

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

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。

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

wwPDB
/ RDF

PDB
ligand

LinkDB

Glycome
DB

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

サポートベクターマシンでの例

分離
超平面

リガンド結合残基

非リガンド結合残基

リガンド結合残基を予測

...SER**D**FLAL**D**LG**G**T...

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

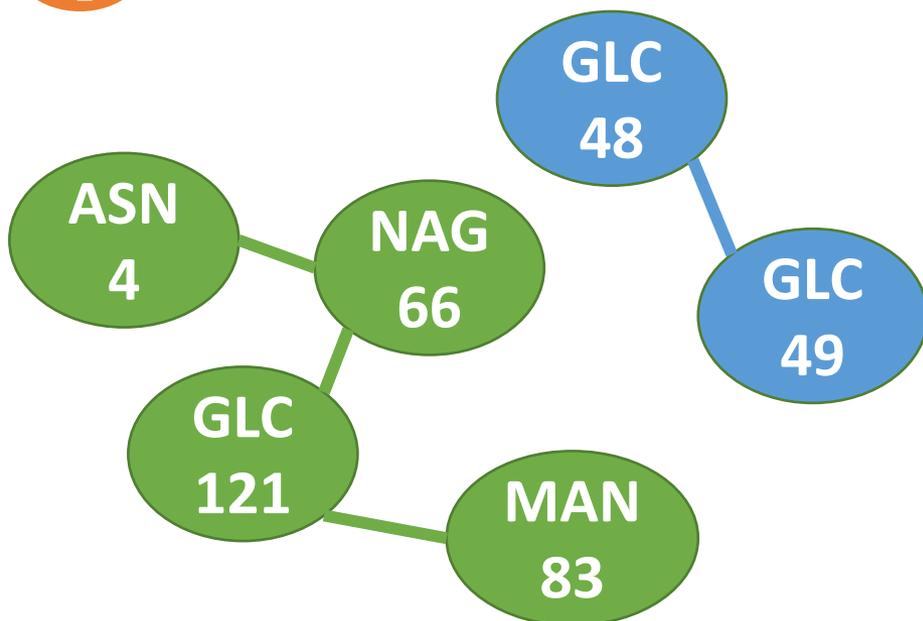
- PDB中の全タンパク質リガンド間原子間距離を網羅したデータベース (PLBSP)にデータ問い合わせを行ったが、データ取得が非常に遅かった。
- PDB中の残基番号とFastaファイル上のアミノ酸位置を対応付けが難しい場合がある。
- PDBファイルだけでは、リガンドであるか残基修飾であるか判断が難しい。

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)。

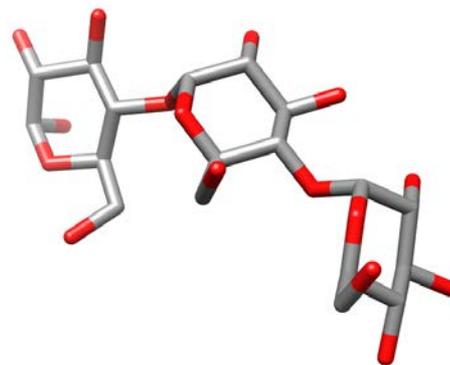
リガンド間共有結合情報のRDF化

1



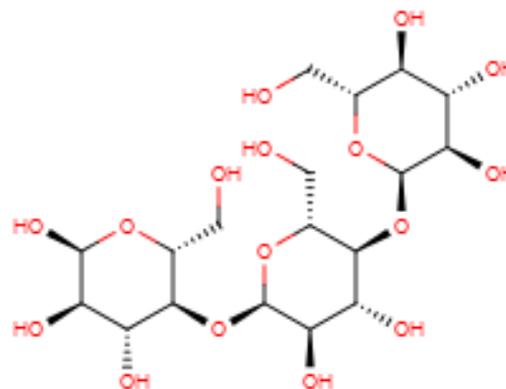
PDBのCONNECTレコードから残基をノード、共有結合関係をエッジとしたグラフデータを生成。経路発見アルゴリズムを用いて、一つの分子を形成するHETATMレコードの組み合わせをグループ化する。

