CA-005

ゲノム編集と機械学習を融合したスマート育種技術の開発
Efficient crop breeding using genome editing and machine learning

小谷野 仁[1)]　黒羽 剛[2)]　吉田 均[2)]
Hitoshi Koyano　Takeshi Kuroha　Hitoshi Yoshida

1　はじめに

1.1　研究の背景

現在, 温暖化を始めとする気候変動により, 農耕地の劣化が進み, 不良耕作地が増大している. また, 世界各地で作物の干ばつ害, 高温障害, 塩害などが頻繁に発生し, 作物の生産量と品質に甚大な被害が及んでいる. 一方で, 食料生産を上回る速度で世界の人口は増加しており, 更に我々の食生活は肉食化が進んでいる. このため, 2050 年には, 人間の食用と家畜の飼料用を合わせ, 現在の約 1.6 倍の穀物が必要になると推計されている. 気候変動と人口増加の下で, 作物生産量の増加は, 食料確保のために世界が共有する課題となっており, 世界各国で農学分野の多くの研究者が, 作物の品種開発に取り組んでいる. しかし, 選抜と交雑を繰り返す育種は, 通常 10 年程度の長い時間を必要とする. このため, 現在, 作物の品種開発に要する時間の短縮化が強く求められている.

一方で, ゲノム編集の研究が 1990 年代後半に始まり, 2012 年に現在主流のツールである CRISPR/Cas9 が開発された. ゲノム編集は同年に *Nature Methods* 誌において Method of the Year 2011 に輝き, 2020 年に CRISPR/Cas9 の開発者がノーベル化学賞を受賞するなどして大変な注目を集め, 現在, 生命科学において必須の技術となりつつある.

近年, ゲノム編集を用いて様々な作物研究がなされており, ゲノム編集による品種開発の効率化が強く期待されている. しかし, 現在, 育種においてゲノム編集を用いて行われているのは「遺伝子の機能喪失による表現型の改良」であり, これは, 従来からの技術である遺伝子ノックアウトでも可能であったことである. 言い換えると, DNA 配列の指定した場所に挿入, 欠失, 及び置換を引き起こすことが出来るゲノム編集の長所が育種において活かされておらず, ゲノム編集を用いた育種は, 従来からの遺伝子の機能のオンオフの二分法のパラダイムの中に留まっている.

1.2　研究の目的

遺伝子の機能喪失を利用する育種では, より良い表現型が得られる機能喪失した遺伝子の組合せを探すことになり, 限られた部品の組合せで表現型を改良することになる. 遺伝子の発現量を調節して新たに表現型を創出することが出来るならば, 品種の可能性は一気に広がる. そこで, 目的の形質に関与する遺伝子の発現調節領域にゲノム編集を行って, 様々なパターンの挿入, 欠失, 及び置換からなる変異を導入し, その遺伝子の発現量を精密に調節して, 望む表現型を作り出すという育種の方法論を考える. 言い換えると, ゲノム編集を用いて, 育種の方法論を「遺伝子の機能喪失による表現型の改良」から「遺伝子の発現量の精密調節による最適な表現型の創出」に進めることが出来ないかを考える.

しかし, 遺伝子の発現調節領域の配列に網羅的なゲノム編集を行って, 大量のゲノム編集個体を作出し, 表現型を調節していくのは, 費用や時間の点で極めて困難である. 結局のところ,「所望の表現型を実現するには、その形質に関与する遺伝子の発現調節領域のどこをどう編集すれば良いのか分からない」ことが, 現在, 育種においてゲノム編集がそのポテンシャルを発揮することが出来ない最大の理由となっている.

そこで, 本研究では, 遺伝子の発現調節領域の幾つかの変異配列とそれらの下での表現型の組のデータに基づいて学習し, 新規の変異配列の下での表現型を予測する方法の開発に取り組む. これにより, 網羅的なゲノム編集実験をすることなく, 出来るだけ少ない実験で発現調節領域の理想的な変異配列を同定することを可能にすることにより, 上記のボトルネックを解消することを目指す.

ガイド RNA の設計から DNA への変異の導入, ゲノム編集個体の栽培, 後代の個体の栽培を経て, それらの表現型の評価までには膨大な時間と労力が掛かるため, ゲノム編集を用いて効率的な品種開発を実現したいという場合に, 学習機械が学習に用いることが出来る変異配列と表現型のデータの数は 10 個程度であり, ゲノム編集育種の現場でビッグデータが利用可能になる見込みは将来的にもない. 発現調節領域が 100 から 200 塩基からなるとすると, 探索空間の大きさは 4^{100} から 4^{200} である. このように, 非常に少数のデータからそれらが含む情報を出来るだけ多く抽出し, 上記の大きさの探索空間の中で高精度の予測を行う必要があることが, 他の多くの予測問題と比較した時の本研究における予測問題の特徴となっている.

2　利用したデータセット

本節では, 次節における本研究の提案方法の提示に先立って, 本研究で用いたデータについて説明する. 提案方法の学習と予測精度の検証には, イネの *TAW1* 遺伝子と *SD1* 遺伝子の発現調節領域の野生型配列と変異配列, 及びそれらの下での表現型のデータを用いた. データに関する詳細な情報は Kuroha *et al.* [1] に述べられている. *TAW1* と *SD1* は, それぞれ, イネの着粒数と草丈を制御する遺伝子であり, ここでは, 表現型として 1 次枝梗数当たりの 2 次枝梗数と草丈に着目した.

