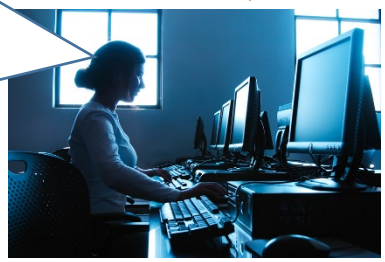
# 機械学習を用いた タンパク質-リガンド結合部位 予測ツールの自動生成 パイプラインの開発

東京大学大学院　農学生命科学研究科
応用生命工学専攻　博士課程
番野雅城

# これまでの（教師付き）機械学習を用いた予測ツールの研究開発の流れ

- どのデータベースの選択すべきか
- データベース間の相互参照の解決
- どういった観点でデータを集めるべきか

公共データベース
（PDB,UniProt など）



バイオインフォマティシャン

システム開発

特徴ベクトルの設計

機械学習アルゴリズムの選定

ユーザーインターフェースの設計

ドメイン知識の獲得

ウェット研究者との議論

論文・文献調査

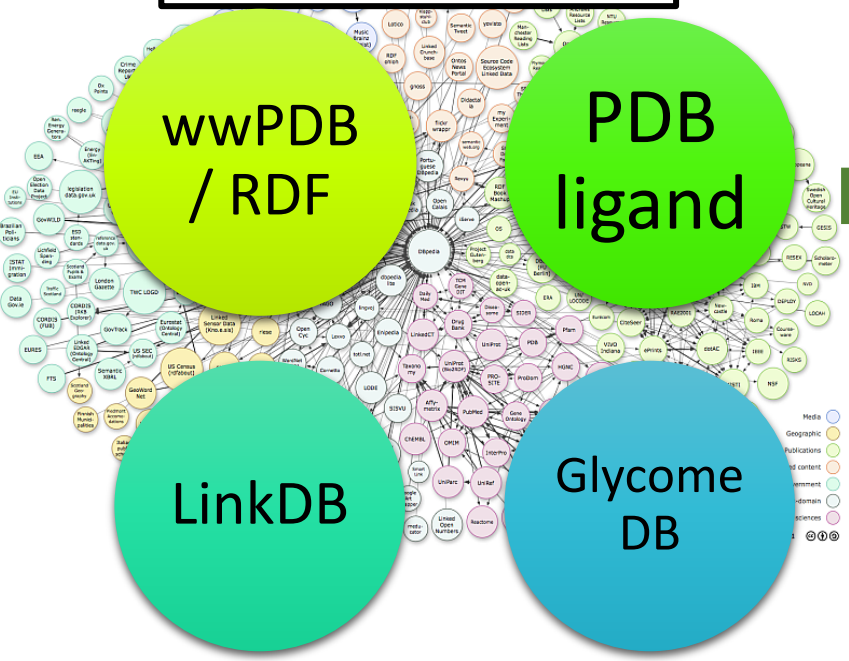
- 技術的な部分の工程は共通部分が多い。
- 何を予測したいかという要望は研究者によって細かな相違があり、全ての要望をすくい上げるのは難しい。

# 目標

本研究では、研究者自身の興味の対象（リガンド種、機能など）に合わせて結合部位予測ツールを生成するパイプラインを開発する。

これにより、既存のツールでは対応できない個別の問題にも対応できる予測ツールの生成が可能となる。

**複数のデータベースを統合検索**




特徴抽出
（残基ごとに特徴ベクトルに変換）

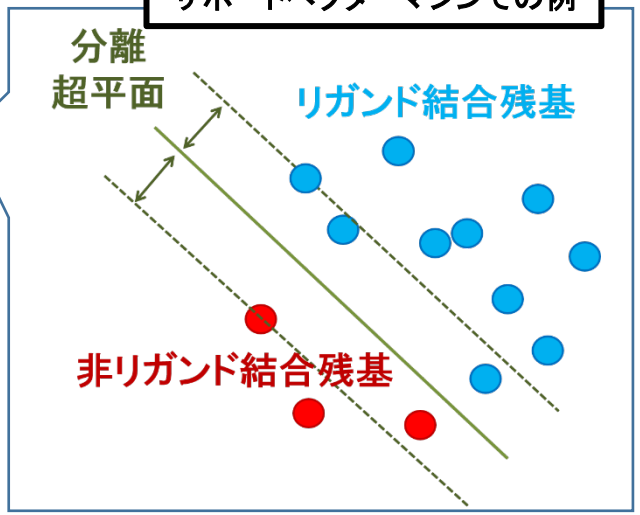
学習　予測

機械学習

リガンド結合残基を予測

...SERDFLALDLGGT...