2



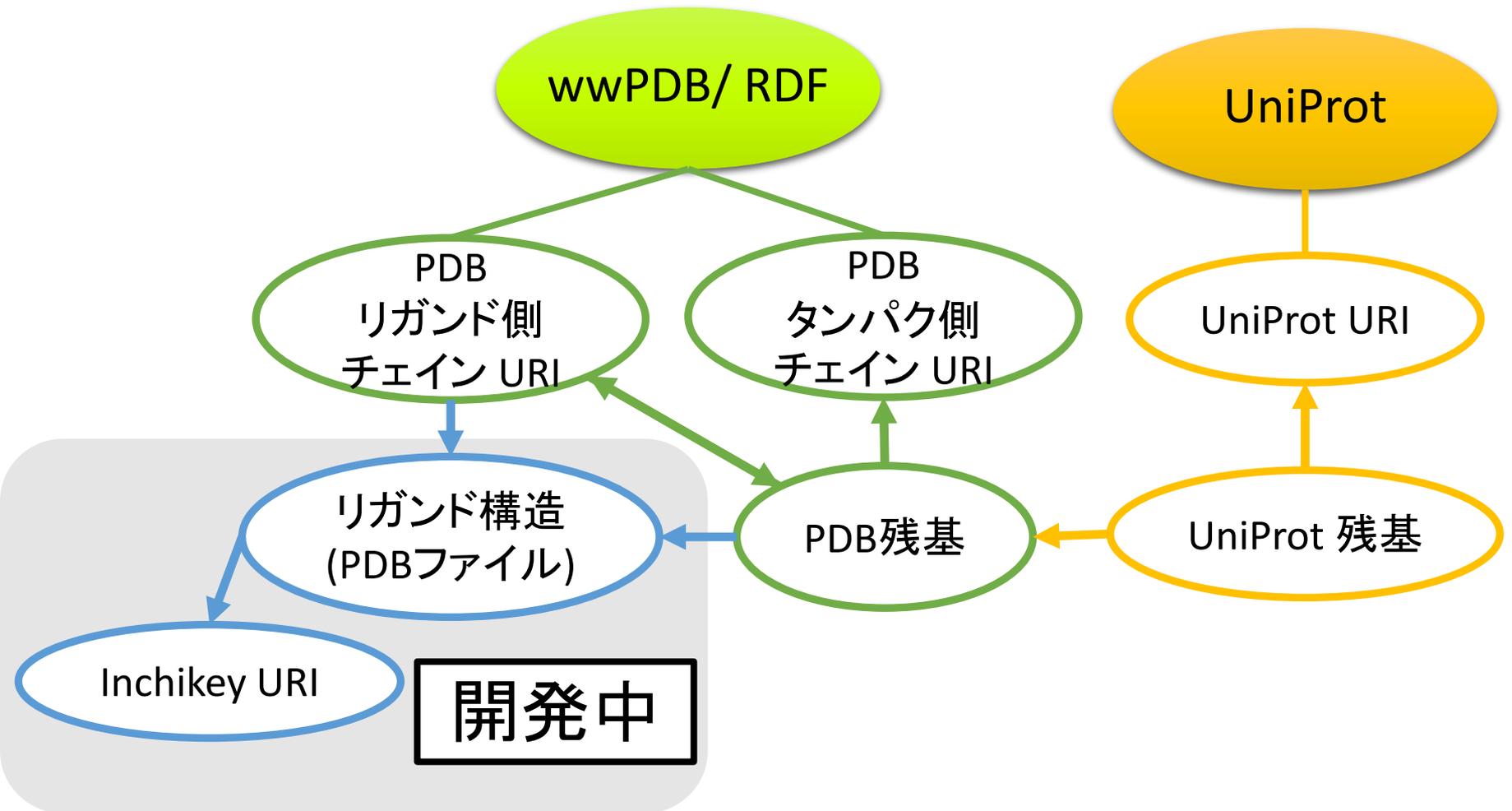
PDBファイル中からグループ化されたHETATMレコードを抜き出し、一つのPDBファイルとして保存。

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

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

2

正例データセット
糖結合残基

...GSER...GD**FLALD**LGGTNF ... RVL...L...