TAW1 遺伝子を用いて, 利用したデータについてもう少し説明する. *TAW1* はイネの第 10 染色体の 17,888,297 から 17,889,724 に位置し, この研究では, その上流にある 113 塩基からなる発現調節領域の野生型配列とゲノム編集を行って得た 26 個の変異配列, 及びこれらの下での表

1)　農業・食品産業技術総合研究機構, 農業情報研究センター Research Center for Agricultural Information Technology, National Agriculture and Food Research Organization
2)　農業・食品産業技術総合研究機構, 生物機能利用研究部門 Institute of Agrobiological Sciences, National Agriculture and Food Research Organization

現型のデータを用いた. この領域はサイレンサーであって, *TAW1* の発現を抑制する働きをすることが知られている. 表 1 に, 本研究において用いられた配列データの例として, この発現調節領域の野生型配列と 4 つの変異配列を示す. この表において, () は挿入を, – は欠失を示している. 従って, 変異配列 1 は, ゲノム編集により, 31 番目と 32 番目の塩基の間に 2 つの塩基 TT が挿入された変異配列である. また, 変異配列 2 は, 83 番目から 87 番目の塩基 CATAA が欠失された変異配列である. 同様に, 変異配列 3 と 4 は, それぞれ 79 番目から 100 番目の塩基 ATACCATAAGTAGCTAGGCTTG と 7 番目から 87 番目の塩基 CTGTGGCTGTGTGCCTGCTTCCGCCCTTTCAGGGGCGCTTTTGCTTTGCTTTTTATGGTACCTGTACTGCTCATACCATAA が欠失された変異配列である.

表 1　TAW1 の発現調節領域の野生型配列と変異配列の例.

> 野生型配列 (表 2 における WT) AGCTAGCTGTGGCTGTGTGCCTGCTTCCGCCCTTTCAGGG GCGCTTTTGCTTTGCTTTTTATGGTACCTGTACTGCTCAT ACCATAAGTAGCTAGGCTTGCTTGGCCACCTCT
> 変異配列 1 (表 2 における +TT) AGCTAGCTGTGGCTGTGTGCCTGCTTCCGCC(TT)CTTTC AGGGGCGCTTTTGCTTTGCTTTTTATGGTACCTGTACTGC TCATACCATAAGTAGCTAGGCTTGCTTGGCCACCTCT
> 変異配列 2 (表 2 における ΔCATAA) AGCTAGCTGTGGCTGTGTGCCTGCTTCCGCCCTTTCAGGG GCGCTTTTGCTTTGCTTTTTATGGTACCTGTACTGCTCAT AC-----GTAGCTAGGCTTGCTTGGCCACCTCT
> 変異配列 3 (表 2 における Δ22) AGCTAGCTGTGGCTGTGTGCCTGCTTCCGCCCTTTCAGGG GCGCTTTTGCTTTGCTTTTTATGGTACCTGTACTGCTC-- ----------CTTGGCCACCTCT
> 変異配列 4 (表 2 における Δ81) AGCTAG---------- ---------- -------GTAGCTAGGCTTGCTTGGCCACCTCT

TAW1 の発現調節領域の野生型配列と変異配列に対する表現型のデータが表 2 に示されている. 野生型配列 WT と変異配列 Δ7, Δ77, 及び Δ1k に対しては, 表現型の評価を 11 回以上行ったが, 表 2 では割愛した. 表 2 (続き 1 と 2) の第 5 列から第 7 列に示されている表現型の測定値の平均, 標準偏差, 及び変動係数は, 割愛したデータも含めて計算された数値である. この表から, 表現型の測定値はかなりばらつくことが分かる.

図 1 に *TAW1* の発現調節領域のゲノム編集による 26 個の編集箇所が示されている. 編集領域の付近に示された文字列は, その編集を行って得た変異配列の ID で, これらは図 2 の第 1 列に示された配列 ID に対応している. また, 編集領域の付近に示された数値は, その編集を行って得た変異配列の下での表現型で, これらは表 2 (続き 1 と 2) の第 5 列に示された表現型の測定値の平均に対応している.

3　提案方法

本節において, 本研究において開発した, 変異配列の下での表現型を予測するための方法を述べる.

表 2　野生型配列と変異配列に対する表現型の測定値. 表頭の 1 から 7 は表現型の測定の回数を表す.

配列 ID	1	2	3	4	5	6	7
+TT	0.25	0.38	0.57	0.80	0.67	0.63	
+T	0.29	1.00	0.63	0.38	0.86	0.40	1.11
+A	0.88	0.75	0.17	0.67			
WT	0.50	1.38	0.40	0.57	0.60	0.63	0.40
ΔGT	1.67	1.38	1.43	1.25	1.30		
ΔTT	0.60	0.67	0.88	0.40	0.40	0.29	0.60
ΔTA	1.50	1.13	1.86	1.63			
ΔCT	0.63	0.71	0.88	1.00	0.33		
ΔTAG	1.50						
ΔAGTA	1.57	1.50	2.00	1.33	1.88	1.78	1.57
ΔAGCT	0.50	0.29	0.71	0.20	0.89	0.86	
ΔTTCC	0.43	0.78	0.20	0.80	1.00	1.11	0.60
ΔCATAA	1.38	1.38	1.40	1.33	1.89	1.25	1.80
ΔCTGTG	0.50	0.80	0.60	0.88	1.10	0.40	1.33
ΔTGTGT	0.56						
ΔGCTTT	1.25	0.56	1.56	0.63			
Δ7	0.00	0.60	1.00	0.40	0.75	0.67	0.67
Δ9	1.00	1.56	0.71	0.78	1.33		
Δ10	1.45	1.33					
Δ14	1.29	0.88	1.00	2.00	1.00	1.60	
Δ22	1.40						
Δ26	0.50	0.73	0.86	0.67	0.67	0.75	0.67
Δ28	1.40	1.70	2.22	1.00	1.43	1.63	1.20
Δ37	1.00	0.78	1.40	0.88	0.50	0.71	1.78
Δ77	4.86	3.71	4.11	3.29	3.71	3.57	3.56
Δ81	2.00	2.83	2.50				
Δ1k	2.20	2.71	3.57	3.67	3.17	2.00	3.60

表 2　続き 1. 表頭の 8 から 10 は表現型の測定の回数を, M, SD, 及び VC はそれぞれ表現型の測定値の平均, 標準偏差, 及び変動係数を表す.