サポートベクターマシンでの例




リガンド結合部位データベースからのデータセット生成のワークフローと、機械学習のワークフローをパイプライン化。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。

# 当初計画

10月

- 予測ツール生成ワークフローの開発
- 予測ツール管理バックエンド開発

11月

- データセット生成ワークフローの開発

# 9 ～ 11月の活動報告

- データセット生成ワークフローの構築
  - リガンド結合残基データベースの開発
  - EBI-SIFTSの残基番号情報をRDF化
  - リガンド間共有結合情報のRDF化（現在開発中）
  - データセット生成ワークフロー試作版の作成
- 予測モデル生成ワークフロー試作版の作成
  - データセット生成ワークフローと予測ツール生成ワークフローの結合試験
- 予測ツール登録リポジトリを公開

# 基盤データベースの構築

## 既存の問題点

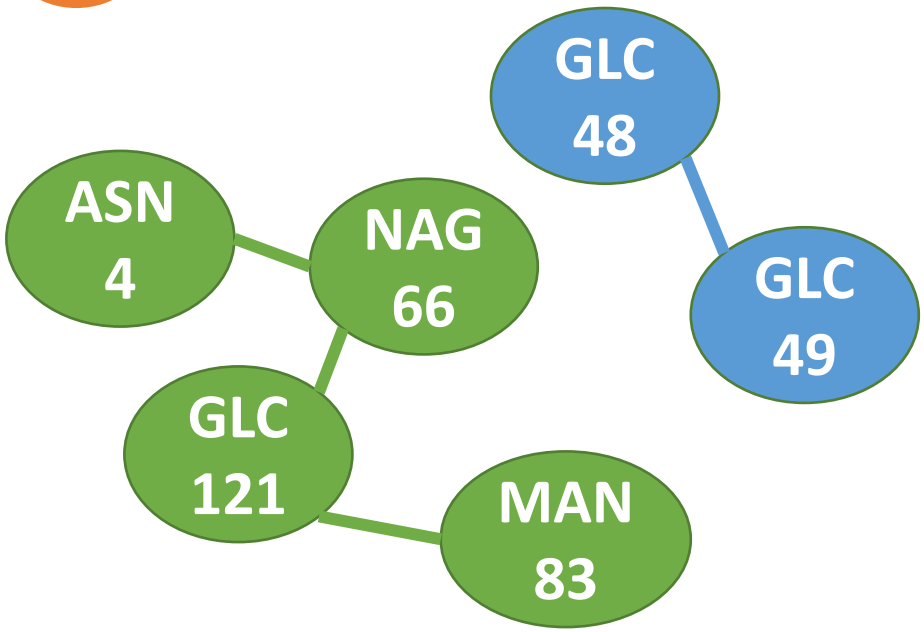
- PDB中の全タンパク質リガンド間原子間距離を網羅したデータベース(PLBSP)にデータ問い合わせを行ったが、データ取得が非常に遅かった。
- PDB中の残基番号とFastaファイル上のアミノ酸位置を対応付けが難しい場合がある。
- PDBファイルだけでは、リガンドであるか残基修飾であるか判断が難しい。

## 解決策

- 5Å以内に存在するPDBファイル中の残基とリガンドの関係を記述した軽量なデータベースの構築した。
- UniProtとPDBの残基の対応関係を人手でキュレーションした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）。

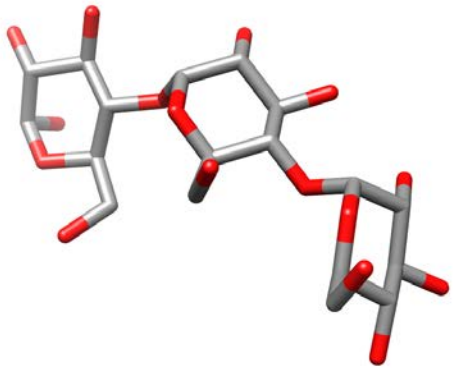
# リガンド間共有結合情報のRDF化

1



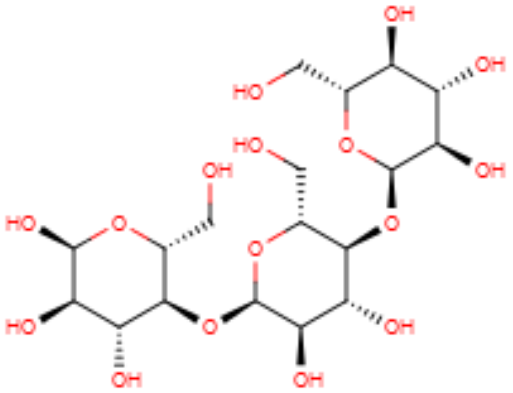

PDBのCONNECTレコードから残基をノード、共有結合関係をエッジとしたグラフデータを生成。経路発見アルゴリズムを用いて、一つの分子を形成するHETATMレコードの組み合わせをグループ化する。

