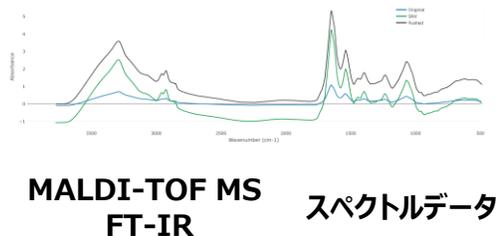




# MALDI-TOF MS・FT-IRデータ解析 同定困難な微生物菌株を識別する

フィルジェン株式会社  
バイオインフォマティクス部

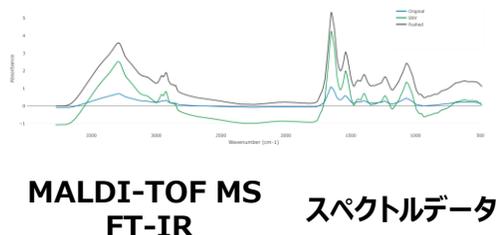
## フィンガープリント法



- データベースに登録された幅広い微生物群から識別
- データベースに登録のない微生物や亜種や株レベル、薬剤耐性・感受性株などより細かな識別に向かない

## プロテオタイピング バイオマーカーを使用

CLOVER  
BioSoft



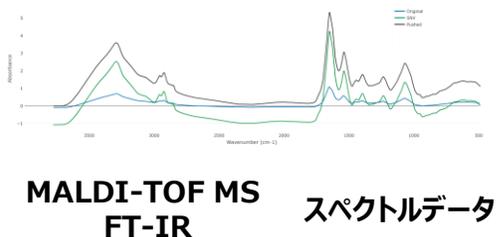
- 微生物や亜種や株レベル、薬剤耐性・感受性株などより細かな識別に対応
- バイオマーカーの検出が難しい場合や、サブタイプ分けに使用できるバイオマーカーがない場合は識別できない

## 機械学習アプローチ

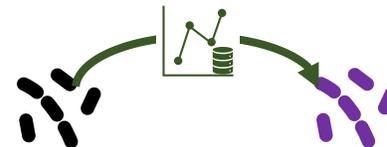
CLOVER  
BioSoft



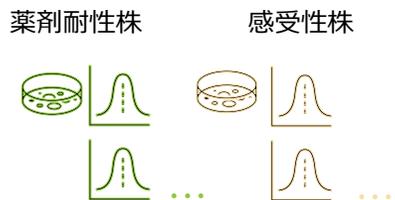
本日はこちらのアプローチをご紹介します。



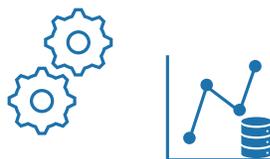
機械学習



作成された予測モデルを使用して  
識別



既知のサンプルから  
識別したいサンプルのカテゴリを作成



機械学習 (ML) による  
トレーニング & 予測モデルの取得



予想モデル構築に使用しなかった  
既知のサンプルを使用しモデルの精度を検証



未知のサンプルを使用して識別

- バイオマーカーの検出が難しい場合や、サブタイプ分けに使用できるバイオマーカーがないケースに対応

CLOVER  
BioSoft



- 非常に直感的で習得が容易  
さらに、Web ベースなのでどこからでもアクセス
- ベンダーに依存しない  
(様々な装置メーカーのデータを処理)
- MALDI-TOF MS と FT-IR データを  
同じプラットフォームで処理できる初のソフトウェア



通常、予測モデルの構築は複雑な工程を要しますが、  
本ソフトウェアでは、スペクトルのアップロードから検証のための予測モデルの構築を  
迅速かつ正確なワークフローで解析できます。

