

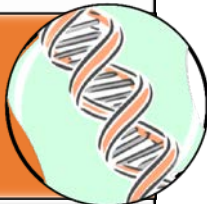
機械学習を用いた タンパク質-リガンド結合部位 予測ツールの自動生成 パイプラインの開発

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

背景： タンパク質-リガンド相互作用

- 分子認識
- 酵素
- 機能制御
- 細胞の支持体

生体内での機能



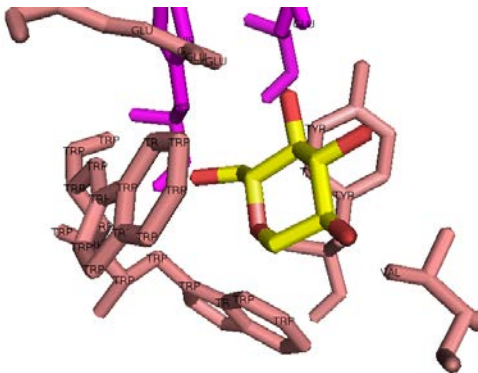
- 農学
 - 機能性食品の開発
 - 有用酵素の設計
 - 安全な農薬の設計
- 環境科学
 - 環境ホルモンの解析など
- 創薬
 - 薬剤ターゲット探索

応用分野



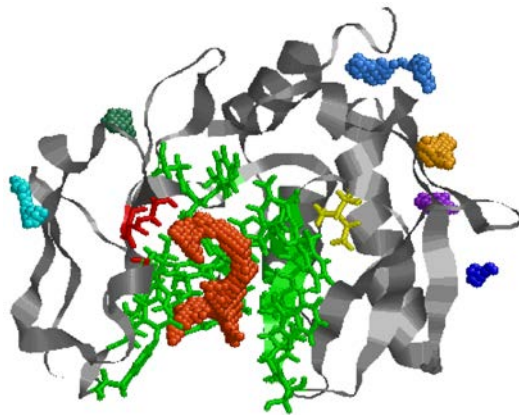
リガンド結合部位を決定・予測する手法

実験的な手法



- X線結晶構造解析
- 核磁気共鳴法
- 生化学的手法
- etc ...

計算機を用いた手法



- ドッキングシミュレーション
- 機械学習的手法

- 実験的手法は、一般に複雑な手順・高額な装置・研究者の熟練等が必要である。
- ドッキングシミュレーション的な手法は、高い精度を持つが単体構造を決定する必要があるなど制約がある。

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



バイオインフォマティクス

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

システム開発

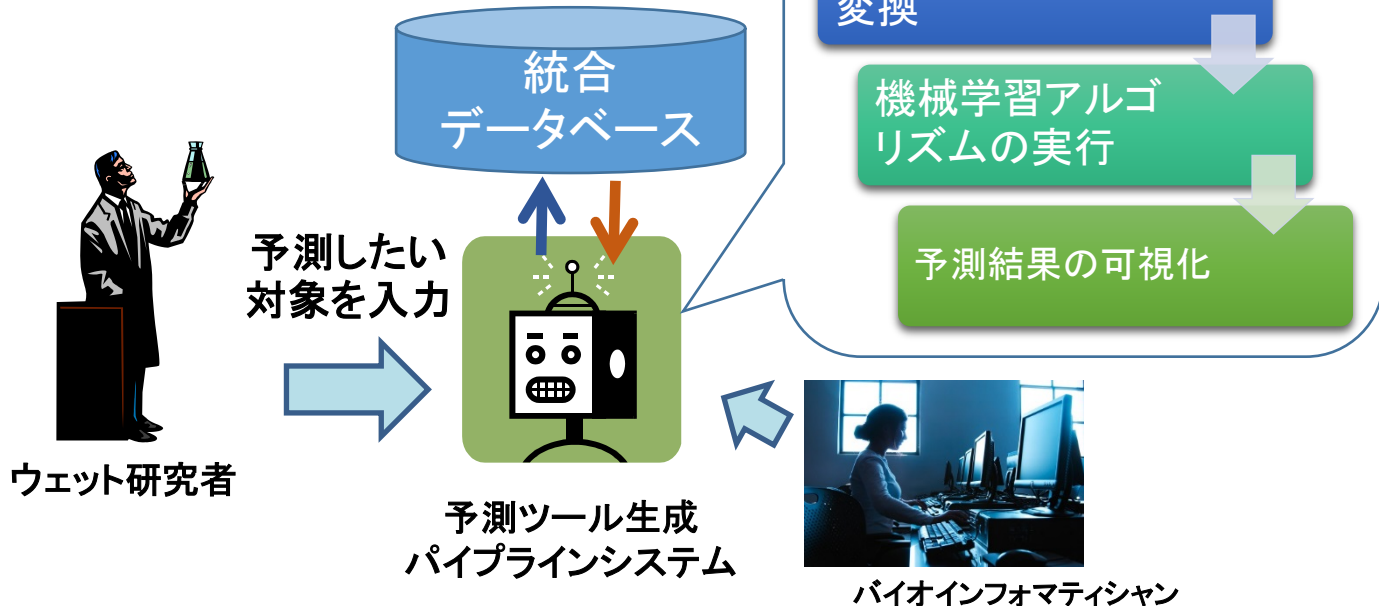
特徴ベクトルの
設計

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

- 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。

本研究の狙い



統合データベースの情報をもとに、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対し、高い精度で予測できる予測ツールの生成が可能となる。

手順1: ユーザーは予測ツール生成システムに 目的のリグンドを入力

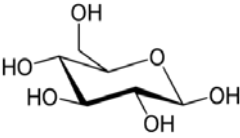
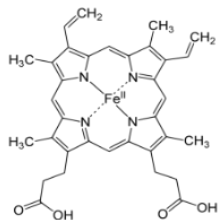
INPUT: 対象リグンド

低分子化合物

金属イオン

Zn^{2+}

糖類



ユーザー

汎用的なパイプライン (Galaxy)

データセット
生成

- データベースから学習に用いるデータセットの抽出

特徴抽出

- 位置特異的スコア行列 (PSSM) 計算

機械学習

- サポートベクターマシンによる学習
- 学習パラメーターの最適化

結合部位予測ツール生成パイプライン

手順2:新規予測ツールに対しユーザーは アミノ酸配列を入力

INPUT: アミノ酸配列
(FASTA)

...SERDFLALALGGT...

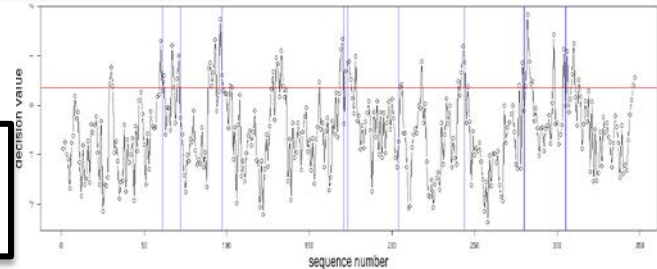
ユーザーが持つ機能未知の
リガンド結合タンパク質のアミノ酸配列



OUTPUT:
リガンド結合残基

...SER**D**FLALDL**G**GT...