2



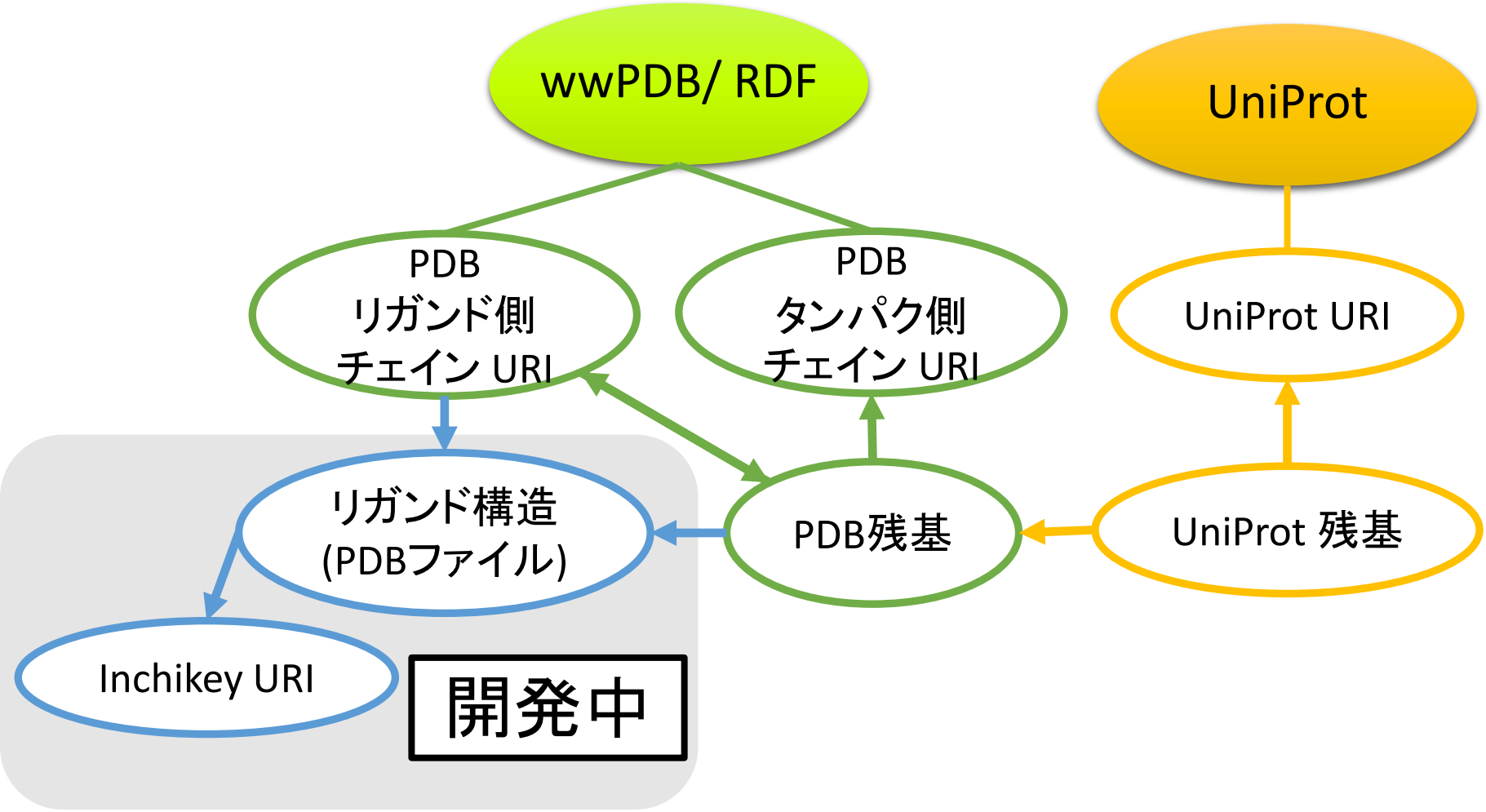
PDBファイル中からグループ化されたHETATMレコードを抜き出し、一つのPDBファイルとして保存。