# 解析行程がパイプライン化

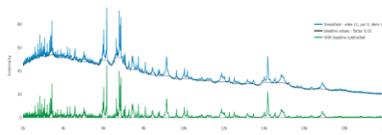


データ  
インポート

実験デザイン  
カテゴリ付け

- ノイズ除去
- ベースライン減算
- スムージング (平滑化)
- マスフィルタリング
- レプリケート指定
- トレーニング/テストデータ指定

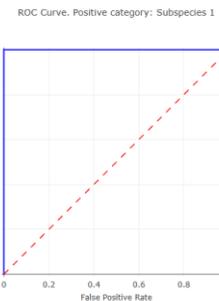
スペクトルデータ  
前処理



ピークマトリクス  
作成

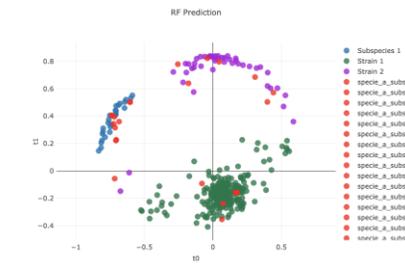
- アライメント
- ブランク除去
- ピーク検出
- 正規化

予測モデル  
作成

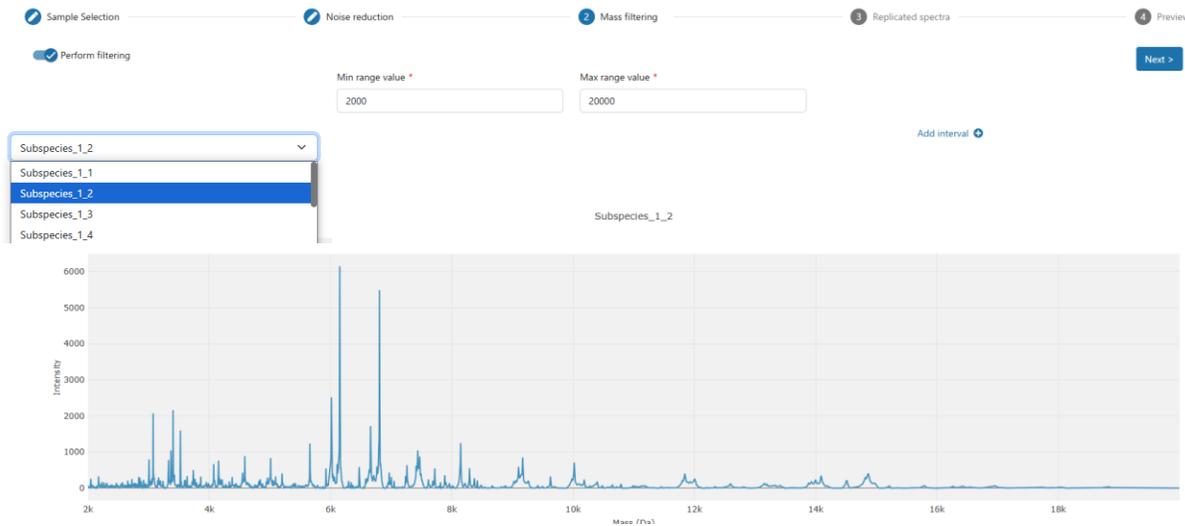


予測モデル  
検証

識別



スペクトルデータの「前処理」や「ピークマトリクス作成」工程ではプレビュー画面が表示されます。



マスフィルタリングの設定画面  
サンプルを切り替えとスペクトルデータも切り替わります。  
この設定の場合は指定された範囲のスペクトルデータがプレビューされます。

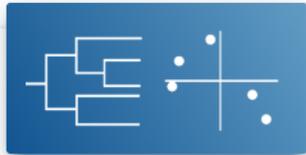


アライメントの設定画面  
指定したサンプルのスペクトルデータが表示されます。  
色分けや3Dの切り替えが可能です。

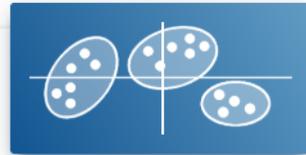
予測用（トレーニング）データが自動的に指定され解析が実行されます。

## Clustering

Identify groups where samples within the same group are more similar to each other than to those in other groups. This process does not require predefined categories as it discovers patterns on its own.



