2024/11/11

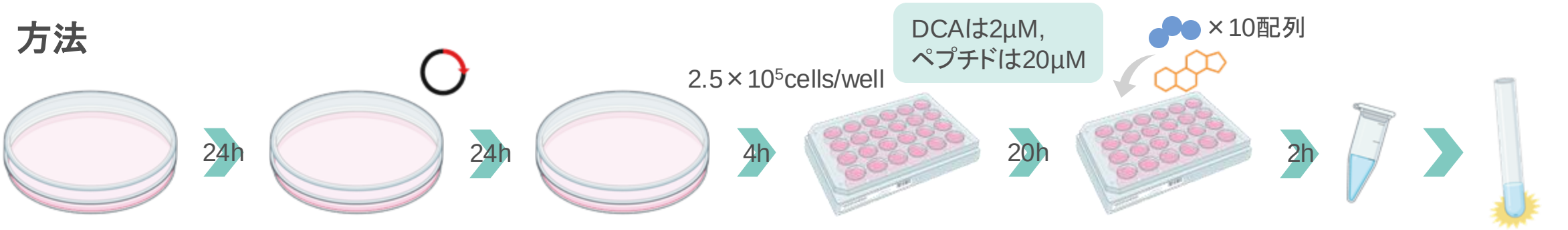
Lucレポーター遺伝子発現系を使ったTGR5アゴニストペプチドの探索

B4 中村 優作

ペプチドアレイで合成した上位10配列のTGR5活性の評価

目的：機械学習によりTGR5活性を**示す**と予測されたペプチドを実際に評価する。

方法



結果

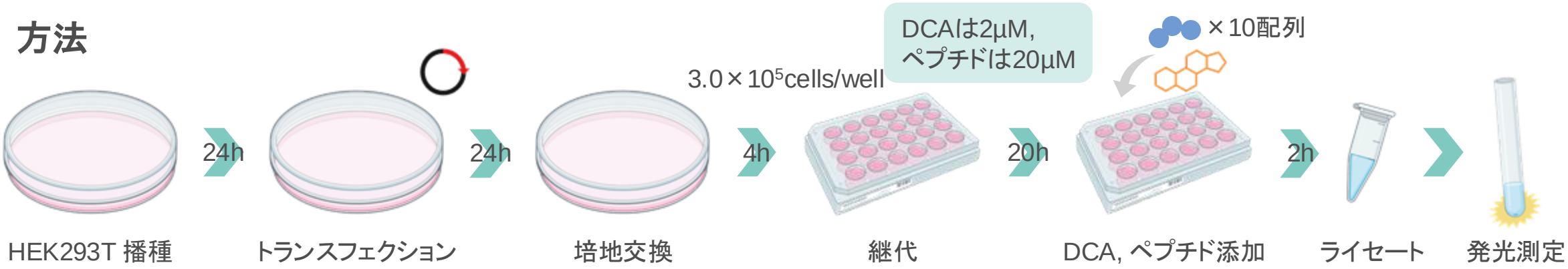
- 全体的に活性が出なかった。
- 誤差が大きく、controlよりマイナスになっているものもある。
  - DCA, ペプチド添加の時点で細胞数が6割程度で少なかった。