予測されたリガンド結合残基



各残基がリガンド結合残基
かを判定する
予測スコアを出力

研究開発内容

データセット生成に特化した二次データベース (PLBSPResidue) の開発

- PDB中のリガンド結合残基情報のRDF化
- リガンド間共有結合情報のRDF化
- PDBとUniProtの間のタンパク質鎖レベルの相互参照情報のRDF化

データセット生成パイプラインの開発

- データベースからデータセットを取得
- 配列冗長性の除去
- PSSMへ変換

予測ツール生成パイプラインの開発

- 特徴ベクトルへ変換
- SVMによる学習の実行
- 遺伝的アルゴリズムによるパラメーター最適化の自動化

システム公開

- UTProt Galaxy の立ち上げ、リガンド結合部位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である UTProt CKAN の立ち上げ

PLBSPResidueデータベースの構築

既存の問題点

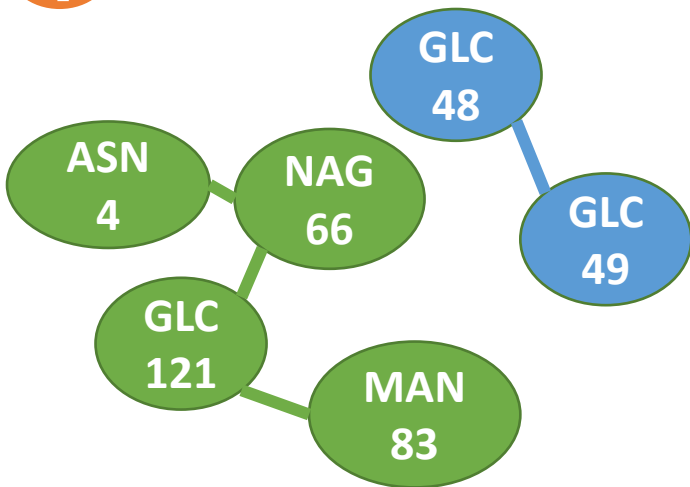
- PLBSPは全原子間距離を網羅しているためレコード数が膨大であり、データ取得が非常に遅い。
- PDB中の残基番号とFASTAファイル上のアミノ酸残基位置の対応付けが難しい場合がある。
- PDBファイルだけでは、リガンドであるか残基修飾であるか判断が難しい。

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。

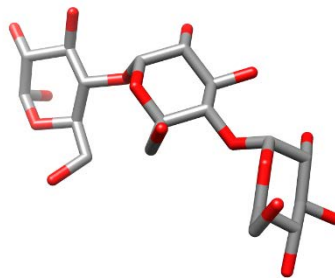
リガンド間共有結合情報のRDF化

1



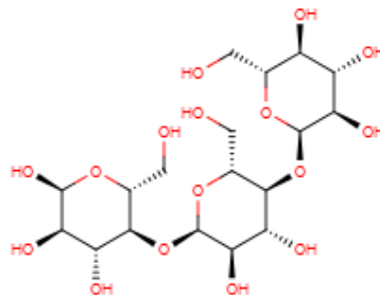
PDBのCONNECTレコードから残基をノード、共有結合関係をエッジとしたグラフデータを生成。経路発見アルゴリズムを用いて、一つの分子を形成するHETATMレコードの組み合わせをグループ化する。

2



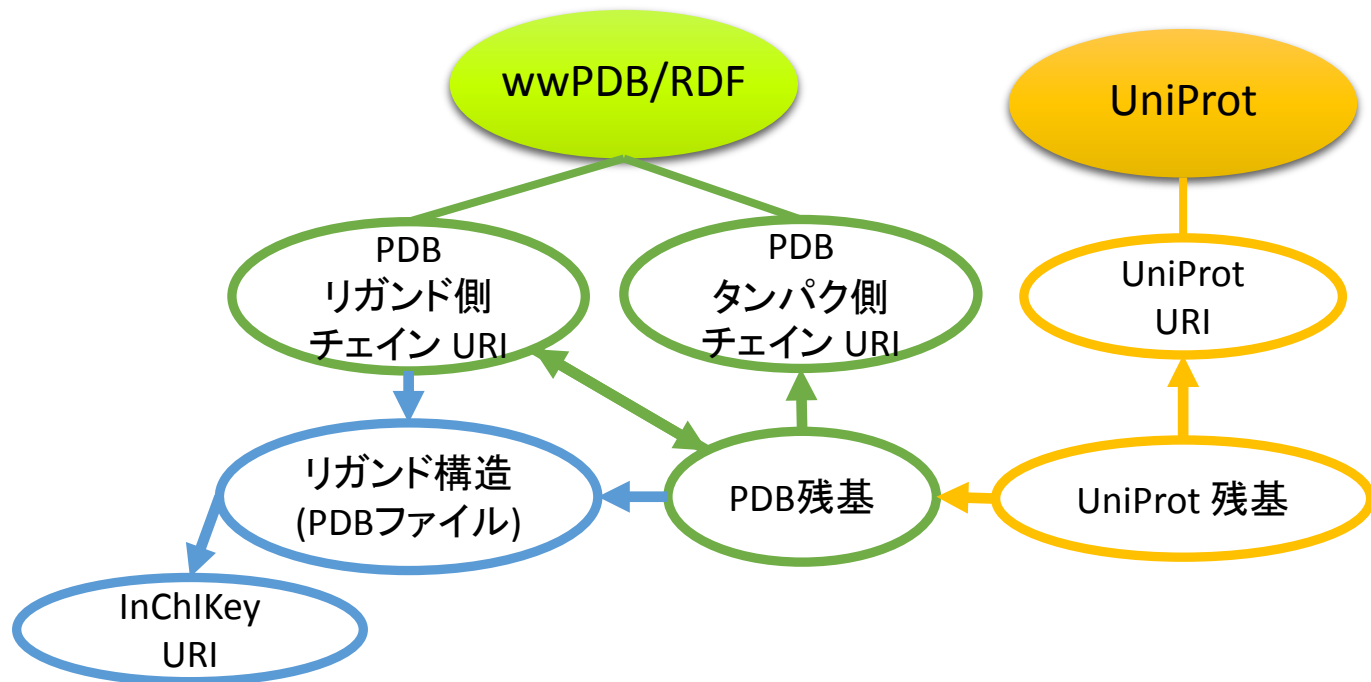
PDBファイル中からグループ化されたHETATMレコードを抜き出し、一つのPDBファイルとして保存。

