# 遺伝子発現予測のための入力変数選択に関する一考察

A STUDY OF INPUT VARIABLES SELECTION FOR GENE EXPRESSION PREDICTION

**宇佐見 仁英** [1)]，**森 直哉** [2)]，**泊 由紀子** [3)]，**渡邊 博之** [4)]
Hitohide Usami, Naoya Mori, Yukiko Tomari and Hiroyuki Watanabe

1) 工博 玉川大学 学術研究所 (〒194-8610 東京都町田市玉川学園 6-1-1 usami@lab.tamagawa.ac.jp)
2) 農博 玉川大学 学術研究所 (〒194-8610 東京都町田市玉川学園 6-1-1 mori-0810@lab.tamagawa.ac.jp)
3) 農学士 玉川大学 学術研究所 (〒194-8610 東京都町田市玉川学園 6-1-1 tomari-0809@lab.tamagawa.ac.jp)
4) 農博 玉川大学 農学部 (〒194-8610 東京都町田市玉川学園 6-1-1 watahiro@agr.tamagawa.ac.jp)

The closed type plant factory comes to a front of new method of agricultural production in our country. These plant factories have various advantages compared with traditional cultivation of the outdoor field. The greatest advantage is freely control of the growing environment without being influenced by the weather and it has a potential for maximizing the plant production. For that purpose, it is necessary to quantitatively grasp the relationship between the growth environments and the growth situations, furthermore to clarify the causal relationship of these factors. In this study, environmental response analysis at the genetic level was carried out using multiple regression model and neural network model as a discussion of basic research. In particular, we mainly discussed about narrowing the environmental factors in input variables selection problem.

***Key Words:*** *Evidence Based Plant Cultivation, Multivariate analysis, Input variables selection, LSTM*

## 1. はじめに

閉鎖型植物工場の最大の利点は，植物栽培のための環境を自在に制御することができることである．この利点を活かし生産性を上げるには，従来の天候をにらみながら人間の勘と経験に頼った圃場での作物栽培法から，工場生産としての均一で高品質な勘に頼らない植物生産法に転換しなければならない．我々は，勘に頼らない植物生産法として科学的根拠に基づく植物栽培法(EBPC：Evidence Based Plant Cultivation)を提案してきた．EBPCの最大のポイントは，植物栽培を植物の多変量時系列の最適化問題（図１）と捉え，生育等の目的変数を最大化する環境制御法を確立する事である．そのための科学的根拠となる知見として植物の生育と環境との関係を定量的な数値として捉える必要がある．

工場内の温度，湿度，光強度，液肥の養分状態等の物理的な環境は，それぞれのセンサーを活用することにより，デジタルな数字として容易に把握することができる．しかしながら，作物の生育状態，健康状態を観察しその状態に応じて温度等の環境を最適に制御することは難しい．特に，生育状態を非接触で数値として的確に捉えるのが難しく，自動化されていないのが現状である．我々は，生体重測定に歪センサーを用い，栽培環境データと１対１に対応できるオンライン計測可能なシナリオ栽培システムを開発[1)]し検討してきた．生育環境の制御項目は，気温，湿度，放射束密度，二酸化炭素濃度，風速，養液の温度・EC/pHの８項目である．これらの生育環境は，生育ステージ毎にシナリオに基づいて自在に設定でき，生育に影響が大きな環境要因を細かく制御できる．このシステムから得られる生体重・生育環境データを用いて多変量解析によりレタスの成長の環境因子を解析した．特に重回帰分析により生育ステージ毎のレタス成長に対する影響が大きな環境因子を推定した． 結果として，新鮮重量の増加に対して温度・湿度とEC値が強い関係性を指したが，どの因子がどの程度レタス成長に寄与しているかを示す明確な指標は得られなかった．

本論では，図１に示すようにオミックス空間における遺伝子発現と表現型としての生体重等が対応するというモデルで，環境因子により直接的な影響を受ける遺伝子の発現レベルでの環境応答解析を実施した．特に入力変数の影響を評価するために入力変数選択問題として重回帰分析とニューラルネットワークを用いて解析した．