→ 正しく評価できる細胞数で、再度実験を行う。

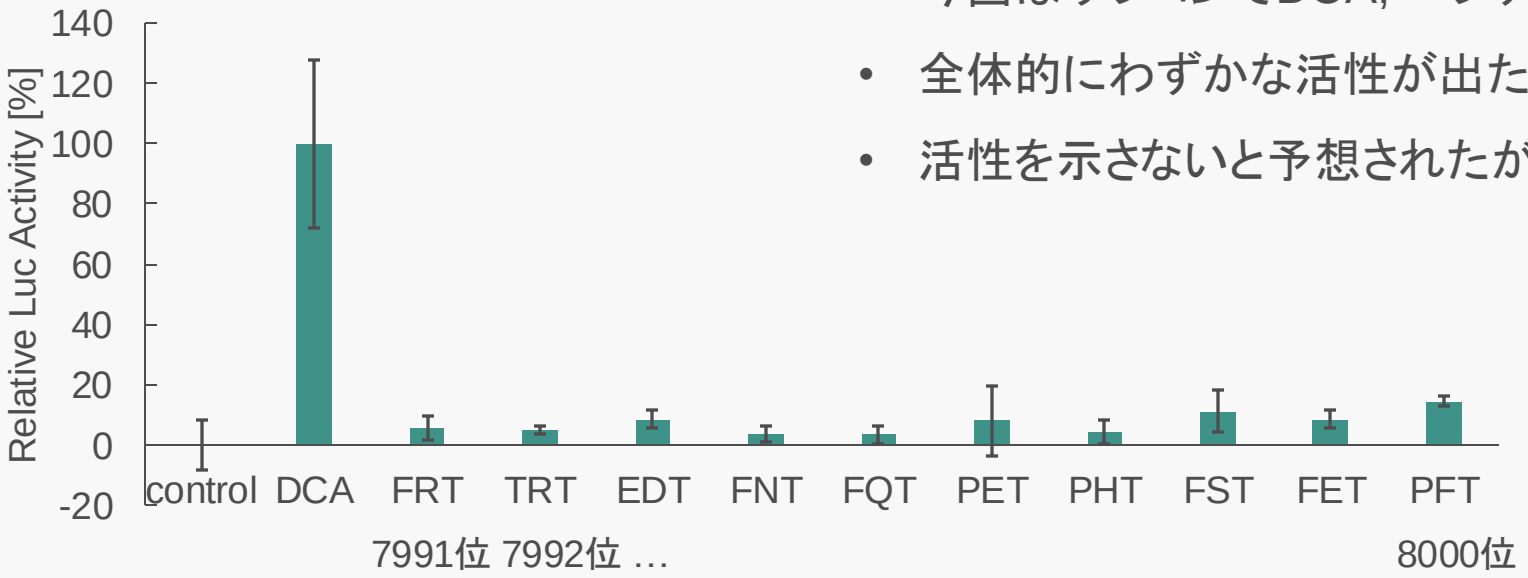
ペプチドアレイで合成した下位10配列のTGR5活性の評価

目的：機械学習によりTGR5活性を**示さない**と予測されたペプチドを実際に評価する。

方法



結果



- 今回はサブコンでDCA, ペプチド添加を行うことができた。
- 全体的にわずかな活性が出た。
- 活性を示さないと予想されたが、上位10配列よりも少し活性が高かった。

予測上位の配列で優位に活性が出ず、予測下位の方が若干活性が高かったことから、
モデルが未知のペプチドに対する予測性能を持っていない可能性が高い。

考えられる原因

- **過学習**
  - たまたま上位に含まれている構造を重要な構造として学習している?
- **データセットが不完全**
  - データ内のTGR5活性の測定方法、測定者が違うことによる影響を受けている?
  - そもそもの重要な構造を学習するためのデータが足りない?
- **構造以外の重要なパラメータ**
  - 現在は構造のみを学習データに使っているが、疎水性値などの生物物理学的特性が必要?

1.データ取得

ChEMBL
BindingDB

TGR5のEC50値の
1773件のデータ

2.前処理

活性あり

上位30%
(EC50 <= 70μM)

活性なし

下位30%
(EC50 >= 1000μM)

3.学習

フィンガープリントを入力

ランダムフォレスト
モデル

4.予測

×8000

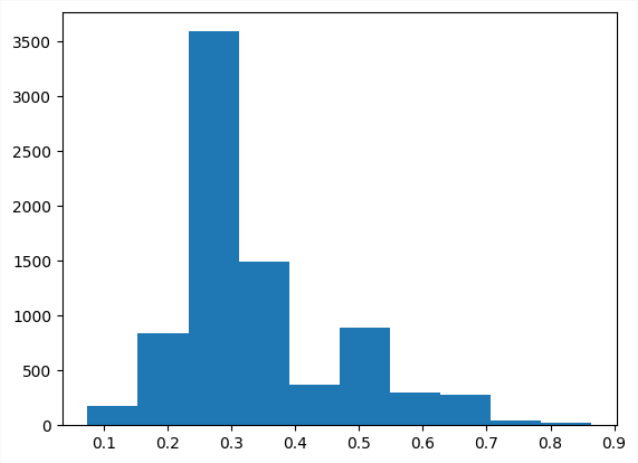
3残基ペプチド

活性予測

モデルの性能

	予測値 活性あり	予測値 活性なし
実測値 活性あり	86	13
実測値 活性なし	11	96

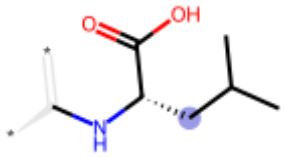
教師データの予測結果(予測精度88.3%)



3残基ペプチドの活性予測結果

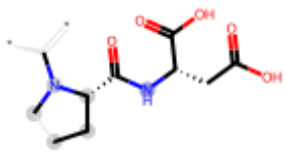
方法 上位10配列の各ペプチドに含まれる部分構造の中で、
ランダムフォレストモデル内での重要度が一番高い構造の、データセット内での出現頻度を調べる

結果



1位

活性あり:33個
活性なし:3個



2位

活性あり:193個
活性なし:60個

3位～10位

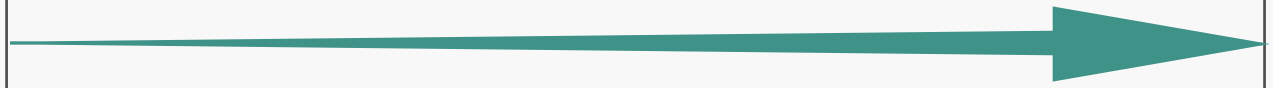
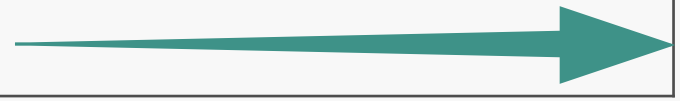
活性あり:261個
活性なし:144個

