Galaxy Analyze Data ワークフロー Shared Data Visualization Help User Using 1.2 MB

Your workflows Create new workflow Upload or import workflow

Name	# of Steps
imported: Predict Ligand Binding Residue ▼	10

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu

Workflow home x

utprot.net:8080/workflow

Galaxy Analyze Data ワークフロー Shared Data Visualization Help User Using 1.2 MB

Your workflows Create new workflow Upload or import workflow

Name	# of Steps
imported: Predict Ligand Binding Residue ▼	10

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu

Edit
Run
Share or Publish
Download or Export
Copy
Rename
View
Delete

メニューが開かれるので「Run」を押す。

ワークフローメニューを押すと先ほど選択したワークフローが表示される。そこから、▼マークを押し、Run を押すと実行画面が開く。

UTProt Galaxy の利用(5)

The screenshot shows the UTProt Galaxy web interface. The browser address bar displays `utprot.net:8080/root?workflow_id=3f5830403180d620`. The main area shows a workflow titled "imported: Predict Ligand Binding Residue" with several steps. Two green callout boxes provide instructions:

- Step 2 のSource Data に先ほどアップロードしたFASTAファイルを選択** (Select the FASTA file uploaded in Step 2's Source Data)
- 上段下段のHETATM Code に予測対象とするリガンドのHETATMコードを入力** (Enter the HETATM code of the ligand to be predicted in the upper and lower HETATM Code fields)

Step1, Step7 のHETATMコードには予測対象となるリガンドのHETATMコードを入力
Step2 で先ほどアップロードしたFASTAファイルを選択し、「Run Workflow」ボタンを押すとワークフローが実行される。**※ 半日から一日で予測ツールが生成される。**

UTProt Galaxy の利用(6)

The screenshot shows the UTProt Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'ワークフロー', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area is divided into three panels:

- Left Panel (Tools):** A search bar and a list of tools under the 'Bilab' category, including 'Get header from Fasta file', 'PSSM generator using psi-blast', 'Remove Sequence Redundancy using CD-HIT', 'GET ligand binding proteins', 'GET fasta using UniProt URI', 'GET binding residue using UniprotURI and HETATM ID', 'Fit Prediction Model using Support Vector machine and Genetic Algorithm', 'Predict Binding Residue using generated Prediction model', 'Create Prediction Model', 'Sugar Binding Residue Predictor', 'Sugar Binding Protein Predictor', and 'Acid Sugar Binding Protein'.
- Center Panel:** A warning message states 'This dataset is large and only the first megabyte is shown below.' Below this, a list of UniProt URLs and their corresponding decision values is displayed. The URLs are all from 'http://purl.uniprot.org/uniprot/P12337'. The decision values range from -0.120607500021 to -0.377665385729.
- Right Panel (History):** A 'History' panel showing the workflow steps. The top step is '39: Predict Binding Residue using generated Prediction model on data 38, data 36, and data 23' with 30,388 lines of output. Below it are steps '38: Create Prediction Model on data 24, data 36, and data 23' and '37: Fit Prediction Model'.

タンパク質名、残基番号、結合部位予測(結果が1 ならば結合部位と予測したと判断)、decision value (結合残基らしさのスコア) がテキスト形式で結果が返される。

UTProt Galaxy の活用(1)

- UTProt Galaxy は次のようなワークフローが実装されている。

ワークフロー名	概略	入力	出力
Predict Ligand Binding Residue	予測対象となるタンパク質のアミノ酸配列のFASTAファイルと、対象となるリガンドのHETATMコードを入力すると、SVMを用いてリガンド結合部位予測ツールを生成し、各残基のリガンド結合残機らしさを示したスコアを返す。	結合部位を予測したいタンパク質のFASTAファイル、対象リガンドのHETATM Code	各残基の結合残基らしさを表した予測スコア
Get Non Redundant Ligand Binding Protein With Ligand Binding Residue	対象となるリガンドのHETATMコードを入力することで、リガンド結合残基データベースから入力リガンドと結合するタンパク質のリガンド結合残基と配列冗長性が除去されたアミノ酸配列を取得する。	対象となるリガンドのHETATMコード	入力したリガンドと結合するタンパク質の配列冗長性が除去されたFASTAファイル、及びリガンド結合残基の一覧
Get Non Redundant Ligand Binding Protein	対象となるリガンドのHETATMコードを入力することで、リガンド結合残基データベースから指定リガンドと結合するタンパク質の配列冗長性が除去されたアミノ酸配列を取得する。	対象となるリガンドのHETATMコード	入力したリガンドと結合するタンパク質の配列冗長性が除去されたFASTAファイル

UTProt Galaxy の活用(2)

『Predict Ligand Binding Residue』の入力画面

Galaxy

utprot.net:8080/root?workflow_id=5969b1f7201f12ae

Galaxy Analyze Data ワークフロー Shared Data Visualization Help User Using 96.5 MB

Running workflow "Predict Ligand Binding Residue"

Expand All Collapse

History

Unnamed history 0 bytes

1 ヒストリーは空です。解析を始めるには、左パネルの「データ取得」をクリック

search tools

Tools

BiLab

Get Data

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Wavelet Analysis

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

NGS: QC and manipulation

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

NGS: Simulation

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Phenotype Association

VCF Tools

Step 1: GET ligand binding proteins from HETATM Code, (version 1.0.0)

HETATM code

MAN

Step 2: PSSM generator using psi-blast (version 1.0.0)

