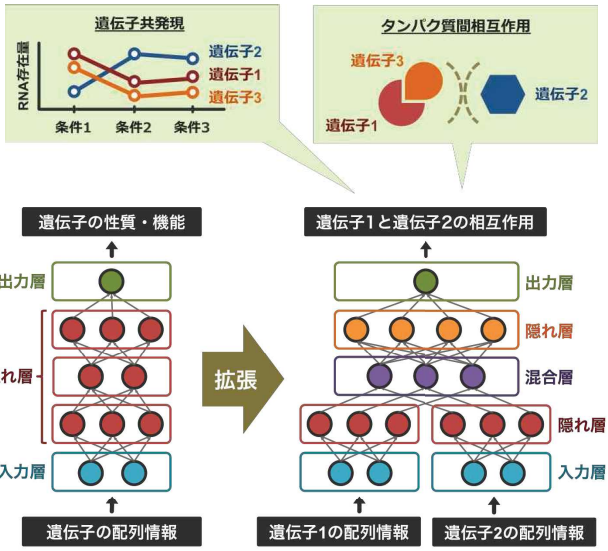
遺伝子相互作用の高精度モデリングに向けたペアワイズ深層学習モデルの開発

青木 裕一
（東北大学）

人工知能と連携して生命現象を解く

研究の目的・背景、研究開発内容

本研究では「2つの要素の間に潜在する関係をデータ駆動的に明らかにするための深層表現学習モデル」を開発する。事例として、生命システムの構成要素である遺伝子に着目し、遺伝子相互作用を生物学的配列情報から推定するための新規表現学習モデルを構築する。本課題において遺伝子相互作用とは、個々の遺伝子が他の遺伝子と連携することによって高度な生物学的機能を達成する分子生物学的現象を指し、RNA段階での量的な関連性である遺伝子共発現や、タンパク質段階での物理的な相互作用であるタンパク質間相互作用を研究対象とする。近年は、高処理能の分析機器や実験技術の発展に伴い塩基・アミノ酸配列や遺伝子相互作用に関する公共データが急速に蓄積されているため、これらを教師データとして活用した深層表現学習が実現可能な状況になっている。そこで本研究では、深層表現学習技術を駆使して「生物学的配列に記された遺伝子相互作用の生起メカニズムを解き明かす」という生命科学の重要課題に挑戦する。

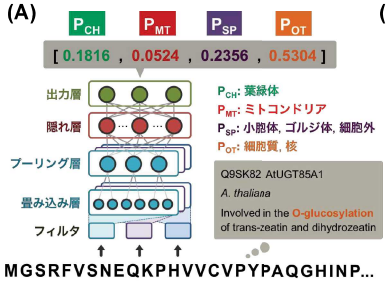


成果

- 畳み込みニューラルネットワークが学習で獲得した内部表現によって、タンパク質細胞内局在の制御メカニズムを説明しうることを実証した。
- 表現学習機能を導入したペアワイズ深層ニューラルネットワークによって、遺伝子相互作用の生起メカニズムに関する仮説の創出に成功した。

結果① ~畳み込みニューラルネットワークを用いた細胞内局在予測モデルの開発~

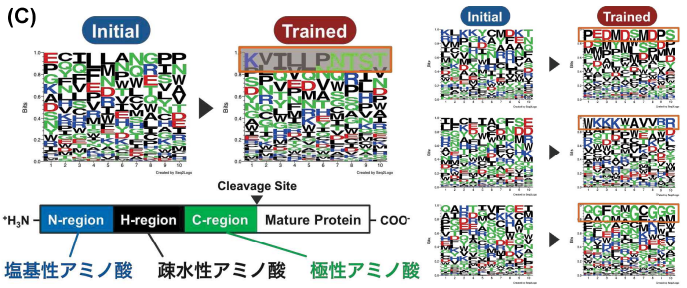
畳み込みニューラルネットワークを用いて、「任意のタンパク質のN末端側アミノ酸配列を入力すると、そのタンパク質が種々の細胞内局在を示す確率を出力する」予測モデルを構築した。入力のアミノ酸配列は One-hot Encoding により0/1行列に変換し、畳み込みフィルタを用いた特徴配列の抽出を行った（A）。モデルを学習するための教師データには、TargetP (http://www.cbs.dtu.dk/services/TargetP/) が提供するベンチマークデータセットを適用した。提案手法は、Non-Plantデータに対して92.4%、Plantデータに対して88.5%と、従来の予測手法 TargetLoc・TargetP・iPSORT を上回る予測精度を示した（B）。予測モデルが学習した畳み込みフィルタを解析した結果、NHC domainなどの既知シグナル配列と類似の配列モチーフに加えて、細胞内局在制御への関与が期待される特徴的なアミノ酸配列パターンが複数見出された（C）。



