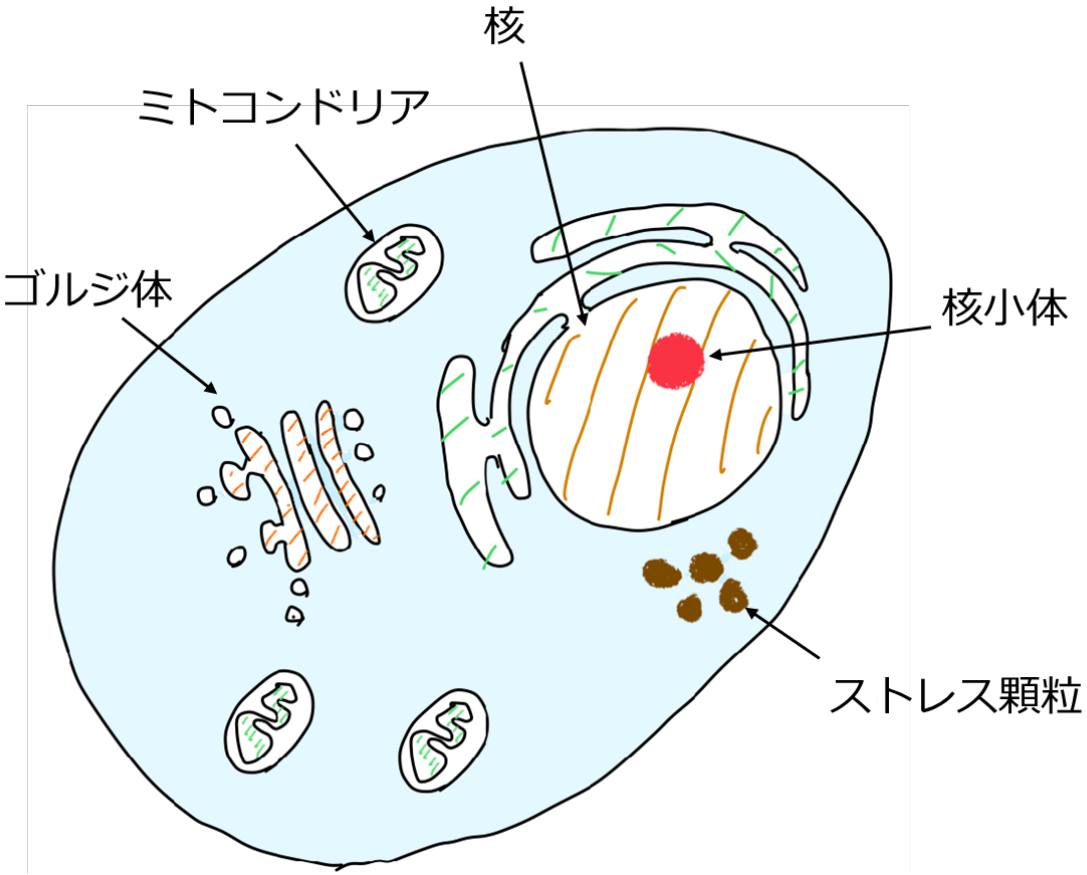
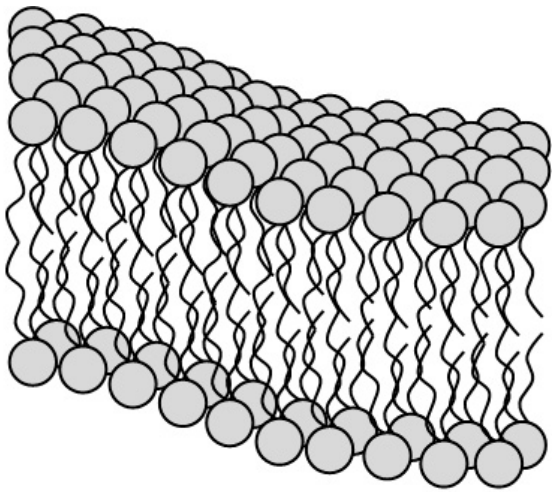
# 機械学習を用いてタンパク質の文法を理解する