考察

- 重要度が高い構造は、活性なしのデータに比べて活性ありのデータで多く出現した。
- このことは予測根拠としては正しいが、実際に実験したところ活性が出なかったため、
  これらは重要な構造ではなかったと言える。(たまたま活性ありのデータに出現していた構造？)
- 活性があるかないかの分類モデルを作成したことで、細かい重要な構造を
  モデルが捉えることができなかったのではないかと考えられる。
- 回帰モデルで予測ができないかの検討をする。

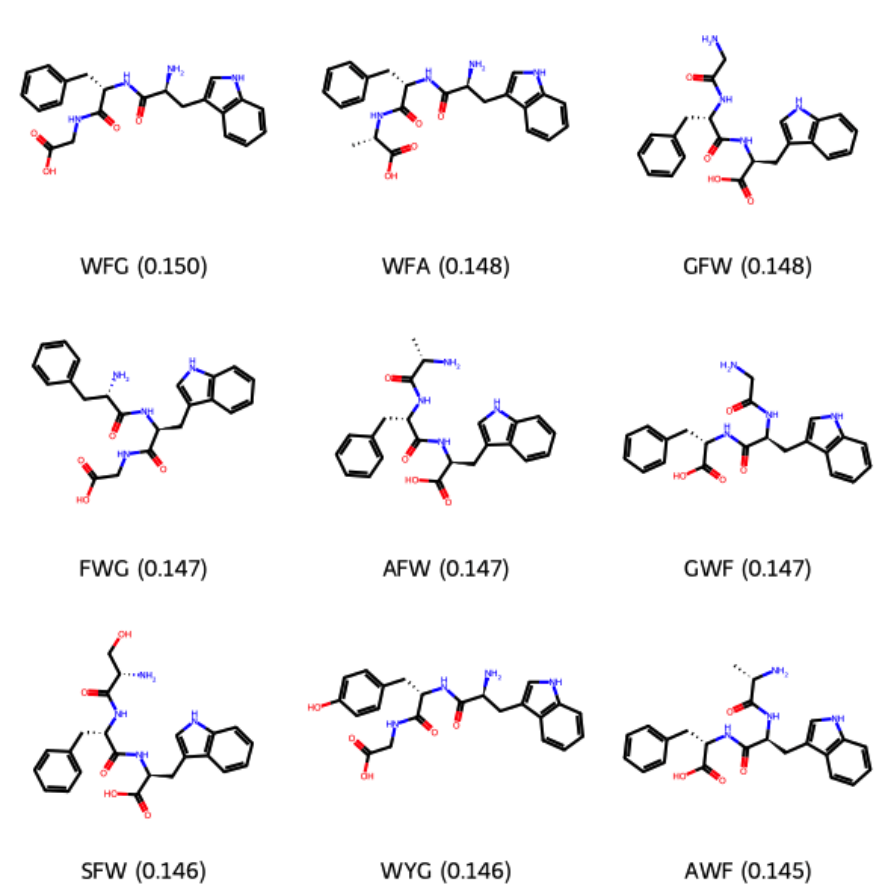
- **モデルの再構築**
  - データセット中の分子について、アッセイ方法を調べ、明らかに他と手法が異なるデータがあれば削除する
  - 回帰モデルで活性予測ができるか試してみる
  - モデルができたら、自分で実際に活性が出なかったペプチドの予測をして、予測結果が活性なしになるかを確認する
- **ペプチドの合成・評価**
  - モデル構築までに時間がかかるため、並行してペプチドの合成、評価を行う
  - 細胞数が少なく、結果のブレが大きかった上位ペプチドを再度評価
  - 清水さんのDCA結合ペプチドをペプチドアレイで合成して評価
  - 構造類似度を予測根拠とするペプチドの合成・評価
- **データが溜まったら、そのデータを用いた活性予測をする**

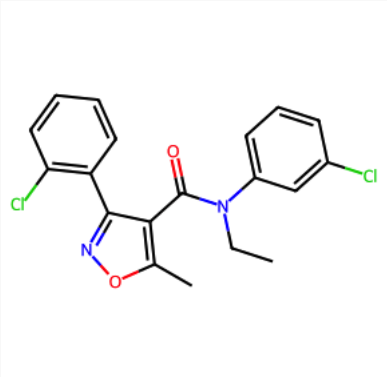
短期予定

	11/11～11/15	11/18～11/22	11/25～11/29
モデル解析・構築			
ペプチド合成(上位ペプチド、DCA結合ペプチド)			
TGR5活性評価			

各フィンガープリント手法で類似度の高かった上位 9 配列

KCF-S





hTGR5 pEC50 = 5.3

ドナーアクセプターペア