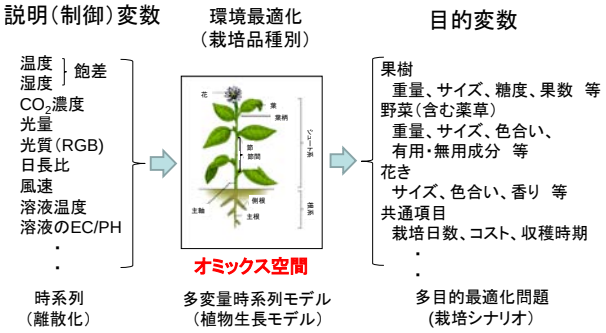


図１ 多変量時系列の最適化問題

## 2. 多変量解析

多変量解析は，複数の結果変数（説明変数＋目的変数）からなる多変量データを統計的に扱い変数間の関連性を明確にする手法で，回帰分析，主成分分析，ニューラルネットワーク，サポートベクターマシーンなど各種手法がある．それぞれ使用目的によって使い分けられているが，その中でも回帰モデルは，最も一般的に使われている基本的な方法といえる．また，昨今のAIブームのDeep learningの母体となったのがニューラルネットワークである．ニューラルネットワークは，需要予測などでの多くの実績を持っている．これら２つの手法での入力変数選択と遺伝子の発現予測を実施した．

### 2.1 重回帰分析

回帰分析とは説明変数と目的変数の間の関係を推定するための統計的手法のことをいう．説明変数が一つなら単回帰分析，説明変数が２つ以上なら重回帰分析と言う．今回，栽培の環境変数は複数あるので重回帰分析となる．n次元説明変数($x_1, \ldots, x_n$)と目的変数 y に関するデータが得られたとする．このときの線形重回帰モデルを式１に示す． a は回帰係数で，最小二乗法等で求められる．

$$y = \sum_{i=1}^{n} (a_i x_i) + b_0 \qquad \cdots (1)$$

重回帰では，回帰係数を算出する時にしばしば過学習による精度不良が発生する．過学習を防ぐため，誤差関数にペナルティーとしての正則化項を加えて最小化する手法が取られる．正則化項としてL1ノルムを取るものをLasso回帰，L2ノルムをとるものをRidge回帰と言う．L1ノルムは，パラメータの絶対値の総和を用いるものであり，L2ノルムはパラメータの二乗の総和である．L1ノルムのLasso回帰はLeast absolute shrinkage and selection operator の略で不要なパラメータ(次元・特徴量)を削ることができる．L2ノルムのRidge回帰は，過学習を抑えることができる．従って，本論では入力変数選択問題に適したLasso回帰を用いることにする．

### 2.2 ニューラルネットワーク

Deep learningでは，多階層のニューラルネットワークが使われるが，今回の解析には３階層のニューラルネットワーク（図２）を用いた．

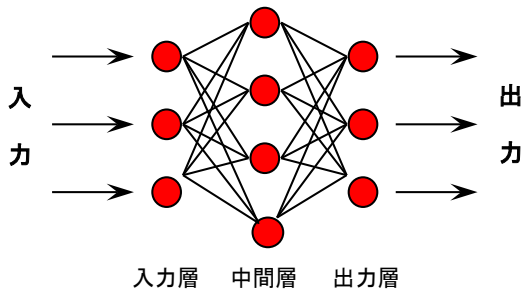


図２　３階層のニューラルネットワーク

Xを入力の重み付け和の結果，　O(h)を活性化関数にXを入力とした時の出力とする．Wは重み係数で，学習によって得られた結果を用いる．入出力関係を式２に示す．

$$x(h)j = \sum_i W(i)ij\, o(i)i, \quad o(h)j = \phi(x(h)j) \qquad \cdots (2)$$

解析ツールは富士通のNEUROSIM/Lを使用した．NEUROSIM/Lでは，目的関数を損失関数（loss function）とし，これを最小化する最適化問題として勾配降下法（gradient descent）を用いて解いている．これを各階層で計算し，層間では誤差逆伝搬法（backpropagation）によって誤差関数を指定された許容誤差範囲内で最小となるW値を決定していく．また，NEUROSIM/Lは競合学習の一種として成長側抑制学習をサポートしており，これは学習パターンの中から隠れている規則性を抽出するための学習法である．本論では，この機能を使用して入力変数選択問題を評価した．また，時系列での遺伝子の発現を予測するために時系列解析に優れたリカレント型ニューラルネットワークでの予測実験も実施した．リカレント型は出力を入力層に再入力させる方式のものである．実装としては，最新のLSTM（Long short-term memory）と呼ばれるリカレント型ニューラルネットワークを用いて遺伝子の発現予測を試みた．LSTMは株価や為替などの時系列の予測問題で多くの実績がある．LSTMのネットワーク概念図[2)]を図３に示す．