3



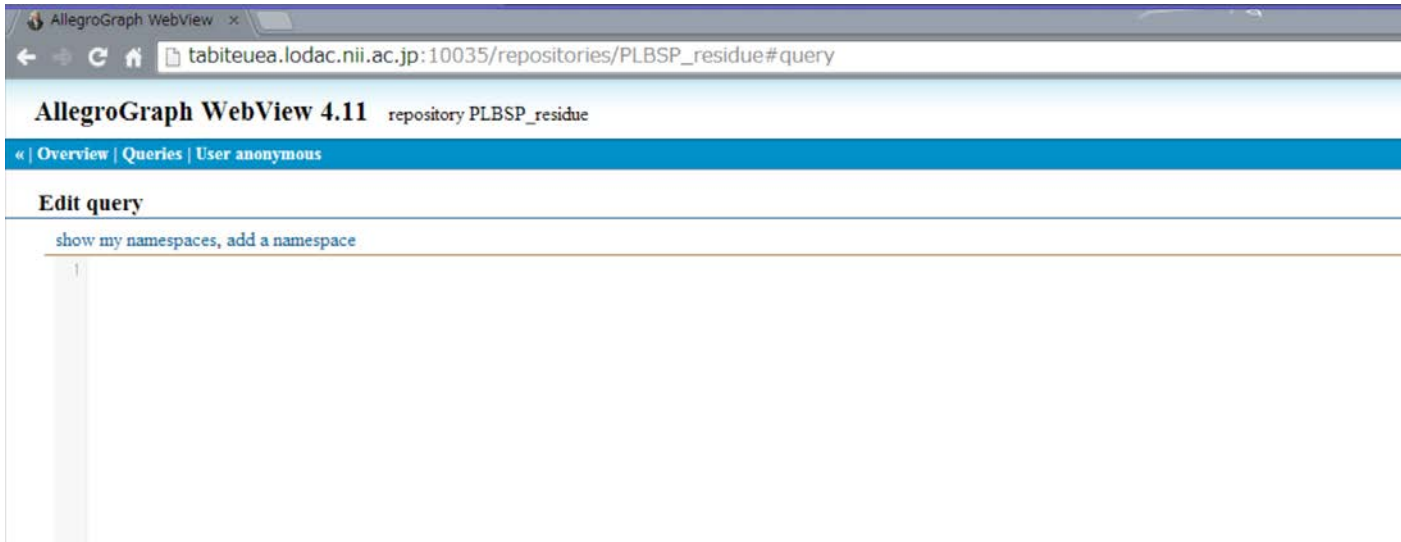
Open Babel を用いて、PDBファイルから InChIKey など構造式情報に変換。

PLBSPResidueのRDFスキーマの概略



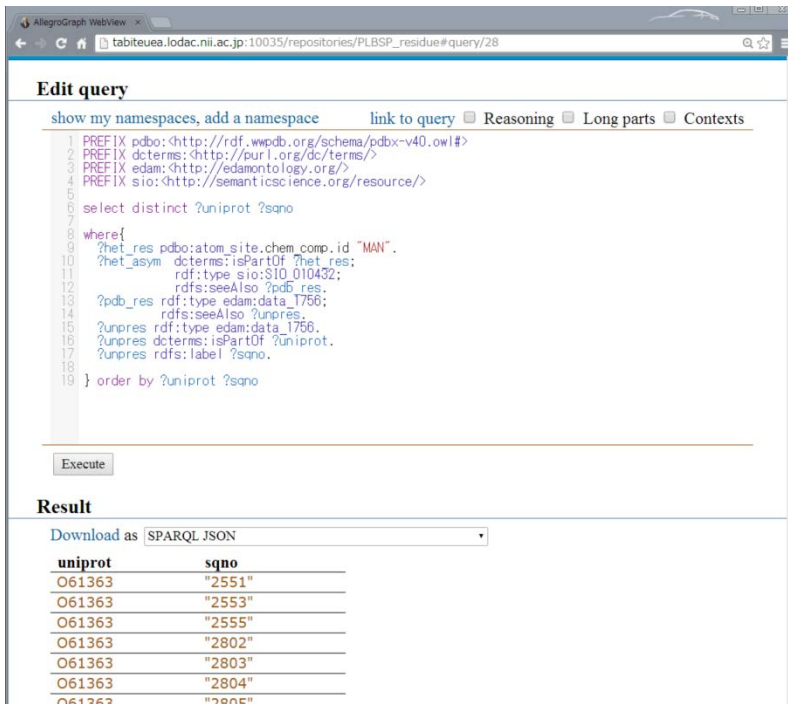
PDBのHETATMコードを指定して、結合タンパク質を問い合わせる形になっている。
結晶構造の解像度やGene Ontology, Family 情報なども検索オプションで指定できる形に
していきたい。

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: `tabiteues.lodac.nii.ac.jp:10035/repositories/PLBSP_residue#query/28`. The page title is "Edit query". Below the title, there are several tabs: "show my namespaces, add a namespace", "link to query", "Reasoning", "Long parts", and "Contexts". The "show my namespaces, add a namespace" tab is selected. 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    rdfs:type sio:SIO_010432;
12    rdfs:seeAlso ?pdb_res.
13  ?pdb_res rdfs:type edam:data_T756;
14    rdfs:seeAlso ?unpres.
15  ?unpres rdfs:type edam:data_T756.
16  ?unpres dcterms:isPartOf ?uniprot.
17  ?unpres rdfs:label ?sqno.
18 } order by ?uniprot ?sqno
19
```

Below the query, there is an "Execute" button. The "Result" section is visible, showing a dropdown menu set to "Download as SPARQL JSON". Below this, a table displays the results of the query:

uniprot	sqno
O61363	"2551"
O61363	"2553"
O61363	"2555"
O61363	"2802"
O61363	"2803"
O61363	"2804"
O61363	"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)

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 a "Reasoning" checkbox which is checked. Other unchecked checkboxes are "Long parts" and "Contexts". The SPARQL query is as follows:

```
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_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
```

Below the query is an "Execute" button. The "Result" section shows a "Download as" dropdown menu with the following options:

- SPARQL JSON (selected)
- SPARQL XML
- SPARQL TTL
- SPARQL Comma-Separated Values
- Comma-Separated Values
- Comma-Separated Values (with fields in N-Triples syntax)

The results table shows the following data:

uniprot	SPARQL JSON
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。

研究開発内容

データセット生成に特化した二次データベース (PLBSPResidue) の開発

- PDB中のリガンド結合残基情報のRDF化
- リガンド間共有結合情報のRDF化
- PDBとUniProtの間のタンパク質鎖レベルの相互参照情報のRDF化

データセット生成パイプラインの開発

- データベースからデータセットを取得
- 配列冗長性の除去
- PSSMへ変換

予測ツール生成パイプラインの開発

- 特徴ベクトルへ変換
- SVMによる学習の実行
- 遺伝的アルゴリズムによるパラメーター最適化の自動化

システム公開

- UTProt Galaxy の立ち上げ、リガンド結合部位予測ツール生成ワークフローを開発
- 予測ツール管理バックエンドである UTProt CKAN の立ち上げ

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

研究開発内容

データセット生成に特化した二次データベース (PLBSPResidue) の開発

- PDB中のリガンド結合残基情報のRDF化
- リガンド間共有結合情報のRDF化
- PDBとUniProtの間のタンパク質鎖レベルの相互参照情報のRDF化

データセット生成パイプラインの開発

- データベースからデータセットを取得
- 配列冗長性の除去
- PSSMへ変換

予測ツール生成パイプラインの開発

- 特徴ベクトルへ変換
- SVMによる学習の実行
- 遺伝的アルゴリズムによるパラメーター最適化の自動化

システム公開

- UTProt Galaxy の立ち上げ、リガンド結合部位予測ツール生成ワークフローを開発
- 予測ツール管理バックエンドである UTProt CKAN の立ち上げ

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

正例データセット
リガンド結合残基

...G**S**ER...G**D**FLAL**D**LG**G**TNF ... RVLL...

負例データセット
リガンド結合残基から5~25残基離れた残基

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(1)

ツール名	$\log_2(C)$	$\log_2(\gamma)$	window	感度 [%]	特異度 [%]	AUC [%]
脂質結合部位 予測ツール	3	-12	18	59.2	75.6	74.2
鉄原子結合部位 予測ツール	10	-11	8	33.2	99.1	87.0
プリンヌクレオチド 結合部位予測 ツール	7	-17	22	31.3	98.0	82.3

脂質結合タンパク質、鉄原子結合タンパク質、プリンヌクレオチド結合タンパク質について、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実行することで結合部位予測ツール生成を行った。

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:

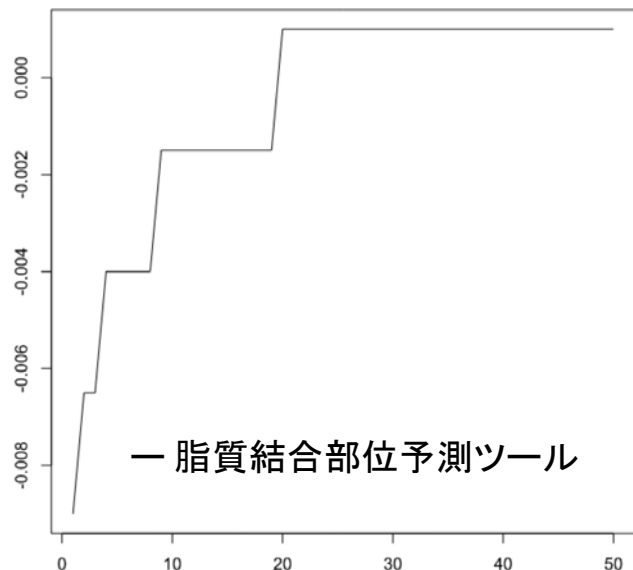
Point mutation 5%, Cross over 80 %

初期集団: 20, 世代数: 20

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(2): グリッドサーチとの比較

最大AUCの差

パラメーター 探索手法	交差検定の 実行回数	最大AUC の差
グリッドサーチ (刻み幅4)	$7^3 = 343$	
遺伝的アルゴリズム (世代数5)	100	-0.0042
遺伝的アルゴリズム (世代数10)	200	-0.0017
遺伝的アルゴリズム (世代数20)	400	0.8×10^{-3}

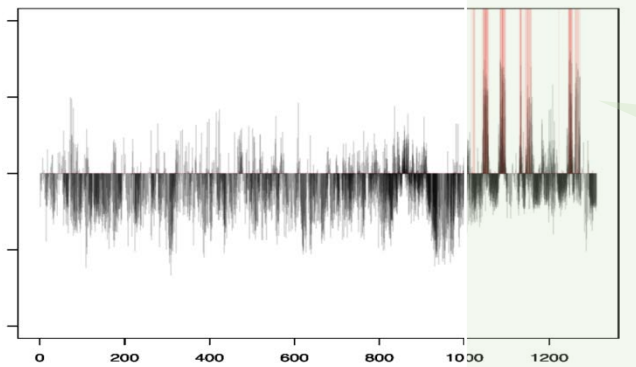


世代数

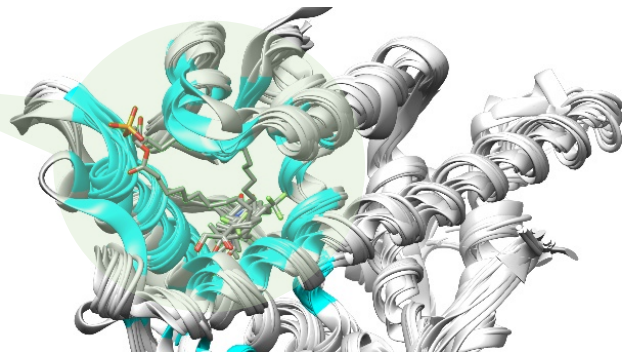
脂質結合タンパク質結合部位予測について、同じパラメーター探索空間に対して、刻み幅4のグリッドサーチを実行し、世代ごとの最大AUCの差をプロットした。
GAを用いたとき、初期世代の段階でもグリッドサーチとの差はかなり小さい。世代数5で計算したとき、最大AUCの差が0.4%以内で必要な計算の工数はおよそ1/3になる。

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(3): 核内受容体 脂質結合ドメインの例

(予測スコアのアラインメント)



(立体構造のアラインメント)



sp|O18473|ECR_HUMAN
sp|P06401|PRGR_HUMAN
sp|P08235|MKR_HUMAN
sp|P10275|ANDR_HUMAN
sp|P20153|USP_DROME

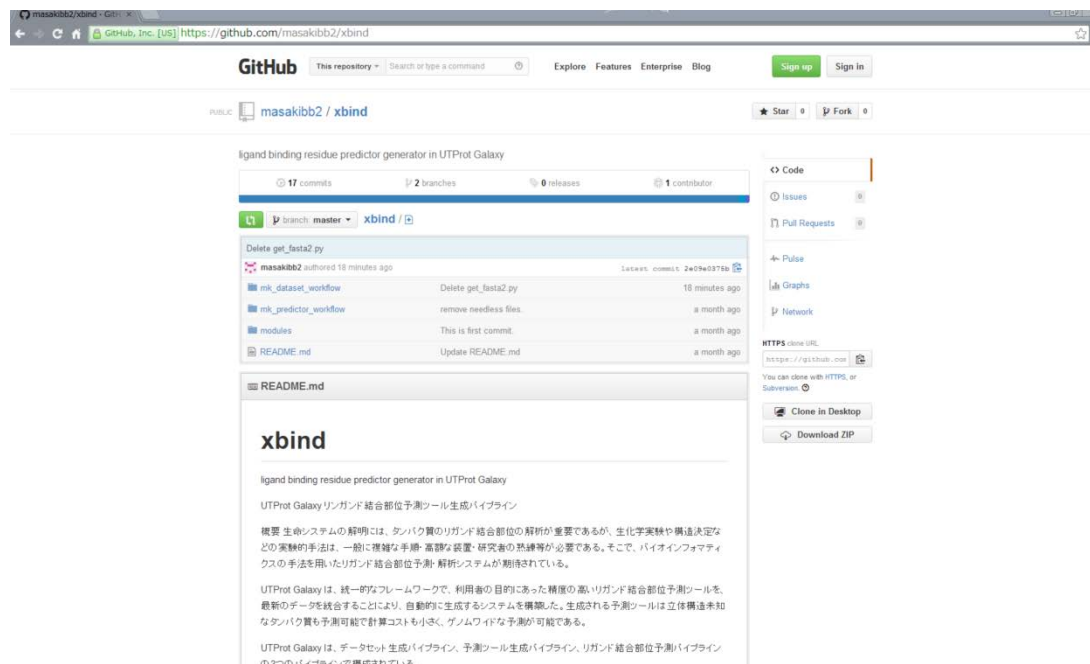
...RCSTVFANQACITRDNRRKAGMYVIEDLLFCRCMSSMNCWYHALLTAIVHSC--RPGLEQCLVVEIQRYVLTNTLRVHLNG---NSASPRGAWFQGLGLTEIRTLQMSNAGLSEKLNKRP...PFLEELVCYADVATAPH
...GKLYAPAPLVNENRMAVE--LCQAGASLQVRLQPEFETIMVLLLSITPKQGLKSGFAEVAR...VWIKELRWKVC--PNSCSGSRVYGLKLLSNHLLSLLEFCFTYERSHAKVEPALVYISDQ
...RMLYAPAPLVNENRMAVE--LCQAGASLQVRLQPEFETIMVLLLSITPKQGLKSGFAEVAR...VWIKELRWKVC--PNSCSGSRVYGLKLLSNHLLSLLEFCFTYERSHAKVEPALVYISDQ