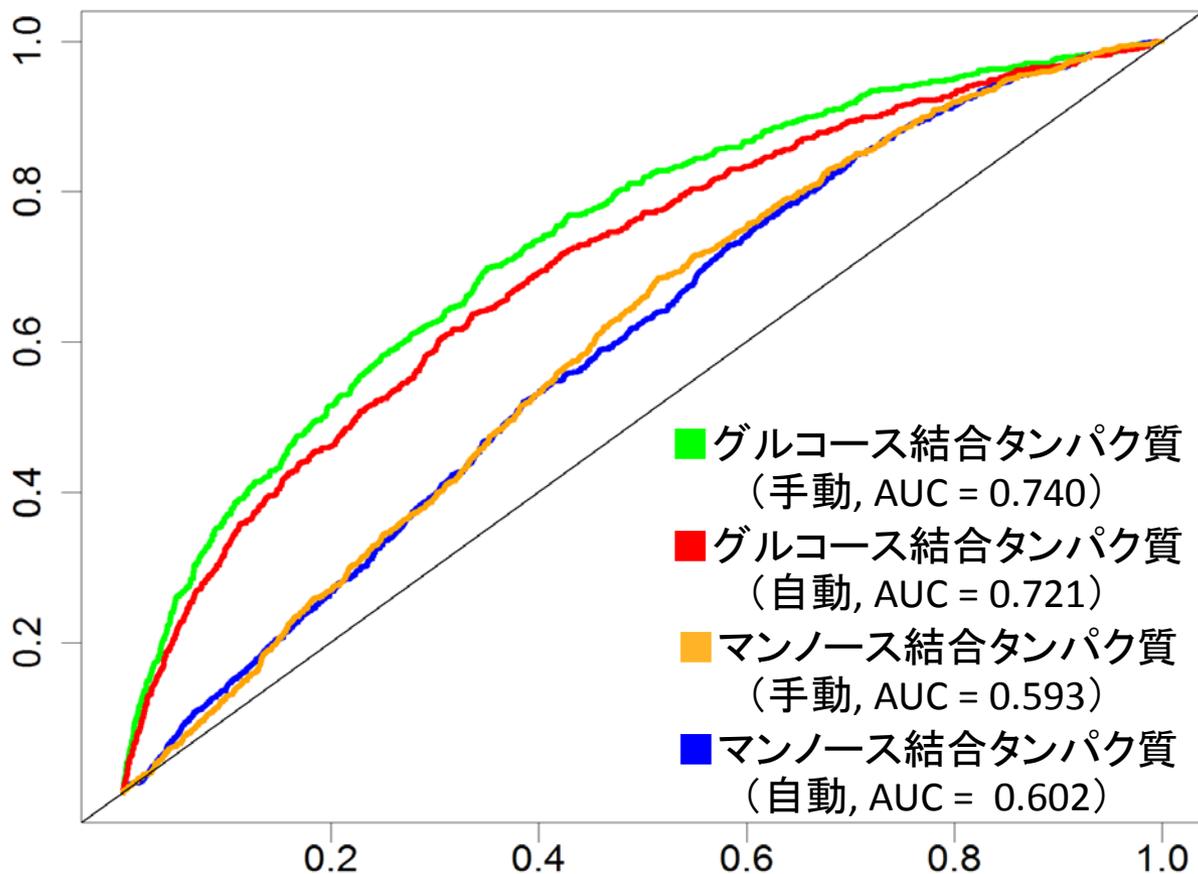
負例データセット

糖結合残基から 5~25残基離れた残基

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 : $c, g, w \in [-14, 14]$

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

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て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

Galaxy

utprot.net:8080

Galaxy Analyze Data ワークフロー Shared Data Visualization Help User Using 0 bytes

UTProt Galaxy

UTProt Galaxy はバイオインフォマティクスの予測/解析ツールの統合環境です。汎用的なバイ
ブライントール Galaxyの独自モジュールとして、現在下記のツールが組み込まれています。

- SBR Predictor
- SBP Predictor
- LBR Predictor (now developing ...)
- LBP Predictor

Galaxy is an open, web-based platform for data intensive biomedical research. The
Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and
Computer Science departments at Emory University. The Galaxy Project is
supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The
Institute for CyberScience at Penn State, and Emory University.