Hierarchical Clustering



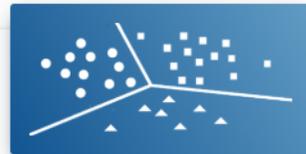
K-Means Clustering

## Classifiers

Train a prediction model for future blind identification. Requires predefined categories to learn from known examples.



PLS-DA



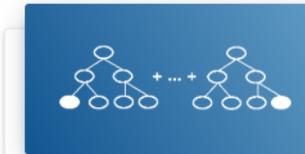
Support Vector Machine  
(SVM)



Random Forest



K Nearest Neighbors  
(KNN)



LightGBM

予測モデル構築のための適切なアルゴリズムを選択できるように、多くのアプローチから選択可能です。

作成後はk-fold Cross Validationなどの機械学習モデルの性能を評価するための指標やROCグラフなどから精度を確認できます。

k-fold Cross Validation

Test with 10% (10-fold CV)

Test with 25% (4-fold CV)

Test with 20% (5-fold CV)

**Test with 10% (10-fold CV)**

Test with 1 item (leave-one-out CV)

Shapley values

Retrieve Analysis

k-fold Cross Validation

Pick Positive Category

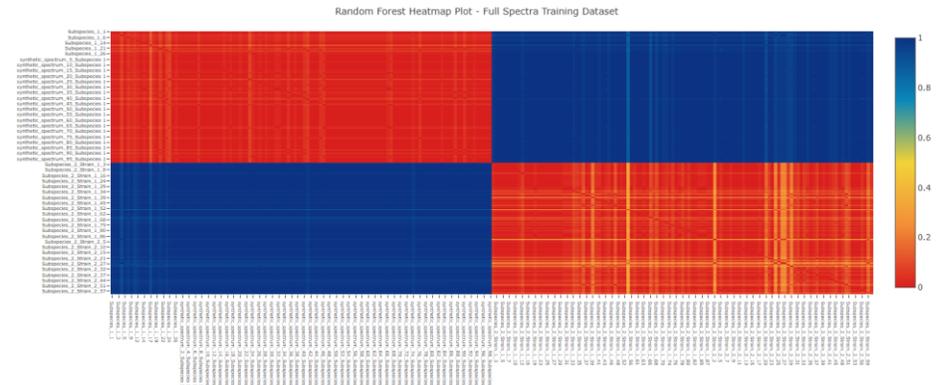
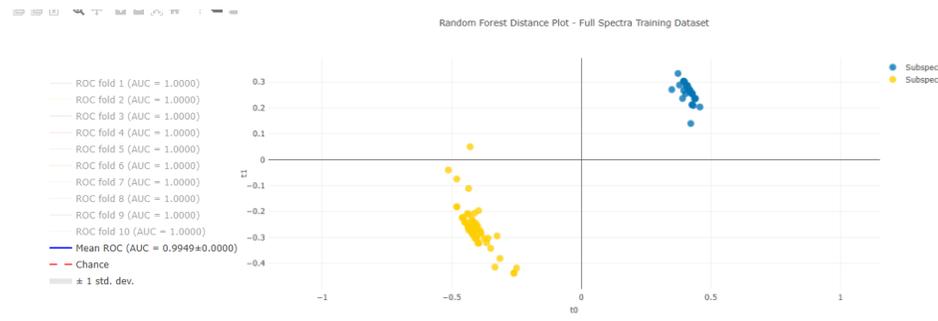
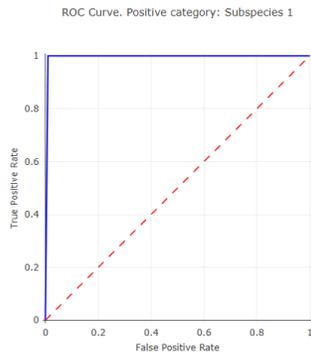
Subspecies 1

10-fold cross validation ⓘ Snapshot

Actual / Predicted	Subspecies 1	Subspecies 2	% Correct
Subspecies 1	119 (TP)	0 (FN)	100 %
Subspecies 2	0 (FP)	119 (TN)	100 %
	100 %	100 %	100 %