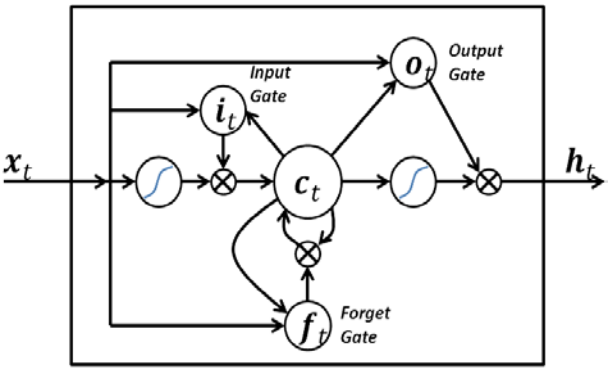


図３　ＬＳＴＭモデル[2)より]

## 3. 入力変数選択問題

昨今のDeep learningでは，ビックデータと称し得られた情報を全てつぎ込んで解析する方法がとられている．玉石混合の大量の情報からノイズを自動的に消し込んで重要となる規則性をAIが見つけてくれる．しかしながら実際の実験データ等では，データ量が限られる場合が多く，間違ったデータやあいまいなデータ等がノイズとなって予測精度を落とす事も多い．入力となる説明変数も観測されたものを全て投入すれば良いわけではなく，場合によっては矛盾した変数となり学習不能となる場合もある．今回，入力変数の影響度を算出する事によって入力変数の重要度を評価し，入力変数の絞り込みの可能性を検討した．　入力変数選択には多くの手法があるが，ここでは回帰分析とニューラルネットワークを用いた．

### 3.1 評価用データ

遺伝子解析をする前に入力変数選択のツールの評価を実施した．scikit-learn[3] に付属している標準的なデータセットのなかから糖尿病のデータを用いて評価実験を行った．このデータベースには，糖尿病患者 442 名の基礎項目（age, sex, body mass index, average blood pressure）と 6 つの血液検査項目（ic, ldl, hdl, tch, itg, glu），1 年後の進行状況 y が入っている．基礎項目と血液検査項目を入力，進行状況を出力としてモデル化した．このデータベースは多くの研究者が利用しており，回帰分析の評価用データベースとしても多くの論文が書かれており，検証用として優れたデータである．糖尿病データを正規化し，その各値の相関をヒートマップ図として図４に示す．また，基礎項目，血液検査項目と進行状況 y の相関係数を表１の「相関」として示す．

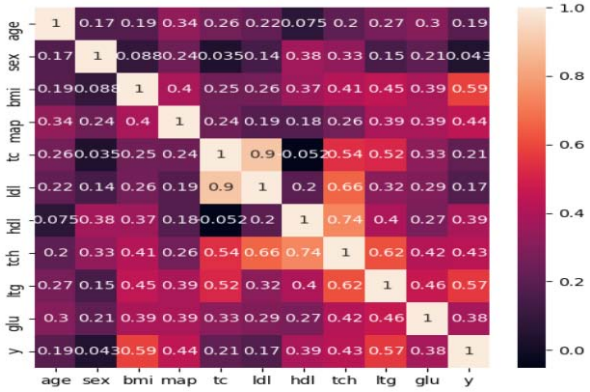


図４　糖尿病疾患に対する相関係数

基礎データとしてSGD (Stochastic Gradient Descent)での回帰係数を求めた．SGDは確率的勾配降下法であり，最小二乗誤差による方法と同様に良く使われている回帰係数の決定法である．結果を表１の「SGD」として示す．

### 3.2 Lasso（回帰分析）

Lasso回帰を用いて糖尿病の関連因子の絞り込みを試みた．正則化項を変化させた時の解パス（Solution Path）を図５に示す．横軸は正則化項で1.0近傍では0でない値に推定される回帰係数が多く，0.0に近づくほど0と推定される回帰係数が増えて行く．このように影響する因子が絞り込まれて行く様子がわかる．