3




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

# 基盤データベースのデータスキーマ




現状では、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。

# データセット生成パイプラインの流れ

INPUT: 対象リガンドのHETATM コード

**リガンド結合タンパク質の問い合わせ**

- PLBSP-Residue
- UniProt URI を取得

**タンパク質のアミノ酸配列の取得**

- UniProt のSPARQL Endpoint
- アミノ酸配列を取得

**配列冗長性の除去**

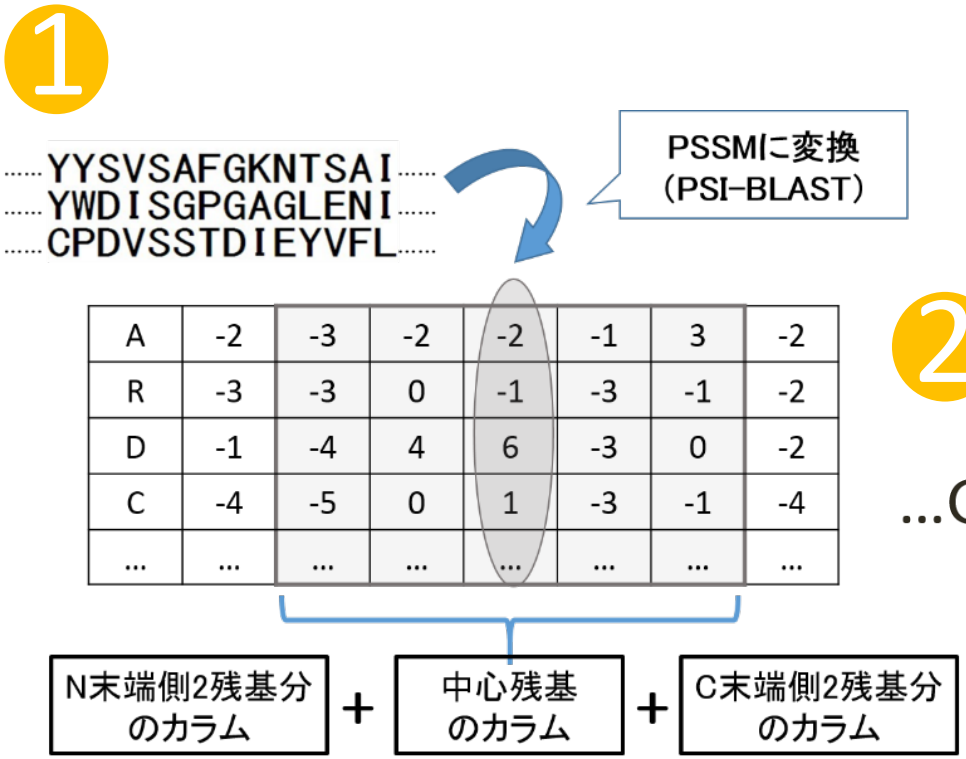
- CD-HIT
- 類似性30%、カバー率50%を条件に配列冗長性除去

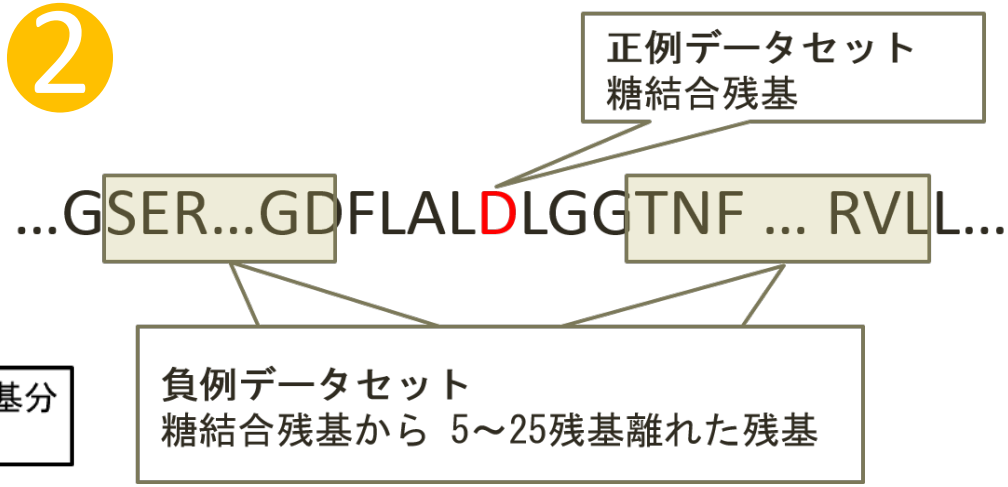
**リガンド結合残基の取得**

- PLBSP-Residue
- 指定したリガンドのFasta上での残基番号を取得