配列 ID	8	9	10	M	SD	VC
+TT				0.5480	0.2014	0.3675
+T				0.6649	0.3286	0.4943
+A				0.6146	0.3106	0.5055
WT	0.71	0.50	0.63	0.6063	0.2627	0.4332
ΔGT				1.4040	0.1620	0.1154
ΔTT	0.57			0.5499	0.1850	0.3364
ΔTA				1.5268	0.3060	0.2004
ΔCT				0.7095	0.2552	0.3596
ΔTAG				1.5000	NA	NA
ΔAGTA	1.00	1.44		1.5637	0.3005	0.1922
ΔAGCT				0.5743	0.2924	0.5091
ΔTTCC	1.00			0.7397	0.3131	0.4233
ΔCATAA	2.11	1.67	1.80	1.6000	0.2914	0.1821
ΔCTGTG				0.8012	0.3342	0.4171
ΔTGTGT				0.5600	NA	NA
ΔGCTTT				0.9965	0.4862	0.4879
Δ7	0.60	0.29	0.50	0.6083	0.3296	0.5419
Δ9				1.0762	0.3612	0.3356
Δ10				1.3939	0.0857	0.0615
Δ14				1.2935	0.4337	0.3353
Δ22				1.4000	NA	NA

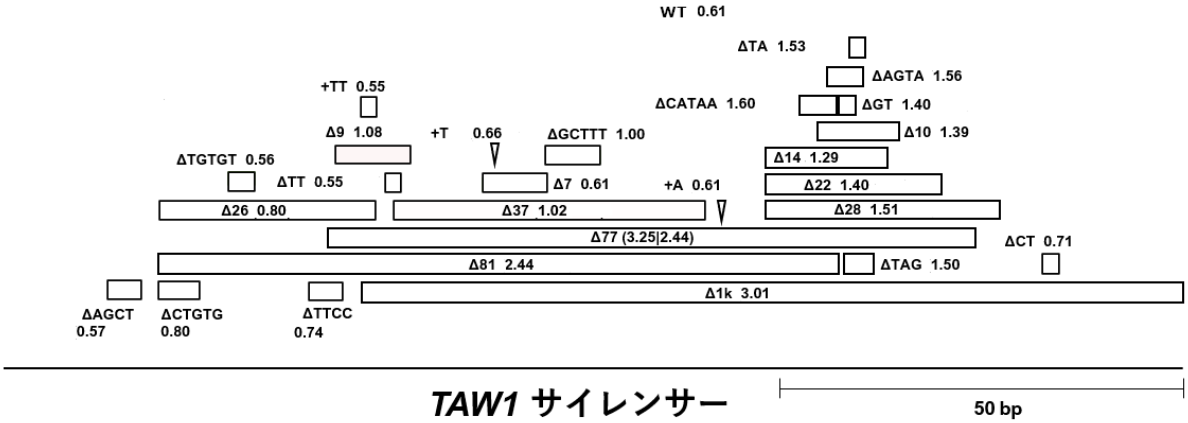

図 1　*TAW1* の発現調節領域のゲノム編集による編集箇所. 長方形が欠失させた領域を, 三角形が挿入した領域を示す.

表 2　続き 2.

配列 ID	8	9	10	M	SD	VC
Δ26	1.20	1.13		0.7955	0.2291	0.2880
Δ28				1.5108	0.3940	0.2608
Δ37	1.10			1.0181	0.4086	0.4014
Δ77	3.40	2.63	3.22	3.2534	0.8420	0.2588
Δ81				2.4444	0.4194	0.1716
Δ1k	2.86	3.00	3.25	3.0067	0.5473	0.1820

3.1　基本原理

学習機械は次の P_1 と P_2 を基本原理として動作するものとし, 第 3.2 節から第 3.5 節において, ある条件が満たされている時にのみ発動する推論規則を追加する. 発現調節領域がサイレンサーである場合を考える. エンハンサーである場合, 以下の不等号が逆になる. エンハンサーかサイレンサーかは, 発現調節領域に少なくとも 1 つのゲノム編集を施して, 転写量 (または表現型) を測定する必要があるが, この情報は既知として話を進める. 配列 s_1 が s_2 に部分列として含まれる時, $s_1 \subset s_2$ と書く. s_1 と s_2 の間の Levenshtein 距離を $d_L(s_1, s_2)$ と書く. 配列 s の下での表現型を $\pi(s)$ を書く. 発現調節領域の野生型配列を s_0 と書く.

$$\mathrm{P}_1 : s_1 \subset s_2 \Rightarrow \pi(s_1) \leq \pi(s_2),$$
$$\mathrm{P}_2 : d_L(s_1, s_0) > d_L(s_2, s_0) \Rightarrow \pi(s_1) \leq \pi(s_2).$$

但し, P_1 は P_2 に優先するものとする.