Confusion Matrix

Accuracy: 100 %      Sensitivity: 100 %      Positive Predictive Value or Precision: 100 %  
Balanced Accuracy: 100 %      Specificity: 100 %      Negative Predictive Value: 100 %  
F1 score: 100 % ⓘ      Error rate: 0 %



検証用（テスト）データを指定してするだけで自動で解析が実行されます。

ここで使用する検証用のデータセットは、すでにカテゴリーが明確（例：亜種AやBと予めわかったサンプル）かつ予測モデル構築に使用しなかったサンプルを使用します。

Select input:

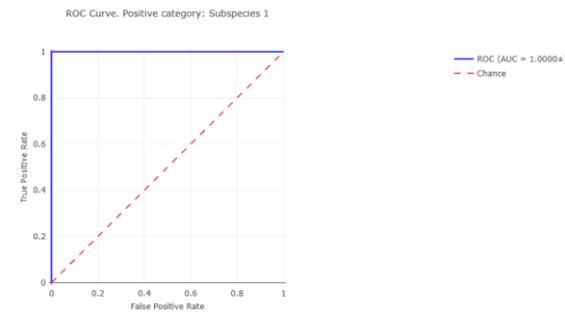
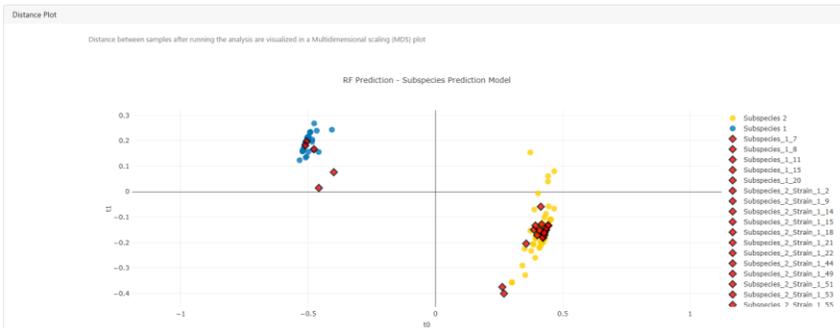
Upload files

Select files

Test Dataset

検証用のデータセットはスペクトルデータの前処理ですでに分けられたテストセットを使うことや新たにデータをインポートすることも可能です。新たにデータをインポートした場合は、予測モデルで使用したサンプルと同じ前処理を1クリックで反映させることができます。

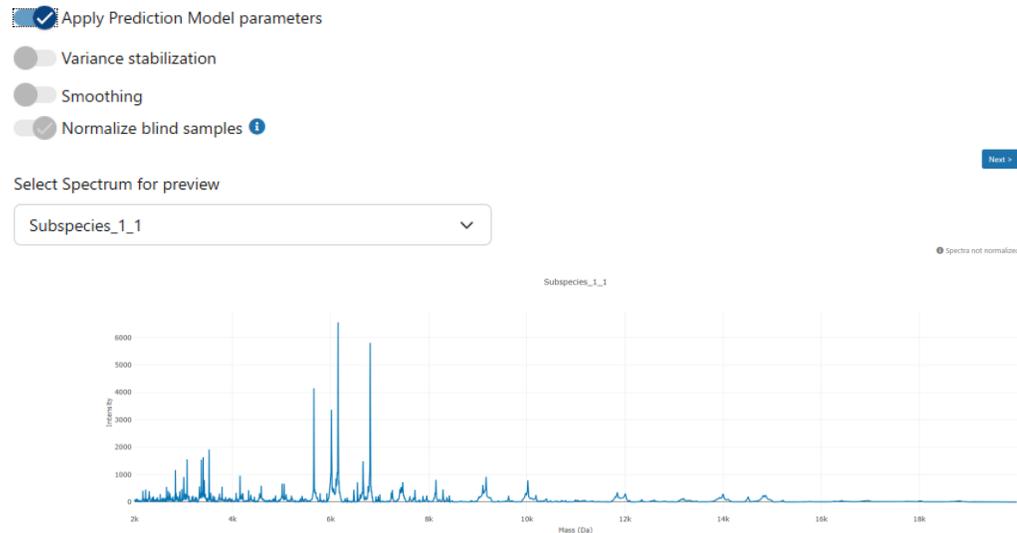
<input type="checkbox"/>	Name	Control Category	Predicted Category (#1)	Probability (#1)
<input checked="" type="checkbox"/>	Subspecies_1_7	Subspecies 1	Subspecies 1	100 %
<input checked="" type="checkbox"/>	Subspecies_1_8	Subspecies 1	Subspecies 1	98 %
<input checked="" type="checkbox"/>	Subspecies_1_11	Subspecies 1	Subspecies 1	96 %
<input checked="" type="checkbox"/>	Subspecies_1_15	Subspecies 1	Subspecies 1	84 %