Source Data

5: Remove Sequence R... on data 4

Times of iterator

2

Step 3: GET fasta using UniProt URI, (version 1.0.0)

Step 4: Remove Sequence Redundancy using CD-HIT, (version 1.0.0)

Step 5: Get header from Fasta file, (version 1.0.0)

Step 6: PSSM generator using psi-blast (version 1.0.0)

Step 7: GET binding residue using UniprotURI and HETATM ID, (version 1.0.0)

HETATM code

MAN

Source Data

Output dataset 'output' from step 5

Step 8: Fit Prediction Model using Support Vector machine and Genetic Algorithm, (version 1.0.0)

Step 9: Create Prediction Model, (version 1.0.0)

Step 10: Predict Binding Residue using generated Prediction model, (version 1.0.0)

☐ Send results to a new history

Run workflow

予測したいタンパク質
のアミノ酸配列 (FASTA
フォーマット)
Get Data → Upload File
を使ってアップロード

上段下段のHETATM
Code に予測対象とす
るリガンドのHETATM
コードを入力

UTProt Galaxy の活用(3)

『Get Non Redundant Ligand Binding Protein With Ligand Binding Residue』 の入力画面

The screenshot shows the Galaxy web interface for a workflow titled "Running workflow 'Get Non Redundant Ligand Binding Protein With Ligand Binding Residue'". The workflow consists of five steps:

- Step 1: GET ligand binding proteins from HETATM Code. (version 1.0.0)
- Step 2: GET fasta using Uniprot URI. (version 1.0.0)
- Step 3: Remove Sequence Redundancy using CD-HIT. (version 1.0.0)
- Step 4: Get header from Fasta file. (version 1.0.0)
- Step 5: GET binding residue using UniprotURI and CD-HIT. (version 1.0.0)

The input section for the workflow is visible, showing two "HETATM code" input fields, both containing the text "MAN". Below these fields, the "Source Data" is listed as "Output dataset 'output' from step 4". There is a checkbox for "Send results to a new history" and a "Run workflow" button.

On the right side, there is a "History" panel showing "Unnamed history" with "0 bytes". A message box states: "ヒストリーは空です。解析を始めるには、左パネルの 'データ取得' をクリック".

上段下段のHETATM
Code に予測対象とす
るリガンドのHETATM
コードを入力

UTProt Galaxy の活用(4)

『Get Non Redundant Ligand Binding Protein』の入力画面

The screenshot shows the Galaxy web interface for running a workflow. The browser address bar shows `utprot.net:8080/root?workflow_id=ebfb8f50c6abde6d`. The Galaxy header includes navigation links: **Analyze Data**, ワークフロー, Shared Data, Visualization, Help, and User. On the left, a sidebar lists various tools under the heading ツール, including Bilab, Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, and Fetch Alignments. The main content area is titled "Running workflow 'Get Non Redudant Ligand Binding Protein'" and contains three steps:

- Step 1: GET ligand binding proteins from HETATM Code. (version 1.0.0)
 - HETATOM code:
- Step 2: GET fasta using UniProt URI. (version 1.0.0)
- Step 3: Remove Sequence Redundancy using CD-HIT. (version 1.0.0)

Below the steps, there is a checkbox labeled "Send results to a new history" and a "Run workflow" button. A green callout box points to the HETATOM code input field with the text: "HETATM Code に予測対象とするリガンドの HETATM コードを入力".

PLBSPResidueデータベースの構築

既存の問題点

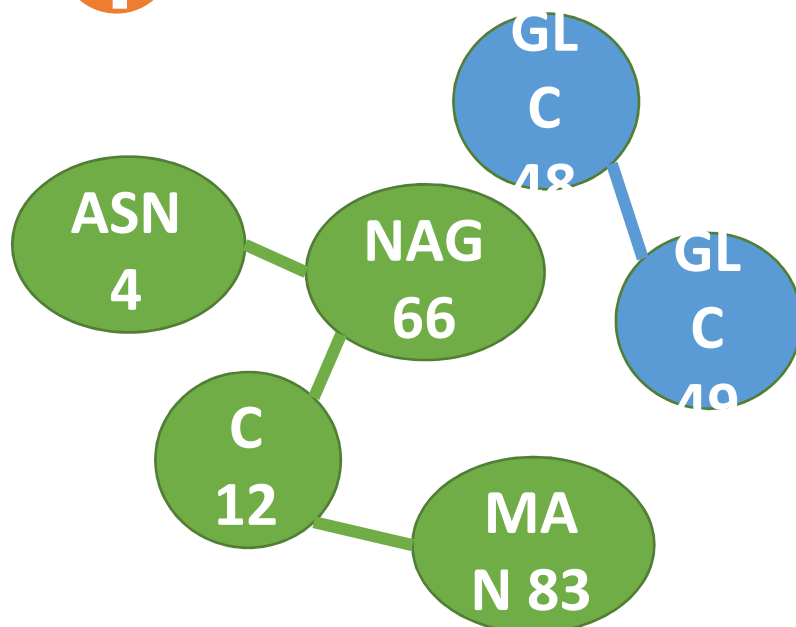
- PDB中の全タンパク質のリガンド間原子間距離を網羅したデータベース (PLBSP) にデータ問い合わせを行ったが、データ取得が非常に遅かった。
- PDB中の残基番号とFASTAファイル上のアミノ酸残基位置の対応付けが難しい場合がある。
- PDBファイルだけでは、リガンドであるか残基修飾であるか判断が難しい。