P_1 と P_2 の下で, ゲノム編集により得られた新規の配列 s が入力された時, 学習機械は次の手順で $\pi(s)$ の予測値 $\hat{\pi}(s)$ を計算する.

ステップ 1. s が含む配列のリスト L_1 を作成する. $L_1 = \emptyset$ の場合があることに注意する.

ステップ 2. s を含む配列のリスト L_2 を作成する.

ステップ 3. L_1 においてゲノム編集の効果が最小である配列

$$s_1^{(\min)} = \arg\min_{s \in L_1} \pi(s)$$

を求める. $s_1^{(\min)}$ は存在しないことがあることに注意する.

ステップ 4. L_2 においてゲノム編集の効果が最大である配列

$$s_2^{(\max)} = \arg\max_{s \in L_2} \pi(s)$$

を求める.

ステップ 5. $s_1^{(\min)}$ が存在するならば,

$$\hat{\pi}(s) = \pi(s_2^{(\max)}) + \left\{\pi(s_1^{(\min)}) - \pi(s_2^{(\max)})\right\} \times \frac{d_L(s, s_1^{(\max)})}{d_L(s_2^{(\max)}, s_1^{(\min)})} \tag{1}$$

によって $\pi(s)$ を予測する.

ステップ 6. $s_1^{(\min)}$ は存在せず, $s_2^{(\max)} \neq s_0$ となる $s_2^{(\max)}$ は存在するならば,

$$\hat{\pi}(s) = \pi(s_0) + \left\{\pi(s_2^{(\max)}) - \pi(s_0)\right\} \frac{d_L(s, s_0)}{d_L(s_2^{(\max)}, s_0)}$$

によって $\pi(s)$ を予測する.

ステップ 7. $s_1^{(\min)}$ は存在せず, $s^* = \arg\min_t d_L(s, t)$ に対して $s^* \neq s_0$ ならば,

$$\hat{\pi}(s) = \pi(s_0) + \left\{\pi(s^*) - \pi(s_0)\right\} \frac{d_L(s, s_0)}{d_L(s^*, s_0)}$$

によって $\pi(s)$ を予測する.

ステップ 8. $s_1^{(\min)}$ は存在せず, $s^* = \arg\min_t d_L(s, t)$ に対して $s^* = s_0$ ならば,

$$\hat{\pi}(s) = \pi(s)$$

によって $\pi(s)$ を予測する.

3.2　ヒューリスティクス 1: 編集有効塩基保存度

第 2 節の図 1 に示されているように, ゲノム編集によって同じ長さの変異を導入した場合でも, 発現調節領域のどの部位を編集したかによって, 表現型の変化分は異なる. Levenshtein 距離では, 編集する 2 つの文字列中の各文字に対して編集費用という量が定義されている. まず, この量を用いて, 上述の問題に対応するように, 第 3.1 節で述べた予測方法を修正する.

編集費用としては, 表 3 に示されている発現調節領域の各塩基の保存度を用いる. この表に示されている塩基保存度は, 類縁種として単子葉類を取って計算された値である. 塩基保存度は類縁種の取り方により変化し, 絶対的な値ではない. 本研究では, 類縁種の取り方を色々と変えて, 予測誤差を最小化する塩基保存度を探すとい

うことはせずに, 塩基保存度はあくまで初期値として用い, 予測誤差を最小化するように書き換えることにする.

図 1 と表 3 から, 塩基保存度の低い塩基の編集は表現型にほとんど影響を与えないように見える. そこで, 具体的には, 編集有効塩基保存度と呼ばれる未知の定数 $c_E \geq 0$ を導入し, 定めた c_E の値以下の保存度を 0 に書き換える. そうすると, 例えば, 2 つの配列 s_1 と s_2 の間の Levenshtein 距離 $d_L(s_1, s_2)$ の計算において, $\cdots + 0.2 \times 1 + 0.4 \times 1 + 0.6 \times 1 + \cdots$ という項が現れ, かつ $c_E = 0.6$ の時, これらの項は $\cdots + 0 \times 1 + 0 \times 1 + 0 \times 1 + \cdots$ となって, $d_L(s_1, s_2)$ に寄与しなくなる. c_E の値を変化させて, 予測誤差を評価することを繰り返し, 予測誤差を最小化する c_E の値を探すステップを, 第 3.1 節の予測方法に組み込む.

表 3　*TAW1* の 113 塩基からなる発現調節領域の, 単子葉類を類縁種にとって計算した塩基保存度. これらの値を Levenshtein 距離の編集費用の初期値として用いた.

塩基	1	2	3	4	5	6	7	8	9	10
保存度	0	0	0	0	0	0	0	0	0	0
11	12	13	14	15	16	17	18	19	20	21	22
0	0	0	0	0	0	0	0	0	0	0	0
23	24	25	26	27	28	29	30	31	32	33	34
0	0	0	0	0	0.2	0.2	0.2	0.2	0.2	0.2	0.4
35	36	37	38	39	40	41	42	43	44	45	46
0.4	0.2	0.2	0.2	0.2	0.2	0.4	0.6	0.8	0.8	1.0	1.0
47	48	49	50	51	52	53	54	55	56	57	58
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
59	60	61	62	63	64	65	66	67	68	69	70
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
71	72	73	74	75	76	77	78	79	80	81	82
1.0	0.8	0.8	0.8	0.6	0.8	0.8	0.8	1.0	1.0	1.0	1.0
83	84	85	86	87	88	89	90	90	91	92	93
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.8
94	95	96	97	98	99	100	101	102	103	104	105
0.8	0.8	0.6	0.6	0.4	0.2	0.2	0	0	0	0	0
106	107	108	109	110	111	112	113				
0	0	0	0	0	0	0	0				