OUTPUT:
対象リガンド結合タンパク質の非冗長なアミノ酸配列
リガンド結合残基の残基番号

# 予測ツールの生成パイプラインの流れ

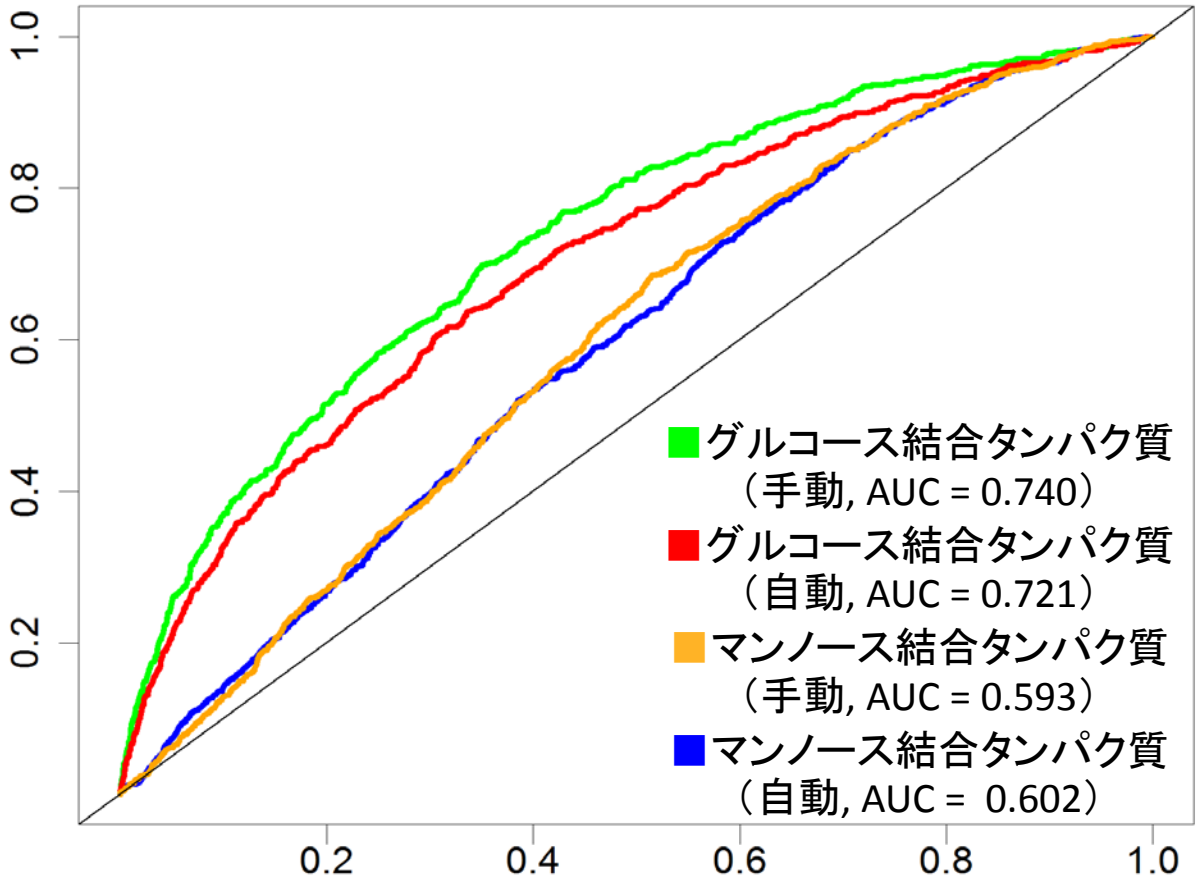



PSI-BLASTを用いてアミノ酸配列をPSSMに変換。中心からｗ残基分のカラムを特徴ベクトルとして使用する。

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結合試験




遺伝子 : c,g,w ∈ [-14,14]

評価関数:

Window = 2 ×(|w| + 1)

$C = 2^c$

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

( 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

動作テストも兼ねて、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で実行することで結合部位予測ツール生成を行った。