解決策

- PDBファイル中で 5Å 以内に存在する残基とリガンドの関係を記述した、軽量の相互作用データベースを構築した。
- UniProtとPDBの残基の対応関係がキュレーションされたEBI SIFTSをRDF化し本データベースに組み込んだ。
- リガンドの共有結合情報は、PDBファイル中のCONNECTレコードから抽出した(次スライド)。

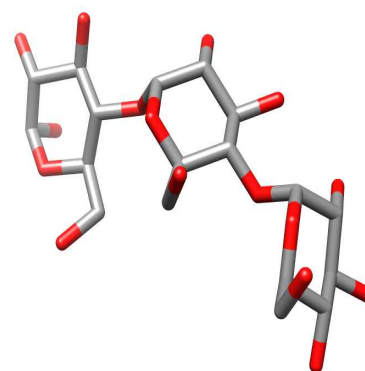
リガンド間共有結合情報のRDF化

1



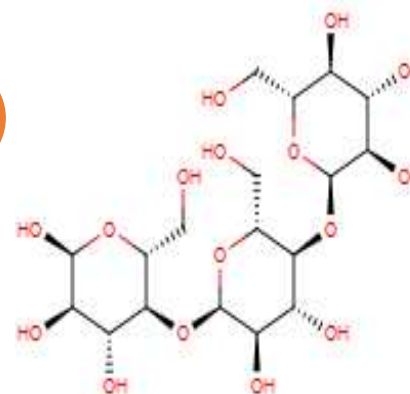
PDBのCONNECTレコードから残基をノード、共有結合関係をエッジとしたグラフデータを生成。経路発見アルゴリズムを用いて、一つの分子を形成するHETATMレコードの組み合わせをグループ化する。

2



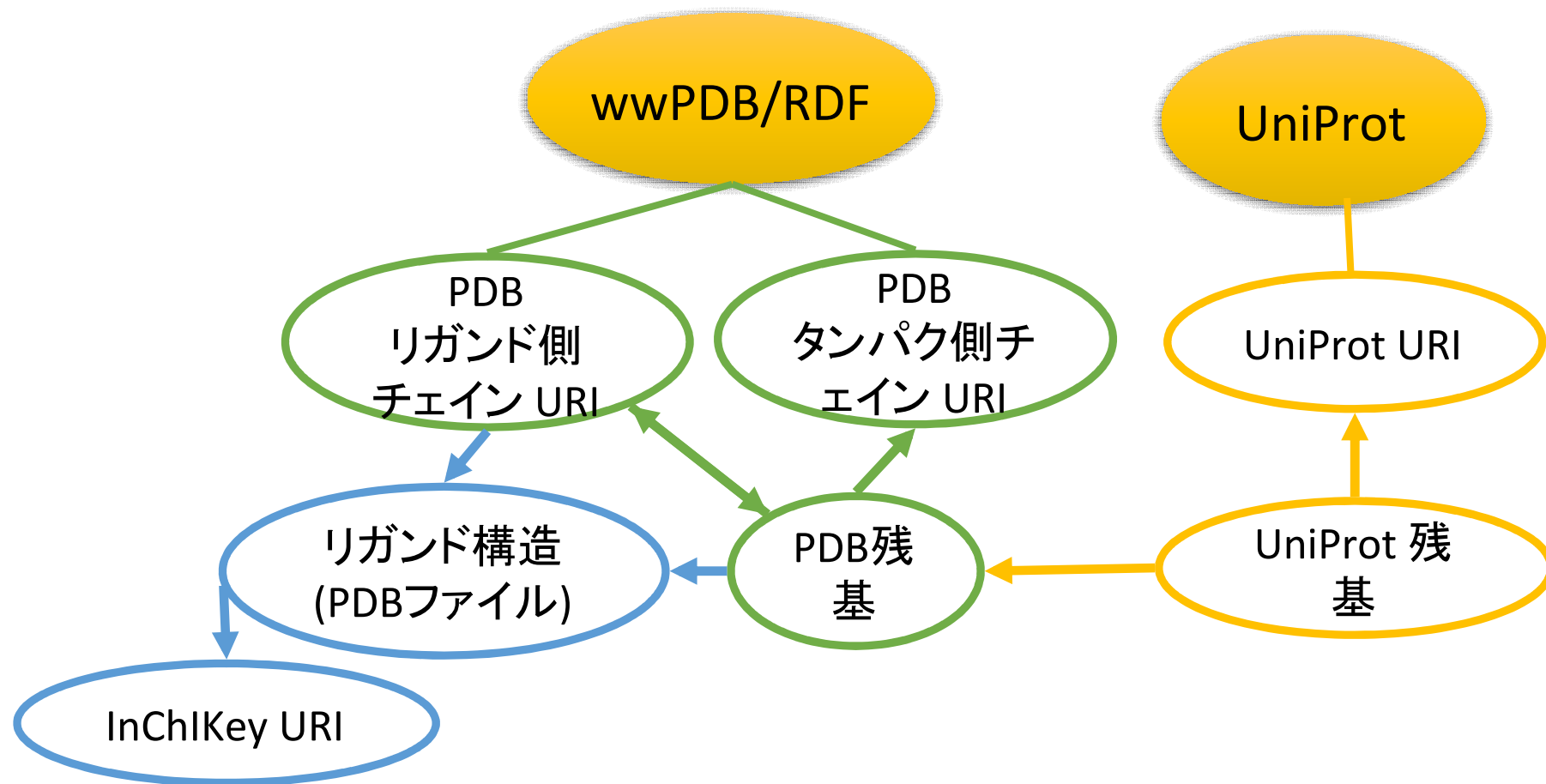
PDBファイル中からグループ化されたHETATMレコードを抜き出し、一つのPDBファイルとして保存。