3.3　ヒューリスティクス 2: 編集効果の飽和

まず, 第 2 節で述べたデータを用いて, 動機を説明する. 表現型のデータは, 小数点以下 1 桁までの値を用いる. $\pi(\Delta\mathrm{TA})$ を予測したいとする ($\pi(\Delta\mathrm{TA}) = 1.5$). ΔTA が含む配列は, ΔAGTA, Δ10, Δ22, Δ28, Δ77, 及び Δ1k の 6 個存在し, これらのうちゲノム編集の効果が最小であるものは Δ22 で, $\pi(\Delta 22) = 1.4$ である. また, ΔTA を含む配列は野生型配列のみで, $\pi(\mathrm{WT}) = 0.6$ である. よって, P_1 から, $\hat{\pi}(\Delta\mathrm{TA})$ の上界と下界は次のように定められる.

$$0.6 = \pi(\mathrm{WT}) \leq \hat{\pi}(\Delta\mathrm{TA}) \leq \pi(\Delta 22) = 1.4$$

更に, $d_L(\mathrm{WT}, \Delta\mathrm{TA}) \ll d_L(\Delta\mathrm{TA}, \Delta 22)$ であるから, 学習機械は, 式 (1) より, $\pi(\Delta\mathrm{TA})$ を 1.4 より 0.6 にずっと近い値として予測する (つまり, 大幅に過小推定する). 図 1 より, Δ28 において欠失された領域では 28 塩基の欠失により表現型が 0.6 から 1.5 に増加するが, $\pi(\Delta\mathrm{TA}) = 1.5$ と $\pi(\Delta\mathrm{GT}) = 1.4$ より, 2 塩基の欠失で表現型が増加する効果は上げ止まることが分かる.

この現象に対応するために, 編集効果の飽和という考え方を導入する. 具体的には学習機械に次の処理をさせる.

ステップ 1. 新規の配列 s が入力された時, 第 3.1 節においてと同様に, まず s が含む配列のリスト L_1 と s を含む配列のリスト L_2 を作成する.

ステップ 2. 各配列 $t \in L_1$ に対して, t を含み, かつ s の長さ以上の長さの配列のリストを作成し, 得られたリストを合併する.

ステップ 3. このリストの中でゲノム編集の効果が最大である配列 (s^* によって表す) を探す.

ステップ 4. そうして, $\pi(s^*)$ を $\hat{\pi}(s)$ の下界とし, 式 (1) を修正して得られる予測量

$$\hat{\pi}(s) = \pi(s_1^{(\min)}) - \{\pi(s_1^{(\min)}) - \pi(s^*)\}\frac{d_L(s, s_1^{(\min)})}{d_L(s^*, s_1^{(\min)})} \quad (2)$$

によって $\pi(s)$ を予測する.

この編集効果の飽和という考え方を用いると, $\hat{\pi}(\Delta\mathrm{TA})$ の上界と下界は

$$1.4 = \pi(\Delta\mathrm{GT}) \leq \hat{\pi}(\Delta\mathrm{TA}) \leq \pi(\Delta 22) = 1.4$$

と定まり, $\pi(s)$ の予測値は $\hat{\pi}(s) = 1.4$ となる.

3.4　ヒューリスティクス 3: 予測値の上界と下界の信頼度の評価と予測量の修正

第 3.1 節の式 (1) によって与えられる予測量は, (A) 表現型の予測値の上界と下界を設定し, (B) (i) 上界 (または下界) を与える変異配列と表現型を予測したい変異配列の間の塩基保存度を重みとする Levenshtein 距離と (ii) 上界と下界を与える変異配列の間の Levenshtein 距離を求め, (iii) それらの比を上界と下界の差に掛けることにより, 予測対象の変異配列の下での表現型が下界の表現型からどれだけ増加したかを見積もる. このため, 予測値の精度は上界と下界の信頼度に依存し, 信頼できる上界と下界を設定できない時には, 例えばカーネル法のような既存の方法からの予測値を用いた方が良いということが起こり得る. そこで, 上界と下界の信頼度を評価する方法を導入し, これらを用いて, 本節で述べてきた方法の予測値ではなく, 既存の方法の予測値を用いる条件を学習させることを考える.

ステップ 1. f_θ をパラメーター θ を持つ既存の予測モデルまたは予測方法とする. そうして, 配列と表現型のデータに変換を施し,

$$\frac{\pi(s_i) - \pi(s_0)}{\max\limits_{1 \leq j \leq n}(\pi(s_j) - \pi(s_0))} = f_\theta\left(\frac{d_L(s_i, s_0)}{\max\limits_{1 \leq j \leq n} d_L(s_j, s_0)}\right), i = 1, \cdots, n \quad (3)$$

という形で f_θ を適用する. k 番目の配列の下での表現型を予測したい場合には, s_k と $\pi(s_k)$ を与えずに, パラメーター θ の値を決定し, その後, s_k を与えて, $(\pi(s_k) - \pi(s_0))/\max_{1 \leq j \leq n}(\pi(s_j) - \pi(s_0))$ を求め, この値から $\pi(s_k)$ の予測値を $\hat{\pi}(s_k)$ を計算する.