...GKLYAPAPLVNENRMAVE--LCQAGASLQVRLQPEFETIMVLLLSITPKQGLKSGFAEVAR...VWIKELRWKVC--PNSCSGSRVYGLKLLSNHLLSLLEFCFTYERSHAKVEPALVYISDQ
...GKLYAPAPLVNENRMAVE--LCQAGASLQVRLQPEFETIMVLLLSITPKQGLKSGFAEVAR...VWIKELRWKVC--PNSCSGSRVYGLKLLSNHLLSLLEFCFTYERSHAKVEPALVYISDQ

(アミノ酸配列のアラインメント)

核内受容体脂質結合ドメインは、C末端領域に α -helixが折り重なった形で結合部位を形成する。結合部位はステロールなどが持つ親水性官能基の周りで水素結合し、それ以外の部分は疎水性相互作用をする。

生成された予測ツールは、 α -helix内に現れる保存された親水性残基の周辺を結合残基と予測することで、ヘリックス内で周期的に現れる結合残基を正しく予測している。

データセット生成、 予測ツール生成パイプラインの利用



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>

研究開発内容

データセット生成に特化した二次データベース (PLBSPResidue) の開発

- PDB中のリガンド結合残基情報のRDF化
- リガンド間共有結合情報のRDF化
- PDBとUniProtの間のタンパク質鎖レベルの相互参照情報のRDF化

データセット生成パイプラインの開発

- データベースからデータセットを取得
- 配列冗長性の除去
- PSSMへ変換

予測ツール生成パイプラインの開発

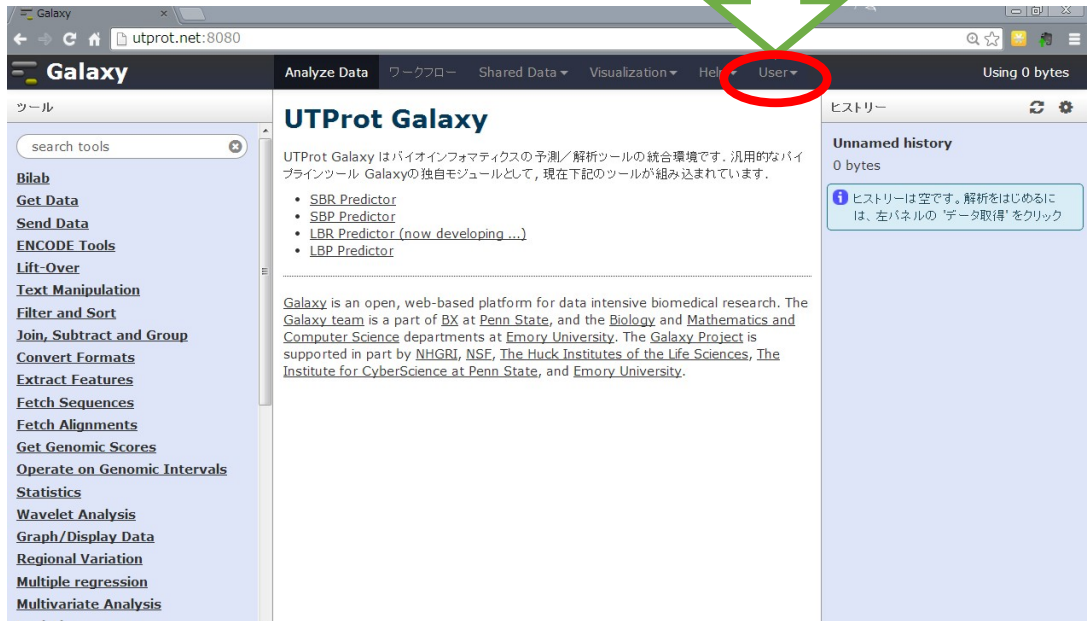
- 特徴ベクトルへ変換
- SVMによる学習の実行
- 遺伝的アルゴリズムによるパラメーター最適化の自動化

システム公開

- UTProt Galaxy の立ち上げ、リガンド結合部位予測ツール生成ワークフローを開発
- 予測ツール管理バックエンドである UTProt CKAN の立ち上げ

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' (Workflows), 'Shared Data', 'Visualization', 'Help', and 'User'. A search bar is located below the navigation bar.

Published Workflows

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	★★★★★		

The detailed view of the 'Predict Ligand Binding Residue' workflow is shown on the right. The workflow steps are:

- 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

The 'Import workflow' button in the top right corner of the workflow view is highlighted with a red circle.

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)

Galaxy

utprot.net:8080/root?workflow_id=3f5830403180d620

Analyze Data ワークフロー Shared Data

ツール

search tools

Bilab

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX table browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server

Upload File (version 1.1.3)

File Format:
Auto-detect

Which format? See help below

File:
ファイルを選択 選択されていません

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:
>148L|E|P0BID|CHAIN|SEQUENCE
MINIFEMLRIDEGLRKIKYDTEGYEIGIC
KPSLNAAKSELDKAGRTNGVITKDDEAKL
FNQDVDAAVRGILR
NAKLKPYVDSLDAVRRAALINVMVFQMGETGV

Here you may specify a list of URLs (one per line) or paste text into this file.

Convert spaces to tabs:
☐ Yes
Use this option if you are entering intervals by hand.

Genome:
unspecified (?)

Execute

230 bytes

1 ヒストリーは空です。解析をはじめるとは、左パネルの「データ取得」をクリック

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

utprot.net:8080/root?workflow_id=3f5830403180d620

Analyze Data ワークフロー Shared Data Visualization Help User

Using 1.2 MB

ツール

search tools

Bilab

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser

The following job has been successfully added to the queue:

2: Pasted Entry

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from "running" to "finished" if completed successfully or "error" if problems were encountered.

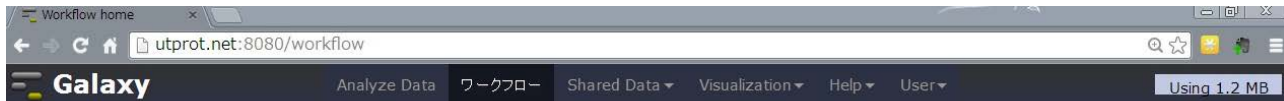
ヒストリー

Unnamed history
230 bytes

2: Pasted Entry

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)



Your workflows

Create new workflow

Upload or import workflow

Name

of Steps

imported: Predict Ligand Binding Residue ▼

10

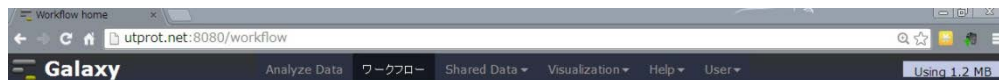
▼マークを選択