3



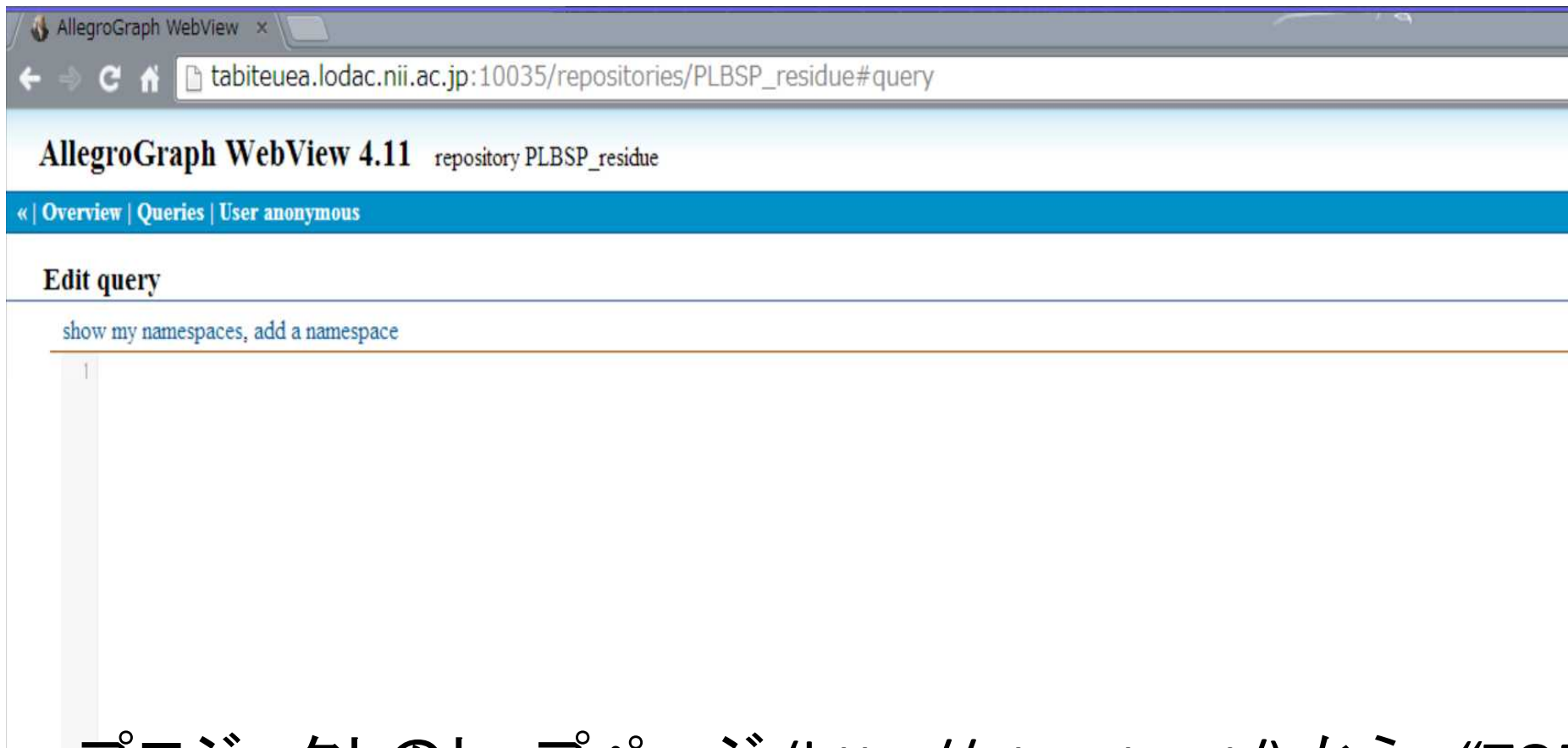
Open Babel を用いて、PDBファイルから InChIKey など構造式情報に変換。