ツール

search 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

ヒストリー

Unnamed history

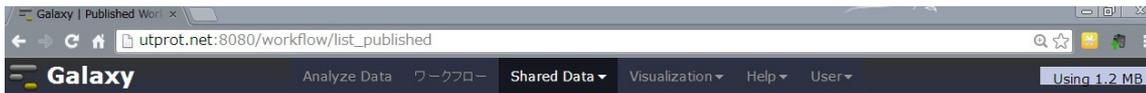
0 bytes

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

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

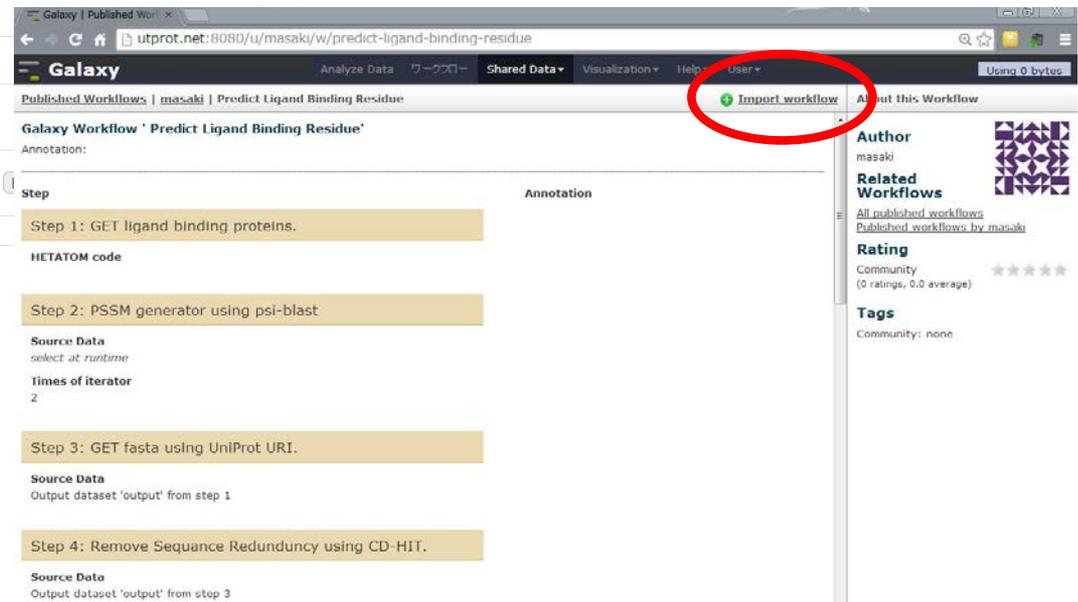


Published Workflows

search name, annotation, owner, and tags

[Advanced Search](#)

| Name | Annotation | Owner | Community Rating | Community Tags | Last Updated ↓ |
|---|---|--------|------------------|----------------|----------------|
| Create Ligand Binding Residue Predictor | | masaki | ★★★★★ | | |
| Get Non Redudunt Ligand Binding Protein | | masaki | ★★★★★ | | |
| Get Non Redudunt 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 | ★★★★★ | | |



<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)

Galaxy

utprot.net:8080/root

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

Send Data

ENCODE Tools

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

UTProt Galaxy

UTProt Galaxy はバイオインフォマティクスの予測/解析ツールの統合環境です。汎用的なバイ
ブライントール Galaxyの独自モジュールとして、現在下記のツールが組み込まれています。

- SBR Predictor
- SBP Predictor
- LBR Predictor (now developing ...)

歴史

Unnamed history

0 bytes

歴史は空です。解析をはじめるとは、左パネルの「データ取得」をクリック

左側メニュー「Get Data」を選択

Galaxy

utprot.net:8080/root

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
- BX table browser
- EBI SRA ENA SRA

UTProt Galaxy

UTProt Galaxy はバイオインフォマティクスの予測/解析ツールの統合環境です。汎用的なバイ
ブライントール Galaxyの独自モジュールとして、現在下記のツールが組み込まれています。

- SBR Predictor
- SBP Predictor
- LBR Predictor (now developing ...)
- LBP Predictor

歴史

Unnamed history

0 bytes

歴史は空です。解析をはじめるとは、左パネルの「データ取得」をクリック

「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)

Galaxy

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|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTGYEYIGL...
KSPSLNAAKSELDKAI GRNTNGVITKDEAEKL
FNQDVDAAVRGILR
NAKLKPVYDSLDAVRRRAALINMVFQMGETGV

Here you may specify a list of URLs (one per line) or paste the contents of a file.

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

Genome:
unspecified (?)