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu



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 displays the UTProt Galaxy web interface. The main workflow is titled "Running workflow \"imported: Predict Ligand Binding Residue\"". The workflow consists of several steps, with Step 1 and Step 2 highlighted by green callout boxes.

Step 1: GET ligand binding proteins, (version 1.0.0). The HETATOM code is set to MAN.

Step 2: PSSM generator using pssm. The Source Data is set to 2: Pasted Entry.

Callout Box 1 (Left): Step2 のSource Data に先ほどアップロードしたFASTAファイルを選択

Callout Box 2 (Right): 上段下段のHETATOM Code に予測対象とするリガンドのHETATOMコードを入力

The right sidebar shows a history panel with an unnamed history entry (460 bytes) and a pasted entry (2: Pasted Entry).

Step1, Step7 のHETATOMコードには予測対象となるリガンドのHETATOM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)

Galaxy

107.22.182.227:8080/root?workflow_id=a799d38679e985db

Analyze Data ワークフロー Shared Data Visualization Help User Using 104.0 MB

ツール

search tools

Bilab

- Get header from Fasta file.
- PSSM generator using psi-blast
- Remove Sequence Redundancy using CD-HIT.
- GFT ligand binding proteins.
- GFT fasta using UniProt URI.
- GFT binding residue using UniProt URI 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
- Acid Sugar Binding Protein

This dataset is large and only the first megabyte is shown below.
Show all | Save

http://purl.uniprot.org/uniprot/P12337 0 1 -0.120807500021
http://purl.uniprot.org/uniprot/P12337 1 1 -1.56897006727
http://purl.uniprot.org/uniprot/P12337 2 1 -0.683495163866
http://purl.uniprot.org/uniprot/P12337 3 1 -1.33442574154
http://purl.uniprot.org/uniprot/P12337 4 1 -0.516701448274
http://purl.uniprot.org/uniprot/P12337 5 1 -0.192014891173
http://purl.uniprot.org/uniprot/P12337 6 1 -1.20649827116
http://purl.uniprot.org/uniprot/P12337 7 1 -0.701368569129
http://purl.uniprot.org/uniprot/P12337 8 1 -0.600251679134
http://purl.uniprot.org/uniprot/P12337 9 1 -0.0636158752309
http://purl.uniprot.org/uniprot/P12337 10 1 -0.297029356832
http://purl.uniprot.org/uniprot/P12337 11 1 -0.820210899063
http://purl.uniprot.org/uniprot/P12337 12 1 -0.656340977844
http://purl.uniprot.org/uniprot/P12337 13 1 -0.261674698358
http://purl.uniprot.org/uniprot/P12337 14 1 -0.722041505063
http://purl.uniprot.org/uniprot/P12337 15 0 0.549536322664
http://purl.uniprot.org/uniprot/P12337 16 1 -0.479266426296
http://purl.uniprot.org/uniprot/P12337 17 0 0.793543686301
http://purl.uniprot.org/uniprot/P12337 18 1 -0.134231408052
http://purl.uniprot.org/uniprot/P12337 19 0 0.740765535453
http://purl.uniprot.org/uniprot/P12337 20 0 0.871240983477
http://purl.uniprot.org/uniprot/P12337 21 0 0.767510308493
http://purl.uniprot.org/uniprot/P12337 22 0 0.615681094552
http://purl.uniprot.org/uniprot/P12337 23 0 0.828590742248
http://purl.uniprot.org/uniprot/P12337 24 1 -0.377665385729

History

Unnamed history
16.4 MB

39: Predict Binding Residue using generated Prediction model on data 38, data 36, and data 23
30,388 lines
format: txt, database: 2

http://purl.uniprot.org/uniprot/P:
http://purl.uniprot.org/uniprot/P:
http://purl.uniprot.org/uniprot/P:
http://purl.uniprot.org/uniprot/P:
http://purl.uniprot.org/uniprot/P:
http://purl.uniprot.org/uniprot/P:

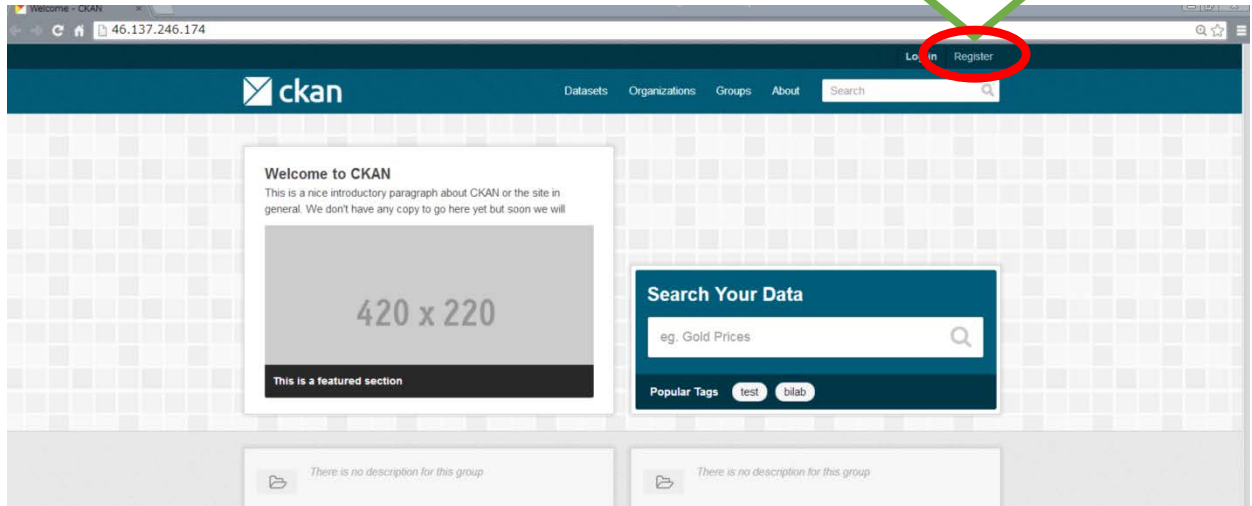
38: Create Prediction Model on data 24, data 36, and data 23

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。

予測ツール管理バックエンド: UTProt CKAN

上段メニューのRegisterで
ユーザー登録

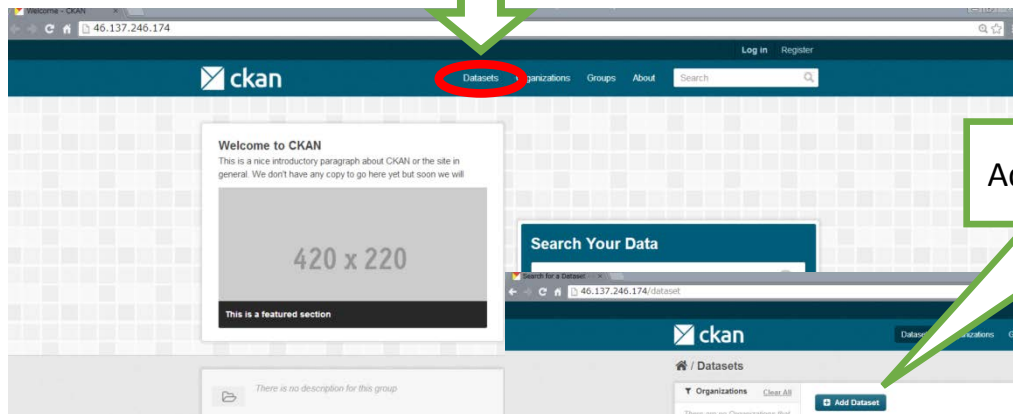


生成した予測ツールは、UTProt CKAN へ登録していくことで共有化できる
(アクセス方法は、<http://utprot.net/> → UTProt CKAN へリンクをたどる)。

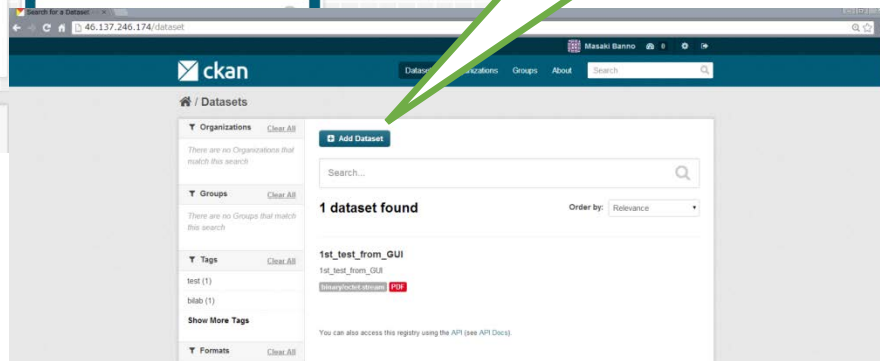
利用のためには、はじめにユーザー登録を行う必要がある。

予測ツール管理バックエンド:UTProt CKAN

上段2段目 Dataset を
クリック



Add Dataset をクリック



結果を登録するためには、上段メニュー2段目の Dataset を押す。
次に来たページで、Add Dataset をクリックする。

予測ツール管理バックエンド:UTProt CKAN

The screenshot shows the CKAN 'Create Dataset' form. The form has three steps: 1. Create dataset, 2. Add data, and 3. Additional data. The first step is active. The form fields are as follows:

- Title:** Glucose binding site prediction tool
- URL:** 46.137.246.174/dataset/glucose-binding-site-prediction-tool
- Description:** Glucose binding site prediction tool generated by UTProt Galaxy.
- Tags:** (empty)
- License:** License Not Specified

Annotations in Japanese:

- Title:** Title には予測器の名称を入れる
- Description:** Description には予測ツールの説明を自由に記述できる。
- Tags:** Tags へは、データセットに対してタグをつけて分類することができる。
- License:** License は登録する予測ツールのライセンス形態を設定する。
- Next Add Data:** 記載内容に問題がなければ、Next Add Data を押す。

CKANのデータセットはWindowsのフォルダのイメージに近く、複数のファイルを一つのデータセットへ登録していく。UTProt CKAN では、一つの予測器に対して一つのデータセットを作成することを推奨している。タグには、利便性の向上のため、“ligand binding site prediction” や、学習に用いたりガンドの分類や名称(“Glucose”、“lipid”など)もできるだけ付けることを推奨する。

予測ツール管理バックエンド:UTProt CKAN

The screenshot shows the CKAN 'Create Dataset' page. The browser address bar displays '7.246.174/dataset/new_resource/glucose-binding-site-prediction-tool'. The CKAN logo and navigation menu are at the top. A green banner at the top of the form area says 'Resource uploaded'. The page title is 'Datasets / Create Dataset'. The form has two steps: '1 Create dataset' and '2 Add data'. The 'Add data' step is active, showing options to 'Link to a file', 'Link to an API', or 'Upload a file'. The 'Upload a file' option is selected. The 'Resource' field contains the URL 'https://xbind.ckan2.s3.amazonaws.com/2014...'. The 'Name' field is 'model description file'. The 'Description' field contains 'Some useful notes about the data'. The 'Format' field is set to 'CSV'. The 'Save & add another' button is highlighted. The 'Next: Additional info' button is also visible.

Resource uploaded

/ Datasets / Create Dataset

What's a resource?
A resource can be any file or link to a file containing useful data.

1 Create dataset 2 Add data

Link to a file Link to an API Upload a file

Resource: https://xbind.ckan2.s3.amazonaws.com/2014...

Name: model description file

Description: Some useful notes about the data

You can use Markdown formatting here

Format: eg. CSV, XML or JSON

This is generated automatically. You can edit if you wish

Previous Save & add another Next: Additional info

ファイルアップロードは Upload fileを選択

Name はアップロードしたファイルの名称を設定

Description はアップロードしたファイルの説明を記述する

Format はアップロードするファイルのフォーマットを指定