ステップ 2. 定数 $U_1, U_2, V_1, V_2 \geq 0$ に対して変異配列 s の表現型の予測に f_θ を用いるための条件 C を次のように定式化する.

$$\begin{aligned}\mathrm{C}: &[d_L(s, s_1^{(\min)}) > U_1 \wedge d_L(s, s_2^{(\max)}) > V_1] \\ &\vee [d_L(s, s_1^{(\min)}) > U_2] \vee [d_L(s, s_2^{(\max)}) > V_2]\end{aligned}$$

そうして, k 番目の配列の下での表現型を予測したいならば, s_k と $\pi(s_k)$ を与えずに, 予測誤差を最小化する U_1, U_2, V_1, 及び V_2 の値を求める.

本研究では, f_θ として多項式カーネルと組み合わせたサポートベクター回帰 (以下 SVR と略記) と多項式回帰モデルを用いた. 上界と下界を信頼するかどうかを分ける定数 U_1, U_2, V_1, 及び V_2 を決定する公式を得るのは難しいため, これらは $[0, \infty)^4$ の中のグリッドサーチで求める.

3.5 ヒューリスティクス 4: 他の遺伝子からのデータを用いた予測精度の強化

今, 幾つかの遺伝子の発現調節領域にゲノム編集を行って, 変異配列を作成し, それらの下での表現型を測定して, 配列と表現型の組のデータを保持しているとする. この時, 新規の遺伝子の発現調節領域の変異配列の下での表現型の予測に, 保持するデータを利用することが出来ないだろうかという考えが出て来る. 表現型のスケールは様々であるから, 式 (3) の左辺のように, 表現型のデータを相対化することが少なくとも必要である. しかし, 遺伝子の機能や転写制御の様式も様々であるから, 式 (3) のように, 表現型と配列のデータに相対化変換を施して, 新規の遺伝子の予測に, 保持する他の遺伝子からのデータを用いても, 多くの場合, 上手くはいかない. 発現調節領域にどれだけの長さの変異を導入したかと, それによって遺伝子の転写量や表現型がどれだけ大きく変化するかの対応は, 遺伝子によって異なるからである.

しかし, 新規の遺伝子の発現調節領域のある変異配列に対しては, 他の遺伝子からのデータを利用することを試みる価値がある. $s_{1,0}$ を遺伝子 1 の発現調節領域の野生型配列, $s_{1,1}, \cdots, s_{1,n_1}$ を変異配列とし, $s_{2,0}$ を遺伝子 2 の発現調節領域の野生型配列, $s_{2,1}, \cdots, s_{2,n_2}$ を変異配列とする.

$$D_L(s_{i,j}) = \frac{d_L(s_{i,j}, s_{i,0})}{\max\limits_{1 \leq k \leq n_i} d_L(s_{i,k}, s_{i,0})}, i = 1, 2, j = 0, \cdots, n_i$$

とおく. 今, s_{2,n_2} の下での表現型 $\pi(s_{2,n_2})$ を予測したいとする. この時, 上述の場合というのは, 定数 $\epsilon_1, \epsilon_2 > 0$ に対して,

$$D_L(s_{1,j}) \in \left(D_L(s_{2,n_2}) - \epsilon_1, D_L(s_{2,n_2}) + \epsilon_1\right), \quad \exists j \in \{1, \cdots, n_1\} \tag{4}$$

$$D_L(s_{2,k}) \notin \left(D_L(s_{2,n_2}) - \epsilon_2, D_L(s_{2,n_2}) + \epsilon_2\right), \quad \forall k \in \{1, \cdots, n_2 - 1\} \tag{5}$$

が成り立つ場合である.

そこで,

$$\Pi(s_{i,j}) = \frac{\pi(s_{i,j}) - \pi(s_{i,0})}{\max\limits_{1 \leq j \leq n_i} (\pi(s_{i,j}) - \pi(s_{i,0}))}, i = 1, 2, j = 0, \cdots, n_i$$

とおき, 次の手順を導入する.

ステップ 1. 第 3.4 節においてと同様に, f_θ をパラメーター θ を持つ既存の予測モデルまたは予測方法とする. そうして,

$$\Pi(s_{i,j}) = f_\theta\left(D_L(s_{i,j})\right), i = 1, 2, j = 1, \cdots, n_i \tag{6}$$

という形で f_θ を適用する.

ステップ 2. 予測誤差を最小化する式 (4) と (5) のパラメーター ϵ_1 と ϵ_2 の値を求める.

4 イネゲノム編集への応用

本節では, 第 3 節で述べた予測方法を用いて, 第 2 節で述べたデータから構築した学習機械の予測精度を検討する.

その前に, 既存の方法からの予測結果を検討しておく. 機械学習の文脈で見て, データの数が極端に少なく, 本予測問題においては, 深層学習など多数のパラメーターを持つ方法は使えない. 既存の学習機械の中では, スペクトルカーネルと組み合わせた SVR の予測精度が最も高かった. *TAW1* と *SD1* のデータセットを用いて, 1 点除外法でスペクトルカーネルと組み合わせた SVR の予測精度を評価した結果が, それぞれ表 4 と 5 に示されている. これらの表の右下に示されているように, *TAW1* の予測における平均相対誤差は 0.3364 であって, *SD1* の予測におけるそれは 0.1926 であった.

表 4 *TAW1* に対するスペクトルカーネルと組み合わせた SVR の予測結果.