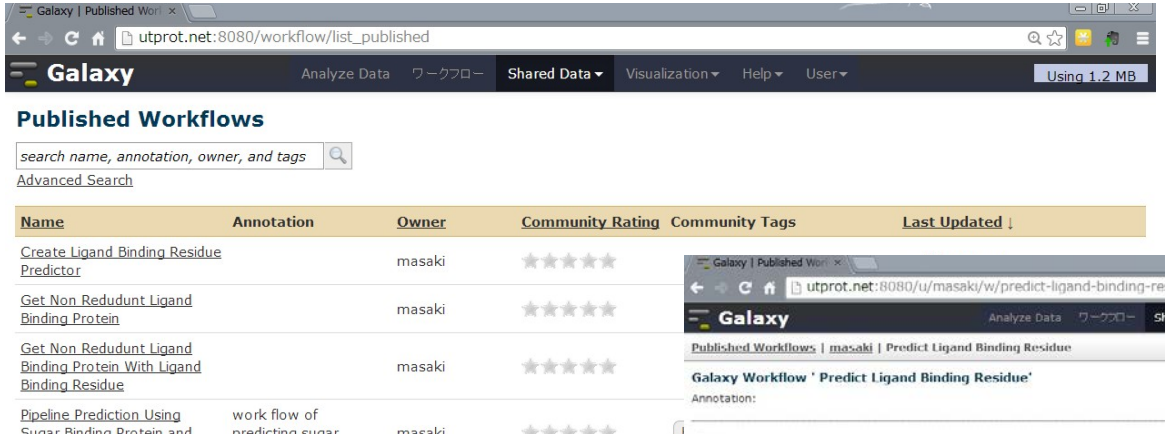
# UTProt Galaxy

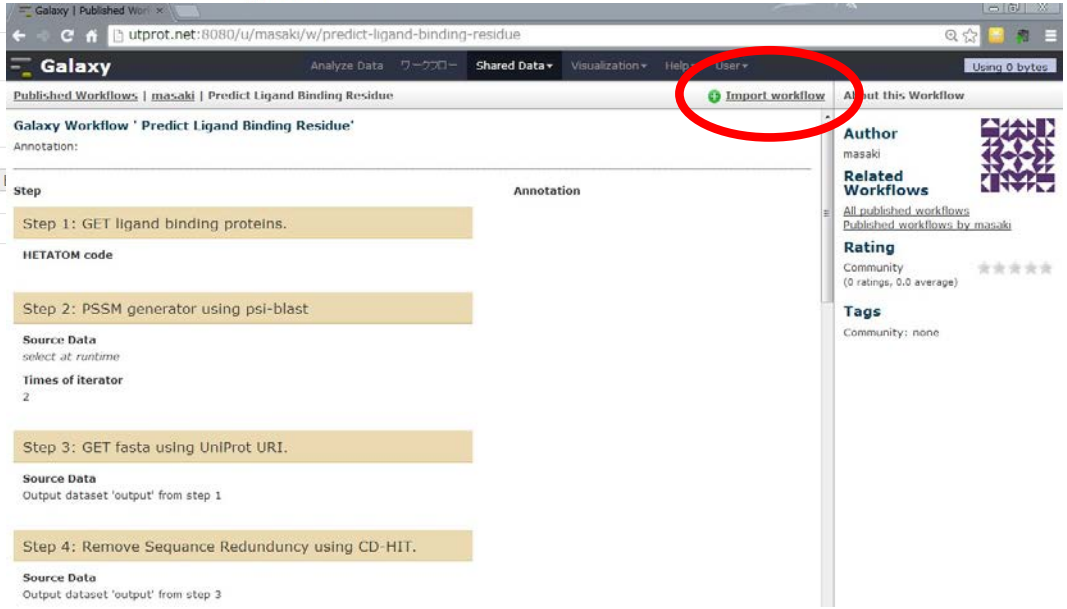
上段メニューのUsers を押して
Registration でユーザー登録


UTPROT Galaxy にユーザー登録することで、
http://utprot.net:8080/workflow/list_published
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）