**京都大学理学研究科 生物科学専攻 M1**

**菱田温規**

# 細胞内の区画化構造



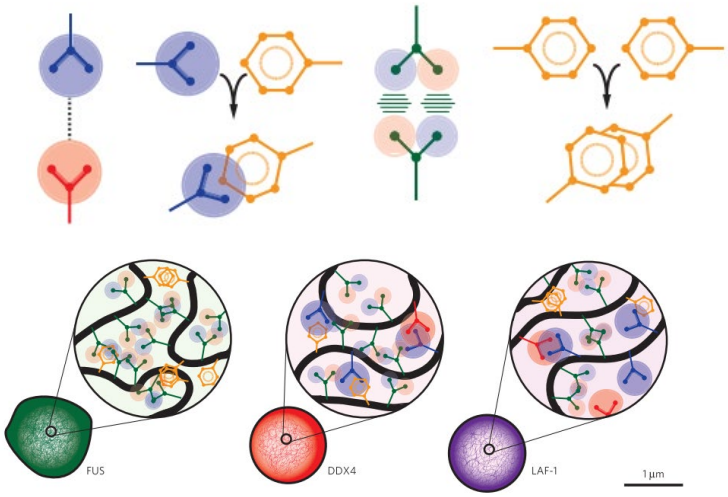



(B) リン脂質二分子膜

https://www.biophys.jp/highschool_old/B-09.html

脂質膜をもつ: 核, ミトコンドリア, ゴルジ体など
膜をもたない: 核小体, ストレス顆粒など

# アミノ酸配列が相分離を引き起こす



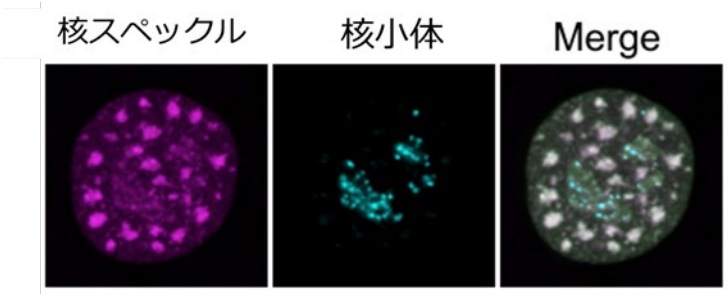

Brangwynne 2015

IDR(天然変性領域) 間の相互作用:

静電相互作用

カチオン-π相互作用

π-π相互作用

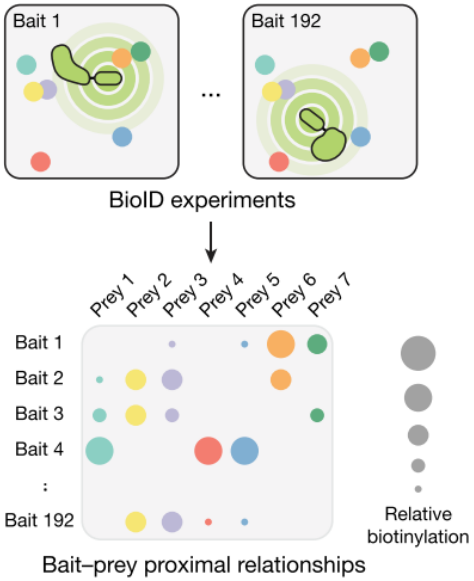



2種の液滴が核内に共存している

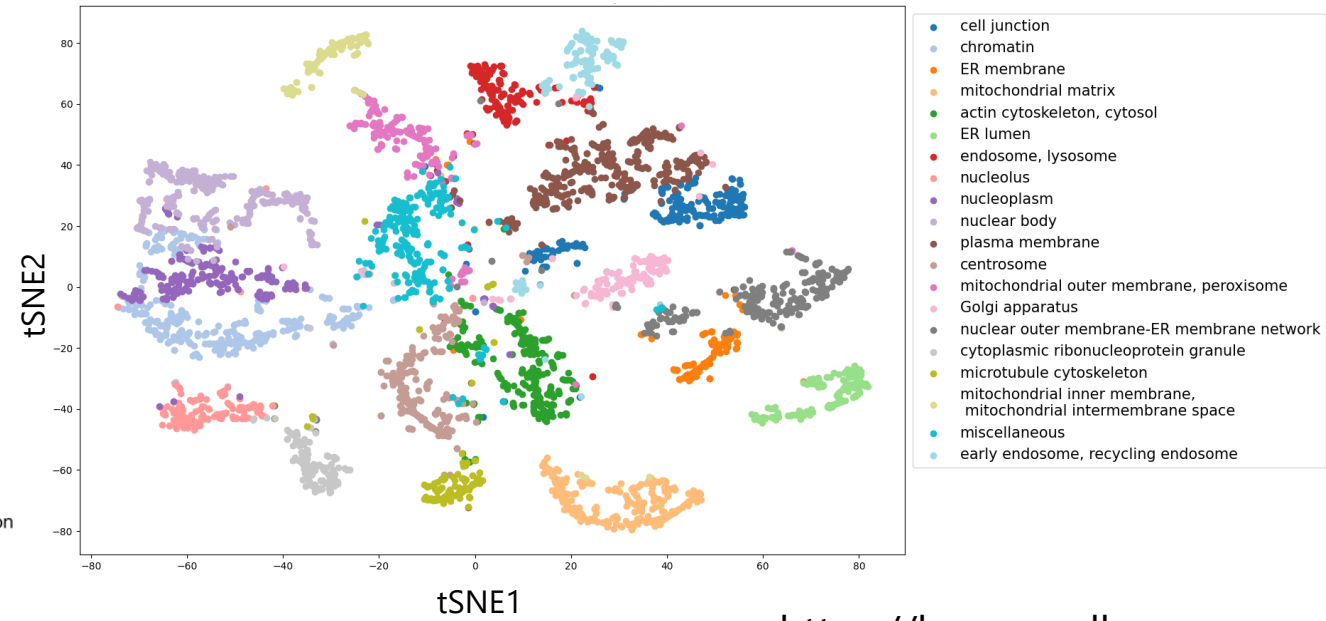
興味があること... 相分離に重要なタンパクの性質は何か？

複数の液滴が共存するメカニズムは？

# タンパク局在位置のデータベース




相互作用からタンパクをclustering



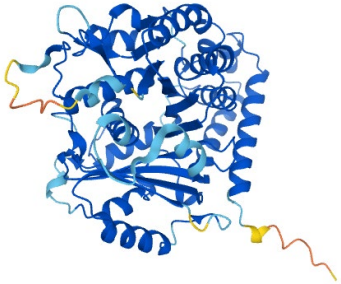

https://humancellmap.org

液滴のクラスタリングが存在する(核小体，核内構造体など)

→液滴にのみ見られるタンパクの特徴量があるか？

例えばIDRの割合など

# Transformerによってアミノ酸配列を数値ベクトルに変換



アミノ酸配列

MRECISIHVG QAGVQIGNAC...

アミノ酸配列

MKIVPPKPFF FEAGERAVLL...

学習済みのTransformer

Embedding

[0.8 0.6 0.5 0.4 ... ]

Embedding

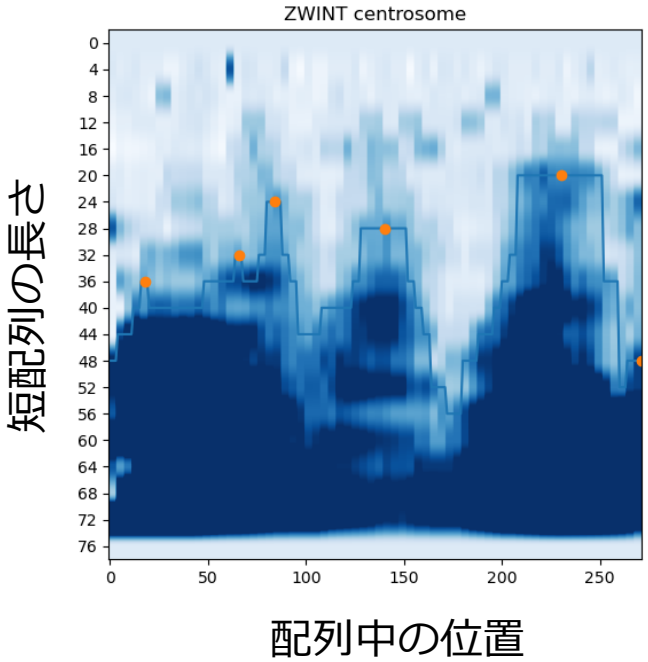
[0.7 0.4 0.1 0.3... ]