PLBSPResidueのRDFスキーマの概略



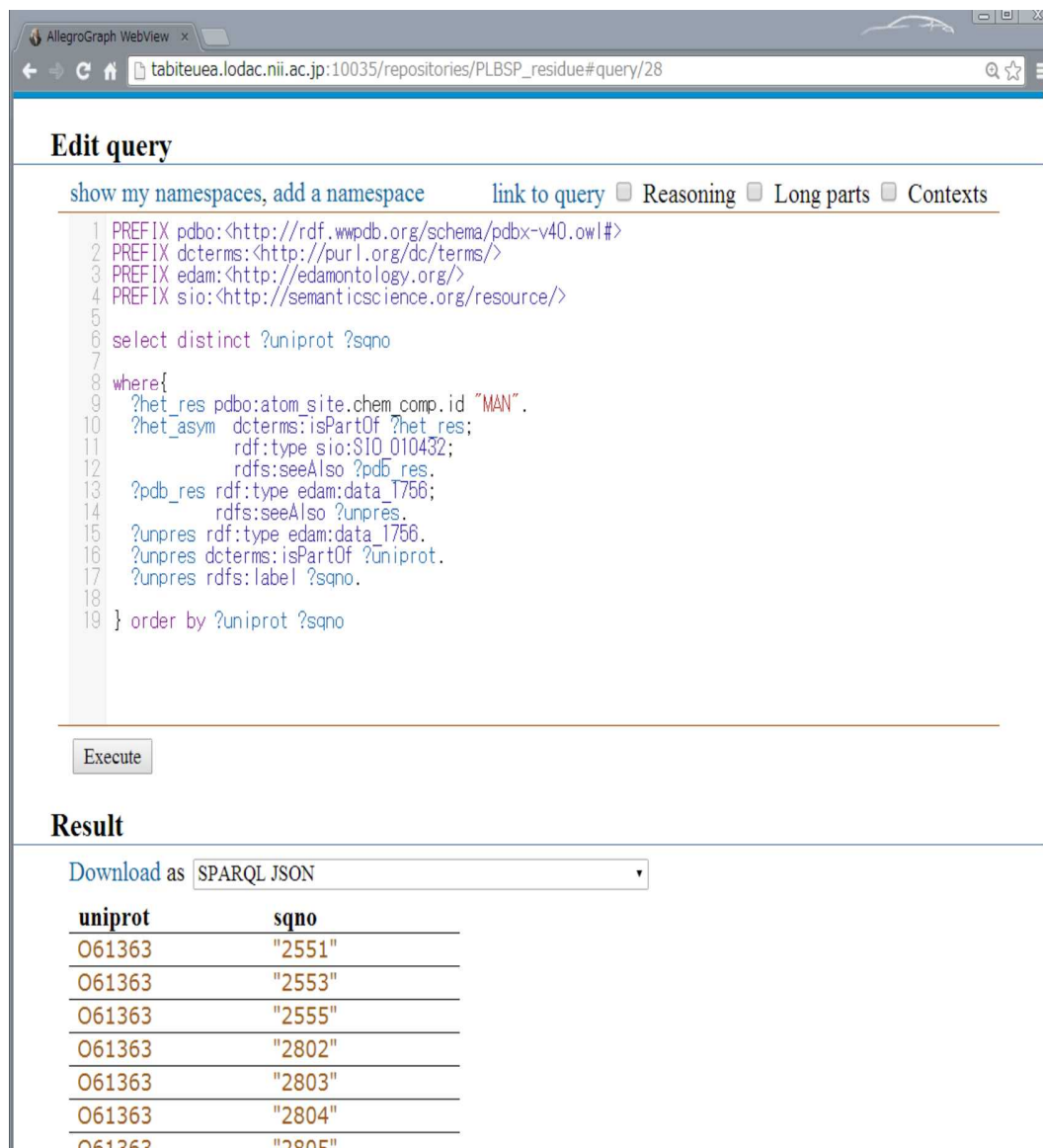
PDBのHETATMコードを指定して、結合タンパク質を問い合わせる形になっている。
結晶構造の解像度やGene Ontology, Family 情報なども検索オプションで指定できる
形にしていきたい。

PLBSPResidueの利用(1)



- プロジェクトのトップページ (<http://utprot.net/>) から “TOP > PLBSP Residue” とリンクをたどることで、今回、構築したデータベースのエンドポイントへアクセスできる。本データベースで動作するクエリは、同Webサイトで順次公開していく。

PLBSPResidueの利用(2)



The screenshot shows the AllegroGraph WebView interface. The browser address bar displays the URL: `tabiteuea.lodac.nii.ac.jp:10035/repositories/PLBSP_residue#query/28`. The page title is "Edit query". Below the title, there are links: "show my namespaces, add a namespace", "link to query", and checkboxes for "Reasoning", "Long parts", and "Contexts". The main area contains a SPARQL query:

```
1 PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>
2 PREFIX dcterms:<http://purl.org/dc/terms/>
3 PREFIX edam:<http://edamontology.org/>
4 PREFIX sio:<http://semanticscience.org/resource/>
5
6 select distinct ?uniprot ?sqno
7
8 where{
9   ?het_res pdbo:atom_site.chem.comp.id "MAN".
10  ?het_asym dcterms:isPartOf ?het_res;
11           rdf:type sio:SIO_010432;
12           rdfs:seeAlso ?pdb_res.
13  ?pdb_res rdf:type edam:data_I756;
14           rdfs:seeAlso ?unpres.
15  ?unpres rdf:type edam:data_I756.
16  ?unpres dcterms:isPartOf ?uniprot.
17  ?unpres rdfs:label ?sqno.
18
19 } order by ?uniprot ?sqno
```

Below the query is an "Execute" button. The "Result" section shows a dropdown menu set to "SPARQL JSON". Below the dropdown is a table with two columns: "uniprot" and "sqno".

uniprot	sqno
061363	"2551"
061363	"2553"
061363	"2555"
061363	"2802"
061363	"2803"
061363	"2804"
061363	"2805"

入力フォームに実行するクエリを入力してExecuteボタンを押すと、画面下の Result 以下にクエリの実行結果が表示される。

PLBSPResidueの利用(3)