http://utprot.net:8080/workflow/list_published




http://utprot.net:8080/u/masaki/w/predict-ligand-binding-residue

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）



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）

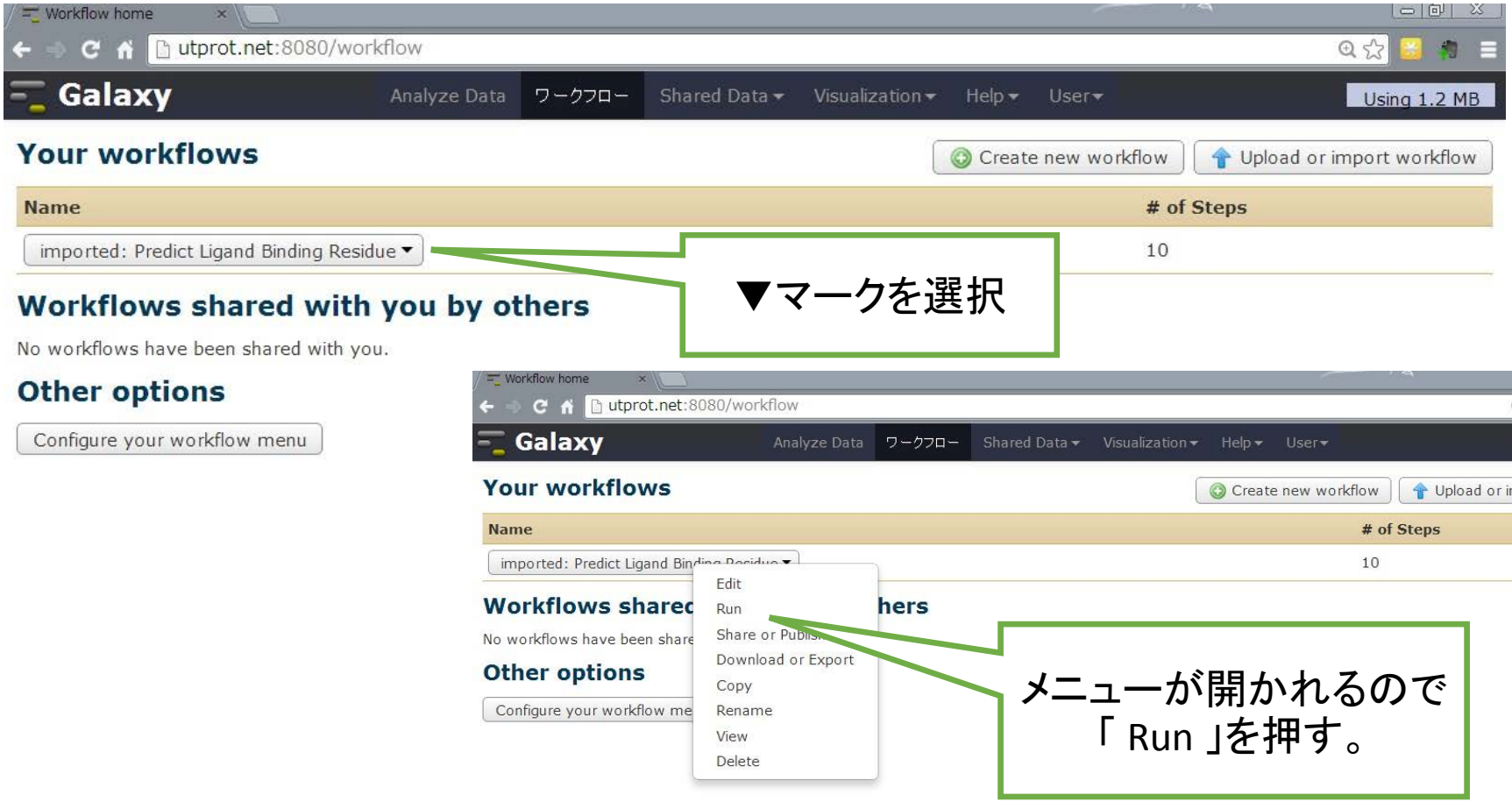

ファイルアップロードかコピー＆ペーストのいずれかの方法でFastaファイルがアップロードができる。

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。

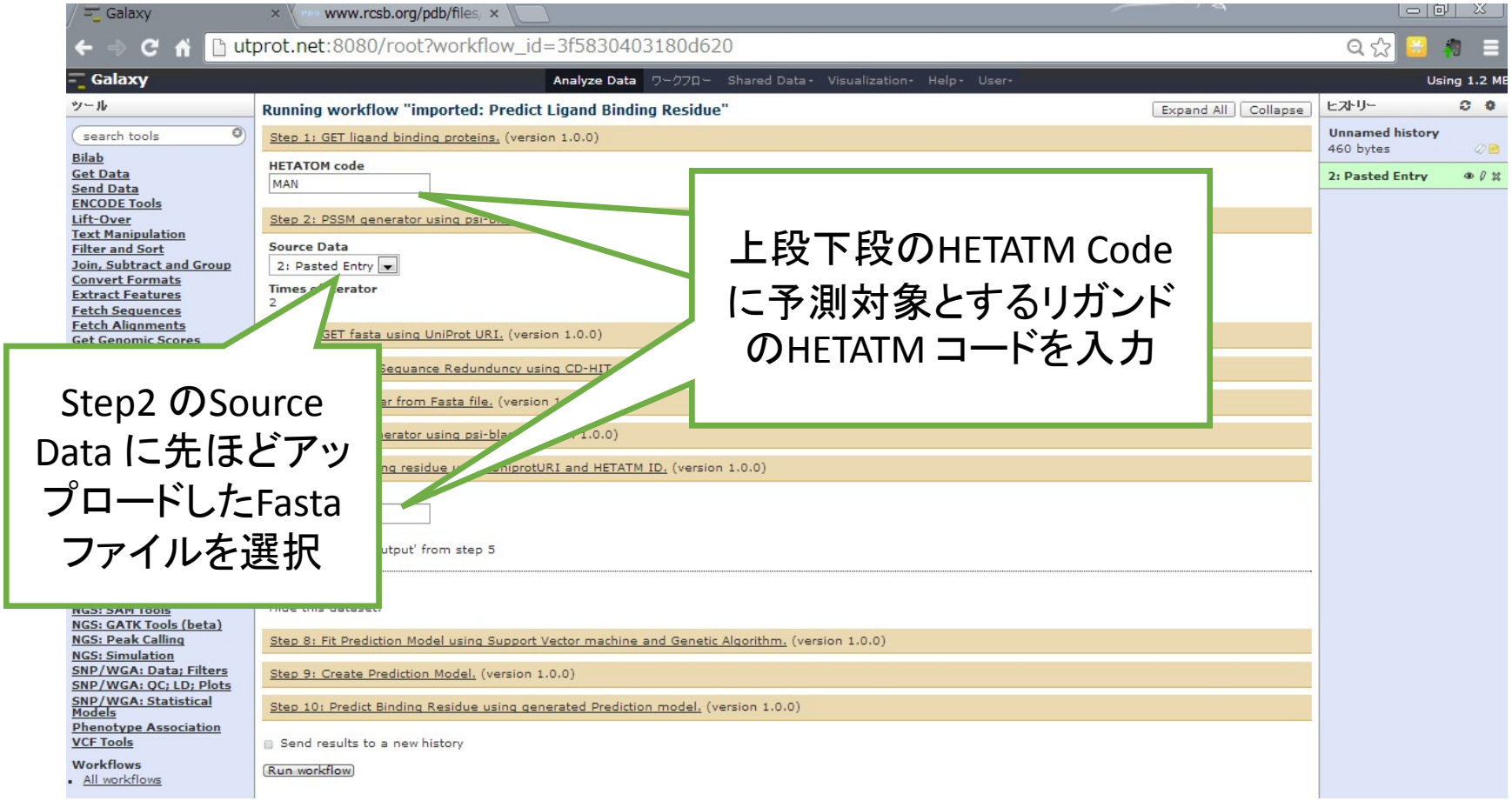

Galaxy 上に保存されたデータはこちらに表示される。

# UTProt Galaxy の利用（4）




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。