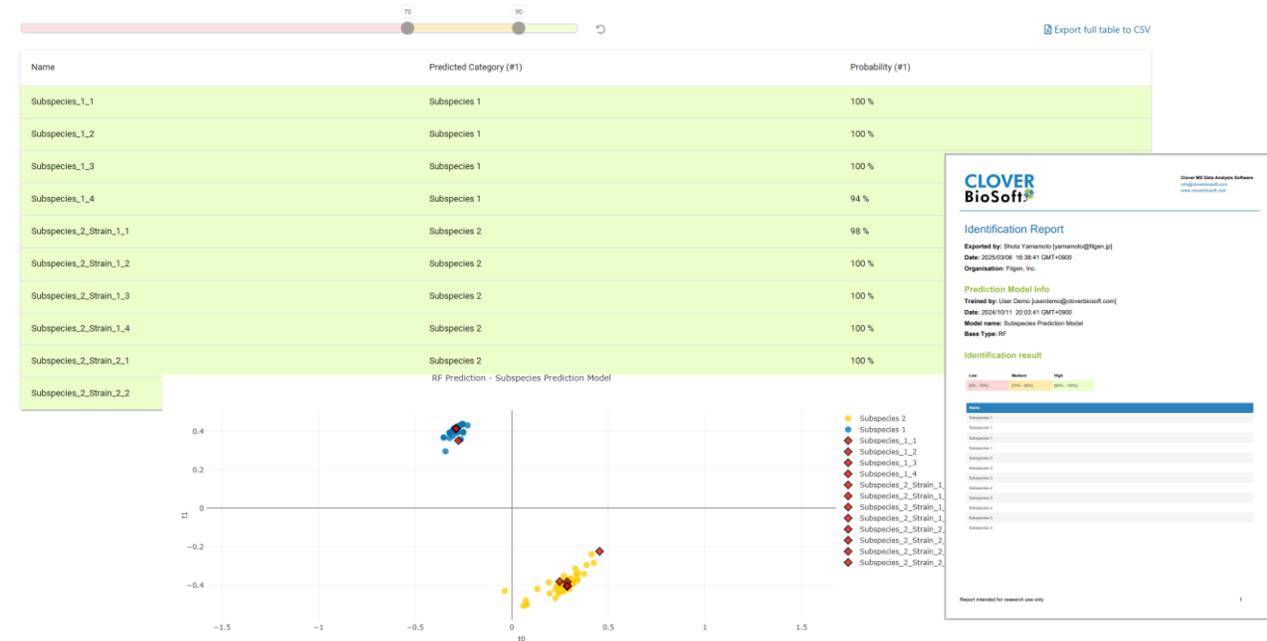
識別されたサンプルが適切なカテゴリーに分けられたか、表、グラフ、レポートなどから確認できます。

識別用データを指定してするだけで自動で解析が実行されます。

ここで使用する識別用のデータセットは、カテゴリーが不明確な（例：亜種AやBが分からない）サンプルを使用します。



使用するスペクトルデータの前処理をモデル構築時と同じにするこや、新たに処理を設定することができます。



識別結果は表、グラフ、レポートなどが取得できます。

お問い合わせ先：フィルジェン株式会社

TEL 052-624-4388 (9:00～17 : 00)

FAX 052-624-4389

E-mail: [support@filgen.jp](mailto:support@filgen.jp)