さらにファイルを追加したのなら save & add another, 追加するファイルが無ければ Next: Additional infoをクリック

続いて個別のデータを登録していく。

予測ツールの再現性を保つため、予測ツールの学習条件・性能評価結果を記述したJSONファイルデータセットとして用いたPSSMファイルと結合残基を記述したファイルは必ず登録する

予測ツール管理バックエンド:UTProt CKAN

What are datasets?

Datasets are simply used to group related pieces of data. These can then be found under a single url with a description and licensing information.

1 Create dataset 2 Add data 3 Additional data

Author: Joe Bloggs

Author Email: joe@example.com

Maintainer: Joe Bloggs

Maintainer Email: joe@example.com

Custom Field: Key Value

Custom Field: Key Value

Custom Field: Key Value

Add Group: ligand binding prediction tool

Previous Finish

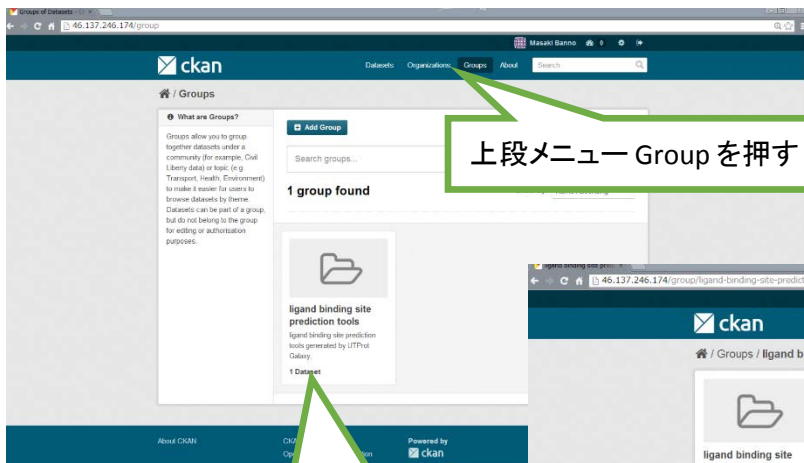
Ligand binding prediction tool のグループに追加する

最後にデータ登録者の情報を登録します。

第三者が予測ツール製作者とコンタクトを取るために用いる情報であり、空欄でも構わない。

利便性の向上のため、Ligand binding site prediction tools のグループに必ず追加する。

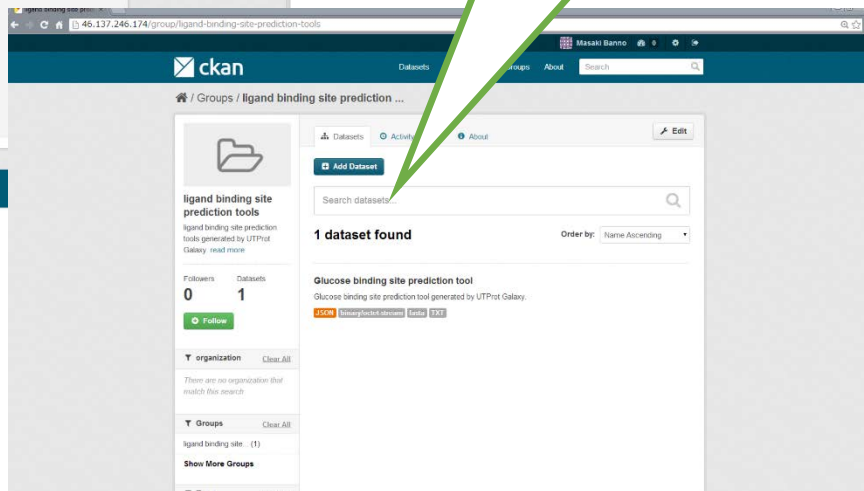
予測ツール管理バックエンド:UTProt CKAN



上段メニュー Group を押す

登録されたタグやdescriptionに
記述された内容に基づき予測
ツールの検索も可能

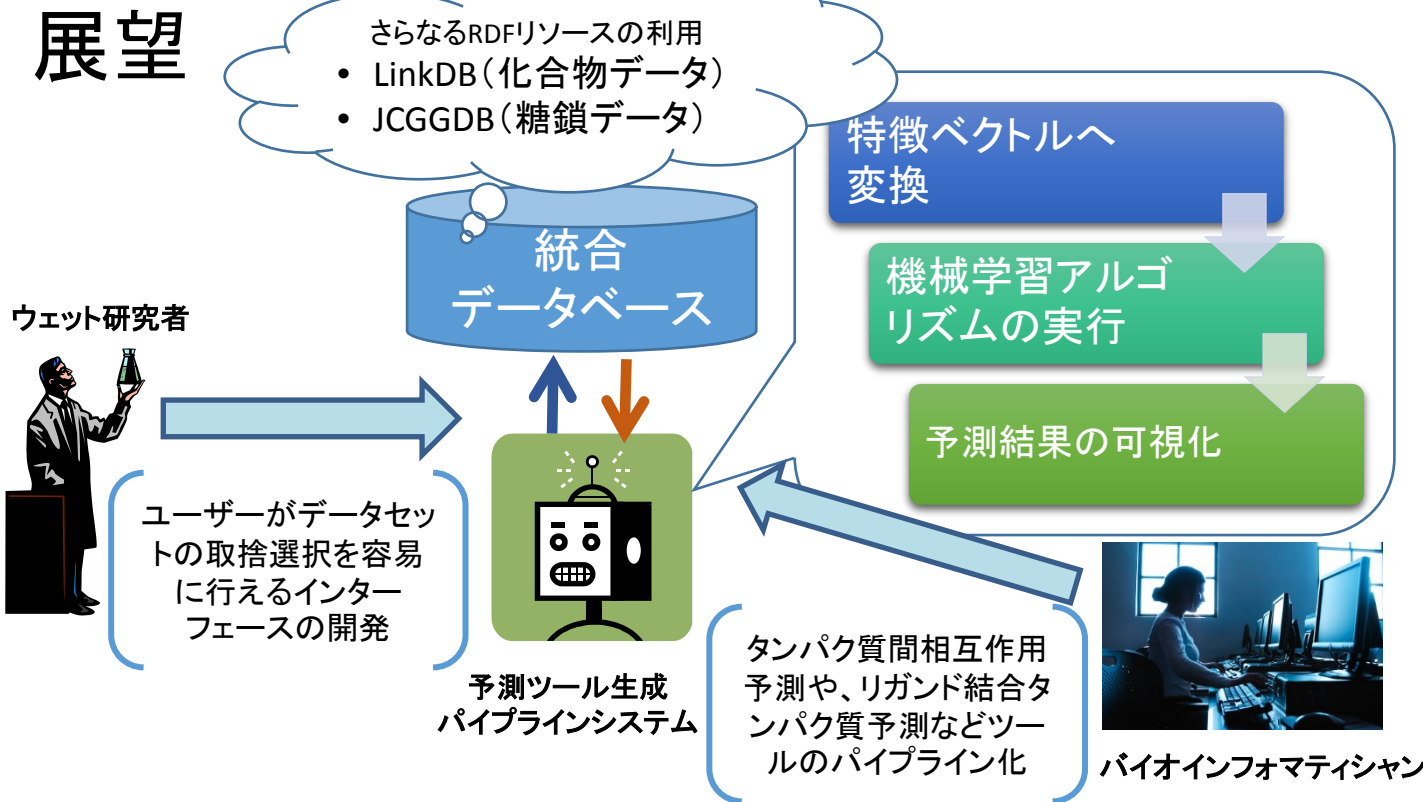
Ligand binding site
prediction tools で一覧取得



登録された予測ツールの一覧は、Group 以下にある ligand binding site prediction tools で見られる。

予測モデル登録の一連の操作を自動化する方法を今後検討していく。

展望



- 他のさまざまな予測問題に対する予測ツールを生成するパイプライン開発
- 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
- 実験者にとって便利な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

UTProt: プロジェクトトップページ

UTProt: ホーム
utprot.net

UTProt
The University of Tokyo Proteins

Products

Analytical Tools [\[all tools\]](#)

Prediction Tools (local tools)

- [SBR-Predictor](#): Sugar Binding Residue Predictor
- [SBP-Predictor](#): Sugar Binding Protein Predictor
- [LBP-Predictor](#): Lipid Binding Protein Predictor
- [GPL-Predictor](#): Predictor of Glycerophosphatidylinositol(GPI) lipid modification of a protein

Tool Integration (web application)

- [UTProt Galaxy](#): The pipeline integrated the predictive/analytical tools towards Interactome. [UTProt Galaxy overview](#)

Graph Database (Linked Open Data) [\[UTProt RDF Platform\]](#)

- [PLBSP](#): Protein-Ligand Binding Site Pair Graph Database
- [SBP](#): Sugar Binding Protein Graph Database
 - version 2014
- [TKIP](#): Tyrosine Kinase Interaction Pair Graph Database
 - version 2014
- [RDF-SIFTS](#): RDFized SIFTS Graph Database
 - version 2014

Image Library with Open License

- [UTProt Image](#)
 - (suspension: 2014/02-present)

News

- JSTライフサイエンスデータベース統合推進事業「統合化推進プログラム(統合データ解析トライアル)」平成25年度新規研究開発課題1に採択されました。詳細は[科学技術振興機構提報書 880号](#) および[NDC統合化推進プログラム\(統合データ解析トリアル\)採択課題](#)を御覧ください。(11, Sep. 2013)

Database (LOD)

[UTProt RDF Platform](#) (4)

Tools

[UTProt Galaxy](#)

[ALL Tools](#)

Link

[UTProt Home](#)

[BLAB](#)

[Department of Biotechnology](#)

[Graduate School of Agricultural and Life Sciences](#)

[The University of Tokyo](#)

[Platform for Drug Discovery, Informatics, and Structural Life Science](#)

[JST NDC](#)

Publish

[Articles](#)

Funding

[Grants](#)

Contact

[Members](#)

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

謝辞

本研究を進めるにあたり、協力してくださった生物情報工学研究室の皆様、

計算資源を提供してくださった国立情報学研究所の皆様、

本研究を支援してくださった研究総括、研究アドバイザー、NBDCの皆様に深く感謝を申し上げます。