Execute

230 bytes

历史信息は空です。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する場合、ファイル選択を選ぶ

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。

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。

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

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.

历史信息

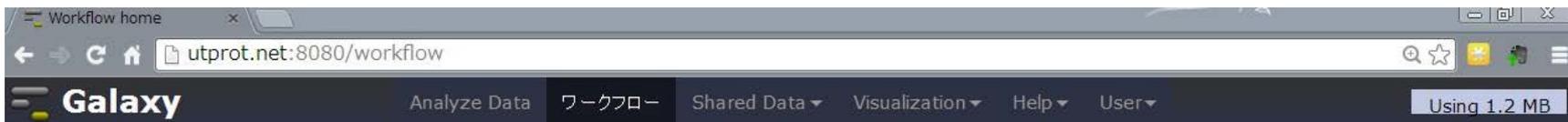
Unnamed history

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 |

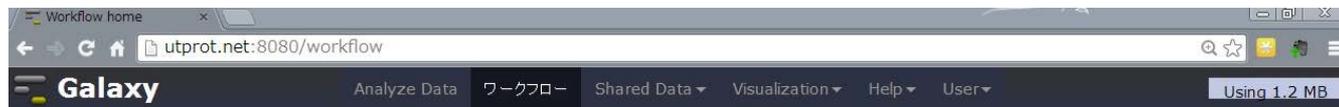
▼マークを選択

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu



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 Galaxy web interface for the workflow "imported: Predict Ligand Binding Residue". The workflow consists of several steps:

- Step 1: GET ligand binding proteins. (version 1.0.0) - HETATM code: MAN
- Step 2: PSSM generator using psi-blast. (version 1.0.0) - Source Data: 2: Pasted Entry
- Step 3: GET fasta using UniProt URI. (version 1.0.0)
- Step 4: Sequence Redundancy using CD-HIT. (version 1.0.0)
- Step 5: Extract from Fasta file. (version 1.0.0)
- Step 6: PSSM generator using psi-blast. (version 1.0.0)
- Step 7: Predict binding residue using UniProt URI and HETATM ID. (version 1.0.0)
- 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)

Annotations in the image:

- 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HETATM Code に予測対象とするリガンドのHETATMコードを入力

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)

Galaxy

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.
- 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
- 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 -0.120607500021
http://purl.uniprot.org/uniprot/P12337 1 1 -1.56937006727
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 -0.516701440274
http://purl.uniprot.org/uniprot/P12337 5 1 -0.192014891173
http://purl.uniprot.org/uniprot/P12337 6 1 -1.20649927116
http://purl.uniprot.org/uniprot/P12337 7 1 -0.701368569129
http://purl.uniprot.org/uniprot/P12337 8 1 -0.800251679134
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 0.815681094552
http://purl.uniprot.org/uniprot/P12337 23 0 0.828590742248
http://purl.uniprot.org/uniprot/P12337 24 1 -0.377665385729
```

ヒストリー

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: ?

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

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

BILAB Data Portal — masaki070540 ログアウト
データセット登録 検索 グループ About
データセット検索

GLC_bind_pred_ver1

表示 リソース (8) アプリ、アイデア等 (0) 履歴 設定 承認 フォロー フォロワー数 (0)

xbind バイブラインで生成したグルコース結合残基予測モデル まだテスト版で糖修飾を除けておらず、オリゴ糖判定も行っていません。

リソース (修正)

| | | |
|--------------------------|---|---------------------|
| <input type="checkbox"/> | 2013-11-22T073259/GLC_bind_all.UniProtURI.txt | txt |
| <input type="checkbox"/> | 2013-11-22T073415/GLC_bind_all.fasta | fasta |
| <input type="checkbox"/> | 2013-11-22T073512/GLC_bind_predictor.model | binary/octet-stream |
| <input type="checkbox"/> | 2013-11-22T073712/GLC_bind_represent.fasta | fasta |
| <input type="checkbox"/> | 2013-11-22T073823/GLC_bind_represent.pssm | txt |
| <input type="checkbox"/> | 2013-11-22T074002/answer_GLC_bind.txt | binary/octet-stream |
| <input type="checkbox"/> | 2013-11-22T074048/ga_summary_GLC_bind.txt | txt |

ライセンス: ライセンスが指定されていません

グループ

BILAB

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