AllegroGraph WebView

tabiteuea.lodac.nii.ac.jp:10035/repositories/PLBSP_residue#query/28

Edit query

show my namespaces, add a namespace link to query ☒ Reasoning ☐ Long parts ☐ Contexts

```
1 PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>
2 PREFIX dcterms:<http://purl.org/dc/terms/>
3 PREFIX edam:<http://edamontology.org/>
4 PREFIX sio:<http://semanticscience.org/resource/>
5
6 select distinct ?uniprot ?sqno
7
8 where{
9   ?het_res pdbo:atom_site.chem.comp.id "MAN".
10  ?het_asym dcterms:isPartOf ?het_res;
11    rdf:type sio:SIO_010000;
12    rdfs:seeAlso ?pdb_res;
13  ?pdb_res rdf:type edam:data_1756;
14    rdfs:seeAlso ?unpres;
15  ?unpres rdf:type edam:data_1756;
16  ?unpres dcterms:isPartOf ?uniprot;
17  ?unpres rdfs:label ?sqno.
18
19 } order by ?uniprot ?sqno
```

Execute

Result

Download as: SPARQL JSON

uniprot	
O61363	"2802"
O61363	"2803"
O61363	"2804"
O61363	"2805"

推論が必要な場合は、Reasoning のチェックボックスをクリックする。

Download as の隣のSPARQL JSONをクリックする

Download as の隣の SPARQL JSONをクリックするとどのようなフォーマットで結果をダウンロードするかを選択できる。

フォーマットを選択が完了したら、Download をクリックすることでダウンロードが開始される。

PLBSPResidueの活用例(1)

PLBSPResidueでは次のようなデータ問い合わせが可能である。

(例)

- PDB-CCDのHETATM コードを入力して、そのリガンドと結合するタンパク質のUniProt ACの一覧を取得する。また、そのタンパク質がリガンドと結合する残基の一覧を取得する。
- ChEBI ID を入力して、特定のリガンド分類(ヌクレオチド、金属イオン、脂質など)と結合するタンパク質のUniProt ACの一覧を取得する。
- PDB-CCDのHETATM コードを入力して、そのリガンドと結合するタンパク質の Pfamドメインを集計する。

PLBSPResidueの活用例(2)

PREFIX pdbo:<<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>>

PREFIX dcterms:<<http://purl.org/dc/terms/>>

PREFIX edam:<<http://edamontology.org/>>

PREFIX sio:<<http://semanticscience.org/resource/>>

select distinct ?uniprot ?sqno

where{

?het_res pdbo:atom_site.chem_comp.id "MAN".

?het_asym dcterms:isPartOf ?het_res;

rdf:type sio:SIO_010432;

rdfs:seeAlso ?pdb_res

?pdb_res rdf:type edam:da

rdfs:seeAlso ?unpres.

?unpres rdf:type edam:da

?unpres dcterms:isPartOf ?uniprot.

?unpres rdfs:label ?sqno.

} order by ?uniprot ?sqno

(例1)

『PDB-CCDのHETATMコードを入力して、そのリガンドと結合するタンパク質のUniProt ACの一覧を取得する。また、そのタンパク質がリガンドと結合する残基の一覧を取得する』

対象としているリガンドのHETATM Codeを調べたい場合は、下記を参照。

PDBChem: <http://www.ebi.ac.uk/pdbe-srv/pdbechem/>

LigandExpo: <http://ligand-expo.rcsb.org/>

ここを、対象とするリガンドのHETATM Code に書き換える。

PLBSPResidueの活用例(3)

PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>

PREFIX dcterms:<http://purl.org/dc/terms/>

PREFIX edam:<http://edamontology.org/>

PREFIX sio:<http://semanticscience.org/resource/>

PREFIX up:<http://purl.uniprot.org/core/>

prefix obo:<http://purl.obolibrary.org/obo/>

```
select distinct ?uniprot ?cc {  
  ?het_asym dcterms:isPartOf ?het_res;  
    rdf:type sio:SIO_010432;  
    rdfs:seeAlso ?pdb_res;  
    rdfs:seeAlso ?chem_comp.  
  
  ?chem_comp pdbo:link_to_chem_comp ?cc.  
  ?cc rdf:type obo:CHEBI_72544.
```

```
?pdb_res rdf:type edam:data_17;  
  rdfs:seeAlso ?unpres.
```

```
?unpres rdf:type edam:data_17;  
  dcterms:isPartOf ?uniprot.
```

(例2)

『ChEBI ID を入力して、特定のリガンド分類(ヌクレオチド、金属イオン、脂質など)と結合するタンパク質のUniProtIDの一覧を取得する(推論が必要なのでReasoningをチェックしてから実行する)』

対象としているリガンド分類のChEBI ID を調べたい場合は、下記のサイトを参照。

ChEBI: <https://www.ebi.ac.uk/chebi/>

ここを、対象とするリガンド分類のChEBI ID に書き換えてください。

PLBSPResidueの活用例(4)

PREFIX pdbo:<http://rdf.wwpdb.org/schema/pdbx-v40.owl#>

PREFIX dcterms:<http://purl.org/dc/terms/>

PREFIX edam:<http://edamontology.org/>

PREFIX sio:<http://semanticscience.org/resource/>

PREFIX up:<http://purl.uniprot.org/core/>

prefix obo:<http://purl.obolibrary.org/obo/>

```
select ?pfam (count( ?uniprot ) as ?cnt ) {  
select distinct ?uniprot ?asym ?pfam {  
  ?auth_asym dcterms:identifier ?sifts.  
  ?sifts edam:has_identifier ?pfam.  
  ?pfam rdf:type edam:data_1138.  
  ?auth_asym pdbo:struct_asym.id ?asym.
```

```
  ?het_res pdbo:atom_site.chem_comp.id "MAN".  
  ?het_asym dcterms:isPartOf ?het_res;  
    rdf:type sio:SIO_010432;  
    rdfs:seeAlso ?pdb_res.
```

```
  ?pdb_res rdf:type edam:data_1756;  
    rdfs:seeAlso ?unpres;  
    dcterms:isPartOf ?asym.
```

```
  ?unpres rdf:type edam:data_1756;  
    dcterms:isPartOf ?uniprot.
```