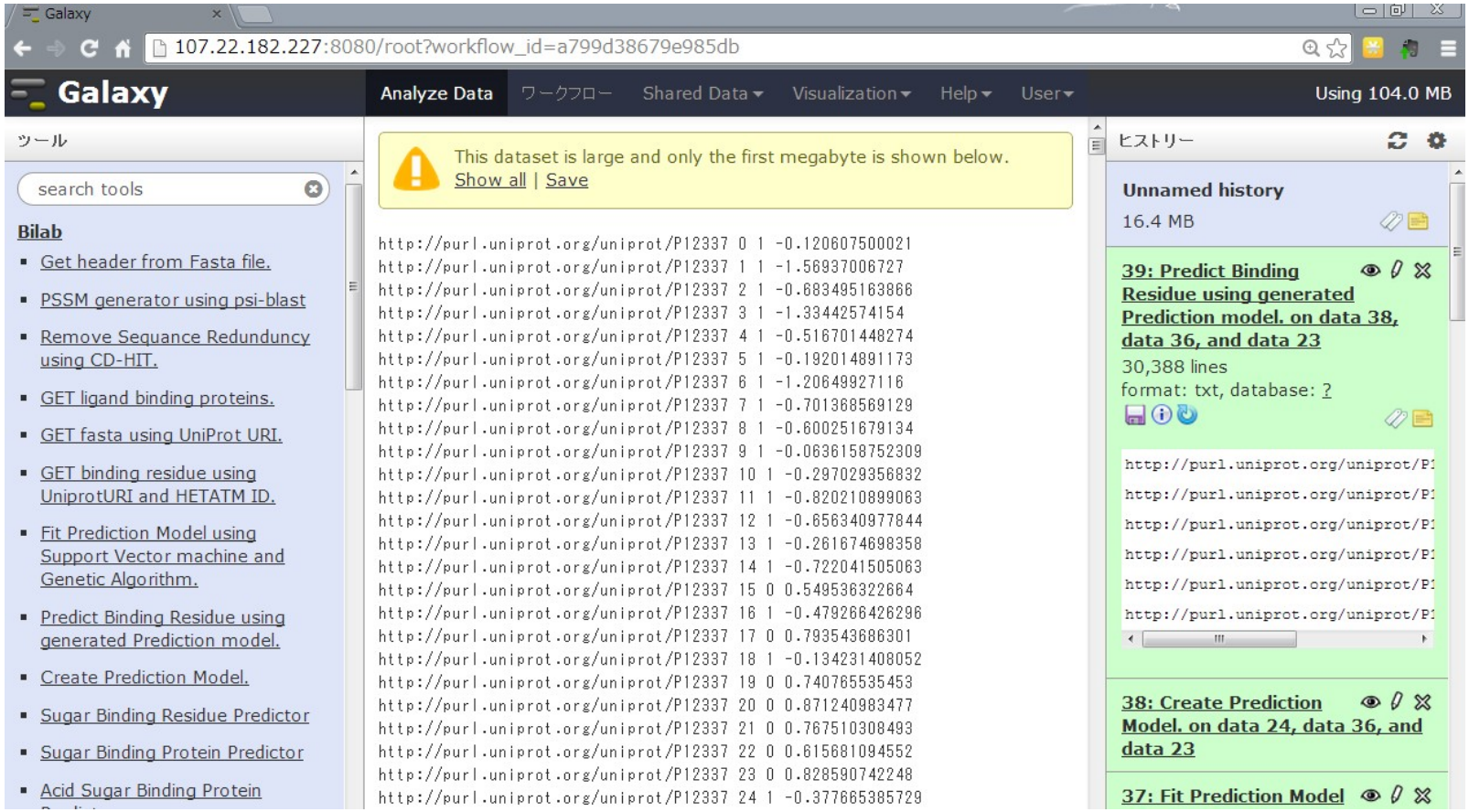
# UTProt Galaxy の利用（5）




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ます。**

# UTProt Galaxy の利用（6）




現状では、タンパク質名、残基番号、結合部位予測（1 ならば結合部位と予測）、decision value (結合残基らしさのスコア) のテキストの形で結果が返ってくる。

# 予測ツール登録リポジトリ


生成された予測ツールはCKANを使って構築した予測ツールデータポータルに登録することで、他ユーザーと共有してもらうことを検討している。
ユーザーは自分の予測ツールにコメントやタグをつけることで、第三者からでも検索が容易となる。

# 今後の予定

- 基盤データベースの整備を進める（リガンドの共有結合情報、UniProtとPDBの残基マッピングを進める）
- SVMのパラメーター探索方法について、複数のデータセットを用いて、より良い最適化手法を探す
- データや予測結果の可視化方法を考える
- 実際の応用事例を蓄積する