配列 ID	測定値	予測値	相対誤差
+TT	0.5480	0.8111	0.4801
+T	0.6649	0.8348	0.2555
+A	0.6146	0.8255	0.3431
WT	0.5423	0.8285	0.5277
ΔGT	1.4040	0.9136	0.3493
ΔTT	0.5499	0.8660	0.5747
ΔTA	1.5268	0.8596	0.4370
ΔCT	0.7095	0.8214	0.1578
ΔTAG	1.5000	0.8762	0.4159
ΔAGTA	1.5637	0.8955	0.4273
ΔAGCT	0.5743	0.8939	0.5565
ΔTTCC	0.7397	0.8667	0.1718
ΔCATAA	1.6000	0.9310	0.4181
ΔCTGTG	0.8012	0.8655	0.0803
ΔTGTGT	0.5556	0.8566	0.5418
ΔGCTTT	0.9965	0.8705	0.1265
Δ7	0.6691	0.9255	0.3832
Δ9	1.0762	0.8658	0.1955
Δ10	1.3939	1.0811	0.2244
Δ14	1.2935	1.2159	0.0600
Δ22	1.4000	1.3450	0.0393
Δ26	0.7955	1.2049	0.5146
Δ28	1.5108	1.3556	0.1027
Δ37	1.0181	1.5705	0.5426
Δ77	3.2924	1.8961	0.4241
Δ81	2.4444	1.7450	0.2861
Δ1k	3.0067	1.6670	0.4456
	平均相対誤差		0.3364

表 5　*SD1* に対するスペクトルカーネルと組み合わせた SVR の予測結果.

配列 ID	測定値	予測値	相対誤差
WT	48.9091	38.2774	0.2174
Δ6	44.5000	38.2791	0.1398
Δ7	38.2500	42.3877	0.1082
Δ10	34.6000	42.3816	0.2249
Δ14	45.3714	38.2516	0.1569
Δ18	42.3800	38.2423	0.0976
Δ40	37.8500	42.3219	0.1181
Δ52	28.6250	42.3096	0.4781
	平均相対誤差		0.1926

次に, *TAW1* と *SD1* のデータセットを用いて, 1 点除外法で, 第 3 節において述べた予測方法の予測精度を評価した結果が, それぞれ表 6 と 7 に示されている. 第 3 節において述べた予測方法の, *TAW1* と *SD1* の予測における平均相対誤差はそれぞれ 0.1287 と 0.0461 であった (表 6 と 7 の右下の数値を参照). よって, 本研究において開発した予測方法は, スペクトルカーネルと組み合わせた SVR と比較して, 予測誤差を *TAW1* のデータセットにおいては約 0.3826 倍に, *SD1* のデータセットにおいては約 0.2394 倍に低下させたことが分かる.

表 6　*TAW1* に対する本研究の提案方法の予測結果.

配列 ID	測定値	予測値	相対誤差
+TT	0.5480	0.5997	0.0944
+T	0.6649	0.5677	0.1461
+A	0.6146	0.5997	0.0242
WT	0.5423	0.5997	0.1059
ΔGT	1.4040	1.2086	0.1391
ΔTT	0.5499	0.6546	0.1904
ΔTA	1.5268	1.2086	0.2084
ΔCT	0.7095	0.5997	0.1547
ΔTAG	1.5000	1.4267	0.0489
ΔAGTA	1.5637	1.4345	0.0826
ΔAGCT	0.5743	0.5423	0.0557
ΔTTCC	0.7397	0.5898	0.2027
ΔCATAA	1.6000	1.4654	0.0841
ΔCTGTG	0.8012	0.5897	0.2640
ΔTGTGT	0.5556	0.5997	0.0794
ΔGCTTT	0.9965	0.9455	0.0512
Δ7	0.6691	0.9373	0.4008
Δ9	1.0762	0.9132	0.1515
Δ10	1.3939	1.4794	0.0613
Δ14	1.2935	1.4585	0.1276
Δ22	1.4000	1.5224	0.0874
Δ26	0.7955	0.7367	0.0739
Δ28	1.5108	1.1688	0.2264
Δ37	1.0181	1.3010	0.2778
Δ77	3.2924	3.2328	0.0181
Δ81	2.4444	2.2531	0.0783
Δ1k	3.0067	3.0673	0.0202
	平均相対誤差		0.1287

表 7　*SD1* に対する本研究の提案方法の予測結果.

配列 ID	測定値	予測値	相対誤差
WT	48.9091	44.9701	0.0805
Δ6	44.5000	47.1828	0.0603
Δ7	38.2500	38.9828	0.0192
Δ10	34.6000	33.3723	0.0355
Δ14	45.3714	46.7512	0.0304
Δ18	42.3800	43.8980	0.0358
Δ40	37.8500	35.6573	0.0579
Δ52	28.6250	27.2270	0.0488
	平均相対誤差		0.0461

コンピューターの中で文字列を生成することは容易であるから, 変異配列を表す文字列を次々に生成し, それらを本研究において開発した予測方法に入力して, それらの下での表現型の予測値を計算させ, 所望の表現型を実現する変異配列を予測することができる. この予測結果の利用が, ゲノム編集から表現型の評価までの時間と労力の削減と, 延いては効率的な品種開発に繋がることを強く期待している.

5　まとめと今後の展開

5.1　現時点でのまとめ

第 1 節において述べたように, 期待の大きなゲノム編集であるが, 育種分野においてはそのポテンシャルを発揮することが出来ずにいる. 本研究では, ゲノム編集を用いた効率的な品種開発を目的として, 望む表現型を創出するための, 遺伝子の発現調節領域の編集箇所を予測する方法の開発に取り組んだ. その下での表現型を予測したい変異配列が, 基準とする配列 (例えば, 野生型配列) から Levenshtein 距離に関してどれだけ離れているかによって, その変異配列の表現型が, 基準配列の下での表現型からどれだけ変化するかを見積もる, というのがアルゴリズムの基本原理で, これに次の 4 つのヒューリスティクスを加えた方法を考えた.