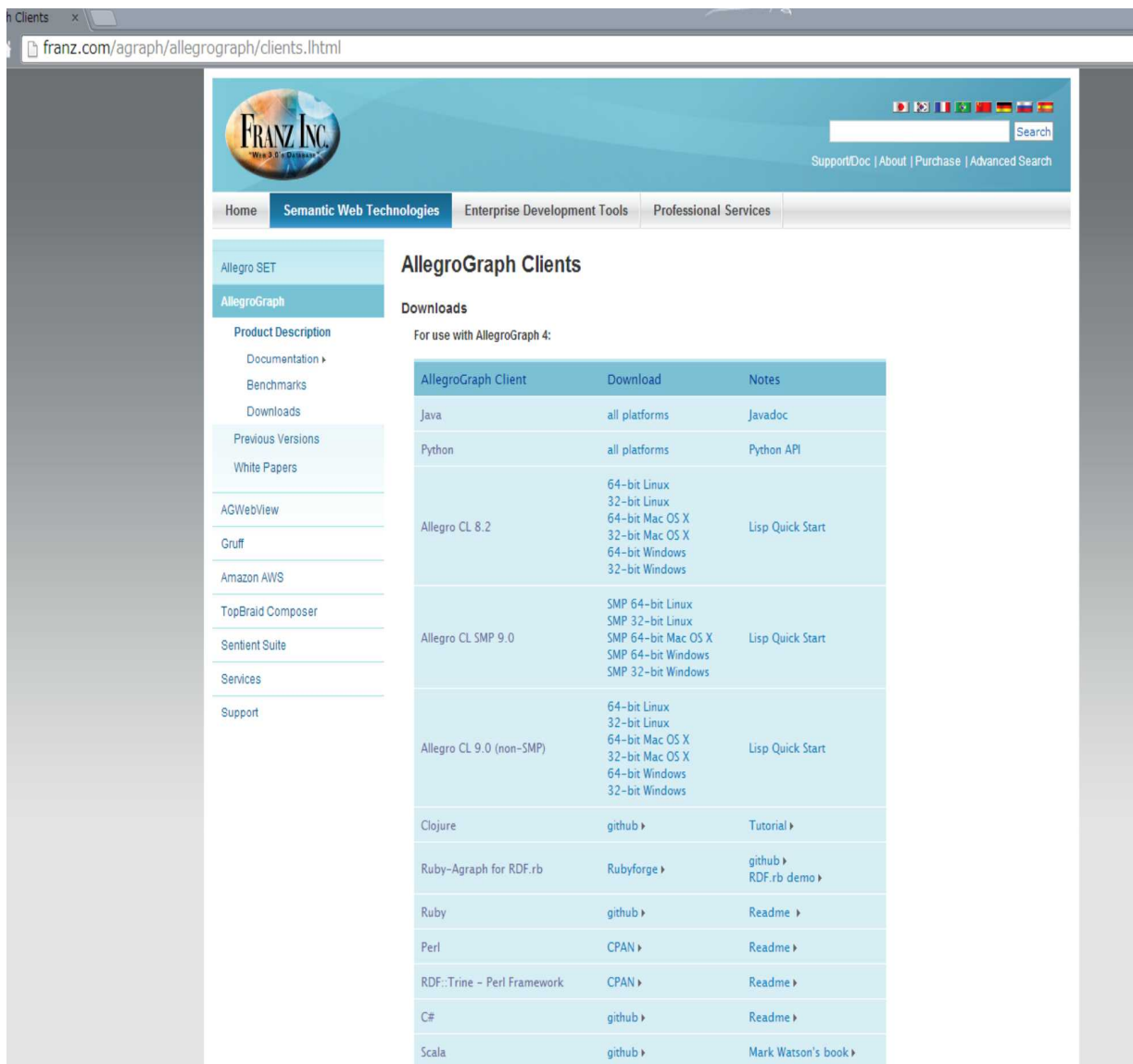
```
}} group by ?pfam order by ?cnt
```

(例2)

『PDB-CCDのHETATM コードを入力して、そのリガンドと結合するタンパク質の Pfamドメインを集計する』

ここを、対象とするリガンド分類のChEBI ID に書き換える。

PLBSPResidueの活用例(5)



The screenshot shows the Franz AllegroGraph Clients page. The browser address bar displays `franz.com/agraph/allegrograph/clients.html`. The page features a navigation menu with links to Home, Semantic Web Technologies, Enterprise Development Tools, and Professional Services. A sidebar on the left lists various resources including Allegro SET, AllegroGraph, Product Description, Documentation, Benchmarks, Downloads, Previous Versions, White Papers, AGWebView, Gruff, Amazon AWS, TopBraid Composer, Sentient Suite, Services, and Support. The main content area is titled "AllegroGraph Clients" and includes a "Downloads" section. Below this, a table lists available clients and their download links.