(B)

Dataset	Method	Category	Sensitivity	Specificity	Accuracy[%]
Non-Plant	TargetLoc	mi	0.91	0.77	82.5 (±1.2)
#Train: 2191 #Test: 547	SVM	SP	0.95	0.92	
		OT	0.91	0.97	
	TargetP	mi	0.89	0.67	80.0 (±0.7)
	NN	SP	0.96	0.92	
		OT	0.88	0.97	
	iPSORT	mi	0.74	0.68	88.5
	Rule	SP	0.92	0.92	
		OT	0.90	0.92	
	This Study	mi	0.79	0.99	83.2 (±1.5)
	CNN	SP	0.99	0.98	
		OT	0.97	0.92	

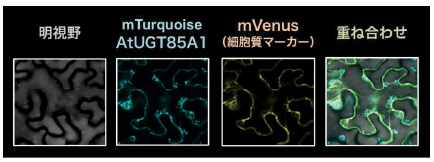
Learning Rate: 0.010, Batch Size: 100, L2 Regularization: 0.002
Filter: 100, # Hidden Unit: 100, Activation Function: ReLU



結果② ~変異導入シミュレーションによる遺伝子工学への応用~

モデルの予測結果を実験的に検証するために、教師データに含まれていないタンパク質AtUGT85A1を対象に、蛍光タンパク質を用いた細胞内局在の観察を行った。細胞質に局在する（P_{OT}が最大）と予測されたAtUGT85A1は、観察実験においても細胞質局在であることが確認された（A）。また、天然配列の各座位を各種アミノ酸に置換した変異配列について細胞内局在を予測する変異導入シミュレーション実験を行い、各アミノ酸置換が細胞内局在に及ぼす影響を定量的に評価する手法を開発した。同一座位における置換でもアミノ酸の物理化学的な性質によって細胞内局在への影響が異なる事例（B-1,2,3）や、複数の置換が相加的に寄与する事例（B-3,4）が見出された。この手法を用いることで、タンパク質細胞内局在の制御が最小限の変異導入によって実現可能となり、細胞内局在の改変による代謝制御など遺伝子工学的な応用が期待できる。

(A)



(B)

		P_{CH}	P_{MT}	P_{SP}	P_{OT}
	MGSRFVSNEQKPHVVCV...	0.182	0.052	0.236	**0.530**
(1)	MGSRFVSNAQKPHVVCV...	0.229	0.084	0.251	**0.436**
(2)	MGSRFVSNDQKPHVVCV...	0.166	0.057	0.247	**0.530**
(3)	MGSRFVSNRQKPHVVCV...	0.288	0.153	0.210	**0.349**
(4)	MGSRFVSNRRKPHVVCV...	**0.329**	0.282	0.156	0.234

A: 中性アミノ酸　E, D: 酸性アミノ酸　R: 塩基性アミノ酸

結果③ ~ペアワイズ深層学習モデルを用いたタンパク質間相互作用予測モデルの開発~

Human Protein Reference Database が提供するタンパク質間相互作用データセットを教師データ（正例・負例 各30000件）として、Conjoint Triad法によってアミノ酸配列から得られる343次元の特徴ベクトルを入力、タンパク質間相互作用の有無を出力とする右の分類モデルを学習した（A）。各特徴ベクトルと任意の重みベクトルの外積について相関係数を算出し、得られた重み付き相関係数を混合入力としてタンパク質相互作用の有無を予測した。学習後の重みフィルタを解析から、ジスルフィド結合の形成によってタンパク質間相互作用の発現に寄与するシステイン残基（Cys, Conjoint Triad Codeは7）を重視する事例などが見出されたことから、提案手法は相互作用に重要なアミノ酸配列特徴の組み合わせをデータ駆動的に明らかにする性能を有していることが示唆された。