Levenshtein 距離の編集費用として発現調節領域の塩基保存度を用いること, また, 指定された類縁種の範囲の下で計算された塩基保存度はあくまで初期値で, それらを予測誤差を減少させるように書き換えることが 1 番目のヒューリスティクスであった.

発現調節領域に長いゲノム編集を施した時の表現型の変化分が, それより短い編集を施した時の表現型の変化分と変わらないという現象が観察された. これは, ある長さのゲノム編集により, 転写因子の結合しやすさが底打ちすることが原因であると推測される. そこで, ゲノム編集効果の飽和という概念を導入して, ある条件が満たされている時には, 基本の予測量の代わりに, この現象に対応した予測量を用いることが 2 番目のヒューリスティクスであった.

本研究で構成した予測量は, ある変異配列の下での表現型の予測値を計算する際に, 保持する全ての配列とそれらの下での表現型の情報を使っても駄目であって, 保持する配列と表現型の情報の中から, 予測対象の配列の下での表現型の予測値の計算に有用な情報を選び出し, それらのみを用いて予測値を計算した方が良い, という考え方の下で構成されている. このため, 予測値の信頼できる上界と下界を設定することが出来る変異配列に対

しては, 本研究において構成された予測量は高い精度の予測値を返すが, そうでない変異配列に対しては, 例えばカーネル法のような既存の方法より誤差の大きい予測値を返すということが起こり得る. そこで, 予測値の上界と下界の信頼度を評価し, 本研究において開発された方法からの予測値と, カーネル関数と組み合わせた SVR など既存の方法からの予測値のどちらを用いるかを分岐される条件を学習させるようにしたことが, 3 番目のヒューリスティクスであった.

SD1 においては, 野生型配列と 7 個の変異配列, 及びそれらの下での表現型のデータが得られていた. よって, 学習機械に 1 点除外法で予測させる場合, 7 組の配列と表現型のデータのみを学習に利用することが出来た. *TAW1* の 27 組の配列と表現型のデータの収集は 3 年を要しており, *TAW1* に対して特別に多くのデータが利用可能であっただけであり, *SD1* の状況は, ゲノム編集育種の現場で学習機械の利用が検討される標準的な状況と言える. このような状況で, 他の遺伝子からのデータを保持している場合, それらを利用することが出来ないだろうかという考えが自然に出て来る. しかし, これまでの解析結果に基づくと, 新規の遺伝子に関する予測を行うのに, 他の遺伝子からのデータを用いても, 予測精度は変わらないか, 悪化することの方が多かった. しかし, 新規の遺伝子の発現調節領域のある種の変異配列に対しては, 他の遺伝子からの情報を利用して, その下での表現型の予測精度を向上させることが出来る. 他の遺伝子からの情報を利用して, 予測精度を向上させることが出来ると期待される条件を学習させるようにしたことが, 4 番目のヒューリスティクスであった.

5.2　今後の展開

最後に, 今後の展開について述べる. 第 1.2 節において述べたように, ゲノム編集は, 現時点では, 育種の主要な道具ではないが, 多くの可能性を秘めている. そこで, ゲノム編集を用いて品種開発を行ってみたいと考えている育種研究者を対象に, 本稿において述べてきた学習機械が動くウェブプラットフォーム DNA Sequence Designer を開発しており (図 2 参照), 完成し次第, 公開する予定となっている.

謝辞

本研究は, イノベーション創出強化研究推進事業「近傍保存配列 CNS のゲノム編集による作物遺伝子発現の精密調整技術の多様な作物への展開」(01005AB1) (2023 から 2025 年度) の支援の下で実施された. 表 3 に示された塩基保存度は, 共同研究者である東北大学大学院生命科学研究科の岩嵜航氏が計算したものである. 本研究に利用させて頂いたことに感謝の意を表する.

参考文献

[1] Kuroha, T. *et al*., Modification of *TAWAWA1*-mediated panicle architecture by genome editing of a downstream conserved non-coding sequence in rice, *Plant Biotechnology Journal*, 2025, https://doi.org/10.1111/pbi.70043

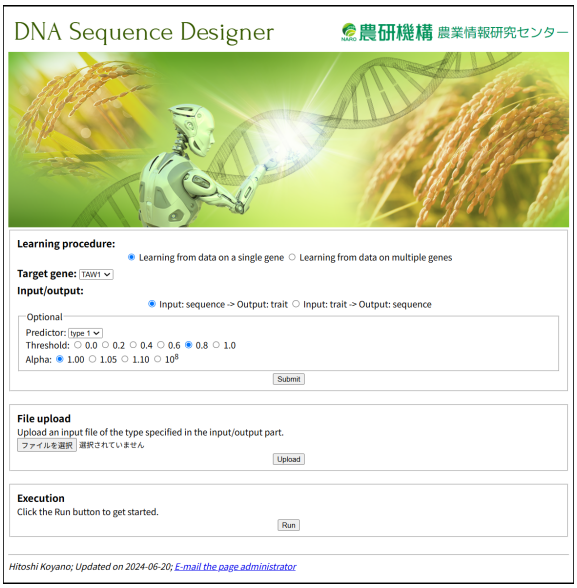

図 2　ゲノム編集育種支援ウェブプラットフォーム DNA Sequence Designer.