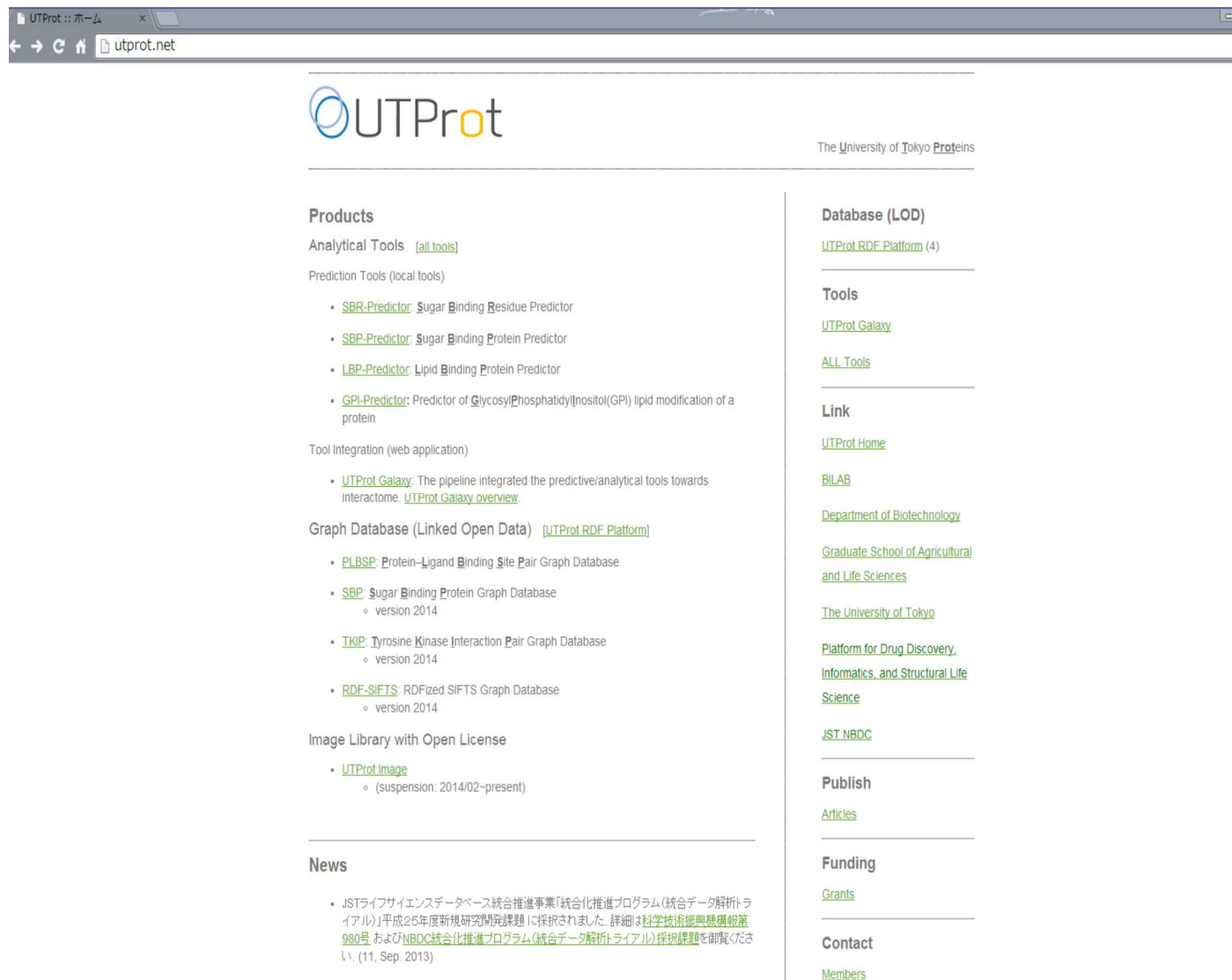
AllegroGraph Client	Download	Notes
Java	all platforms	Javadoc
Python	all platforms	Python API
Allegro CL 8.2	64-bit Linux 32-bit Linux 64-bit Mac OS X 32-bit Mac OS X 64-bit Windows 32-bit Windows	Lisp Quick Start
Allegro CL SMP 9.0	SMP 64-bit Linux SMP 32-bit Linux SMP 64-bit Mac OS X SMP 64-bit Windows SMP 32-bit Windows	Lisp Quick Start
Allegro CL 9.0 (non-SMP)	64-bit Linux 32-bit Linux 64-bit Mac OS X 32-bit Mac OS X 64-bit Windows 32-bit Windows	Lisp Quick Start
Clojure	github ▶	Tutorial ▶
Ruby-Agraph for RDF.rb	Rubyforge ▶	github ▶ RDF.rb demo ▶
Ruby	github ▶	Readme ▶
Perl	CPAN ▶	Readme ▶
RDF::Trine - Perl Framework	CPAN ▶	Readme ▶
C#	github ▶	Readme ▶
Scala	github ▶	Mark Watson's book ▶

本データベースはグラフデータベースシステム AllegroGraphを用いている。データ取得は 開発元である Franz社が無料で提供する専用のクライアントライブラリを用いることで、JavaやPythonといったプログラム言語から問い合わせることも可能である。

多くのデータを機械的に処理する場合はこちらが有効である。

<http://franz.com/agraph/allegrograph/clients.html>

UTProt: プロジェクトトップページ



- 追加情報は、適宜 <http://utprot.net/> へ記述していく。