Classifierで局在予測

Nucleolus

Nuclear body

# 局在を決める部分配列を特定する




アミノ酸配列を短く切り分ける

→ 短配列の予測スコアをすべて計算

色が濃いほど予測のスコアが高い

=予測に重要な配列

画像の中で，頂点になっている点を特定する

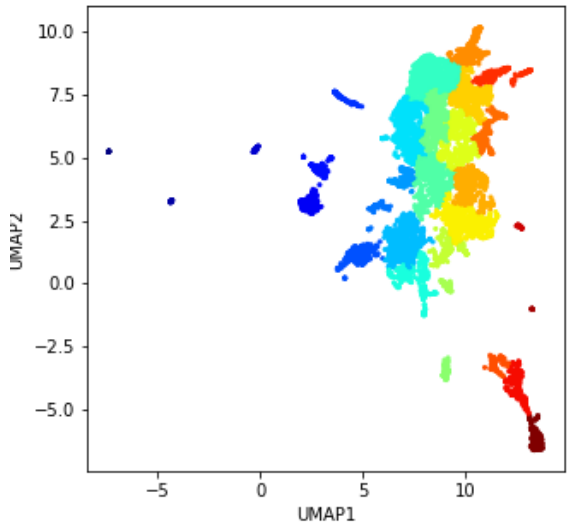
頂点に当たる配列を含むより長い配列はスコアが高くなっている

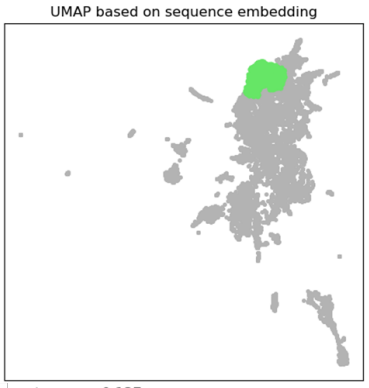
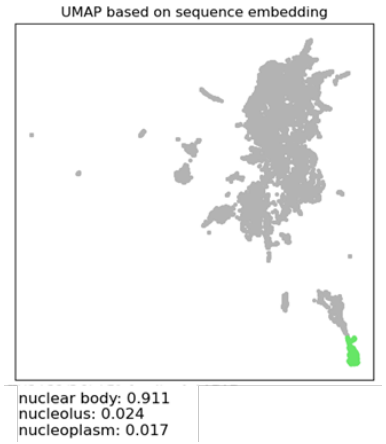
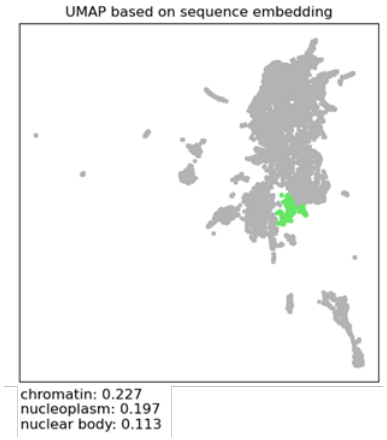
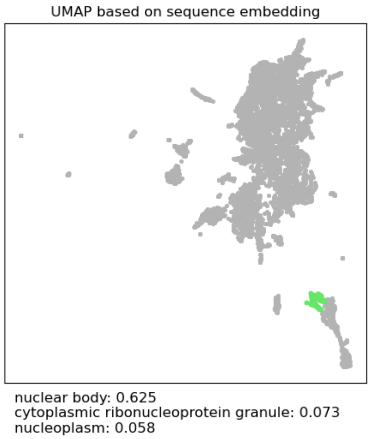
# 重要配列は30ほどのクラスターに分かれる

アミノ酸配列

[MRECISIHVG]
[AGVQIGNAC]
[AVPGGEVCC]
[MASTYGGGA]
⋮
[MKVVPPKPFF]

Embedding clustering




33のクラスターに分かれ,
クラスターごとに特定の
配列シグナルを持っている