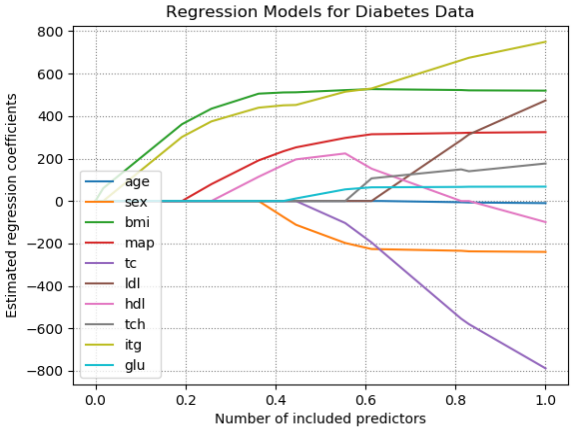


図５　糖尿病疾患に対するLasso回帰

また，alpha=1.0とした時のLassoの回帰係数を表１の「Lasso」として示す．

### 3.3 成長測抑制学習（ニューラルネットワーク）

成長側抑制学習は，学習時に近傍のニューロンに抑制を掛けることによって強い結合（Wが大）がより強く結合する．逆に弱い結合だと周りからの抑制作用により成長できなくなり，最終的には結合が無くなってしまう（W=0）ことになる．今回，抑制項としては0.001（入力層と中間層の間）と0.0005（中間層と出力層の間）を用いた．学習結果を図６に示す．図よりsex, bmi, map, tc, tch, itgが出力に強く影響していることが読み取れる．

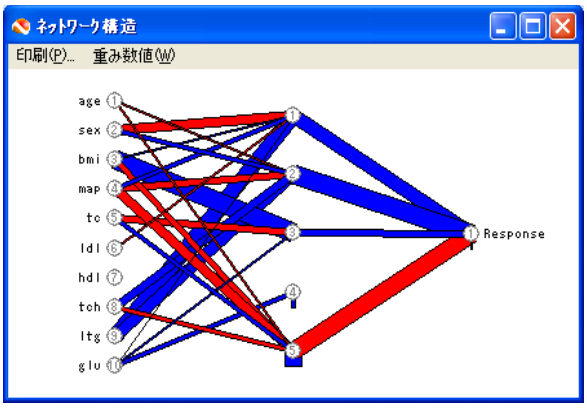


図６　糖尿病疾患に対する成長側抑制学習

また，NEUROSIM/Lでは影響度を計算することができる．影響度は，「どの入力項目がどの出力項目に対してどのように影響しているのか？」を評価するための機能である．各データポイントの微係数を学習データの積算値として算出している．成長側抑制学習後の影響度を表１の「側抑制」として示す．

### 3.4 各手法の比較

相関係数，SGD法による重回帰係数，alpha=1.0とした時のLasso回帰係数，成長側抑制学習を行なった後のNEUROSIM/L の影響度を表１に示す．

表１　入力変数選択のための因子の影響度評価

| 因子 | 相関 | SGD | Lasso | 側抑制 |
|---|---|---|---|---|
| age | 0.055 | −1.62 | −0.15 | 0.02 |
| sex | −0.11 | −11.5 | −9.72 | −0.01 |
| bmi | 0.35 | 26.9 | 26.8 | 0.09 |
| map | 0.08 | 14.4 | 12.8 | 0.11 |
| tc | −0.05 | −26.4 | −7.4 | −0.09 |
| ldl | −0.15 | 11.4 | 0 | −0.06 |
| hdl | −0.03 | 1.21 | 11 | 0.01 |
| tch | 0.27 | 7.34 | 0 | 0.09 |
| ltg | 0.52 | 32.9 | 26.7 | −0.04 |
| glu | 0.042 | 2.09 | 1.3 | 0.11 |

非常にバラツキが大きく，各因子の影響度を読み解くのは難しい状況ではあるが，全体を俯瞰するとsex, bmi, map, tc, itgなどの影響度が高そうであり，既往の研究と一致する傾向はみられる．今回の比較対象であるLassoと側抑制では，sex, bmi, map, tc, hdl, itg と bmi, map, tc, idl, itg gluの影響度が高く，一方，削除項目はそれぞれ「idlとtch」と「sexとhdl」であり，一致する結果は得られなかった．

今回，十分なチューニングをしていないとはいえ重回帰分析，ニューラルネットワークともに曖昧性を多く含んだ結果といえる．この手の問題を解くには，多くの手法で解析した総合的な判断が必要となる．

## 4. 遺伝子発現予測

永野等[4),5),6)]は，複雑に変化する野外環境に対する植物応答を類推するために野外圃場でのイネのトランスクリプトームデータと気象データを統計モデリングによって解析し，その予測結果を報告している．今回，永野等の研究を参考に，入力変数の絞り込みに限定して重回帰とニューラルネットワークによる分析を実施した．今回の実験では，網羅的な解析ではなく，可能性の検討ということで比較検討のために次の２種の遺伝子を選択した．

・Os01g0700100（糖輸送体関連遺伝子）

MtN3 and saliva related transmembrane protein family protein.　(bidirectional sugar transporter SWEET2b )

・Os02g0724000（時計関連遺伝子）

The Rice B-Box Zinc Finger Gene Family (CONSTANS-like protein, heading promotion under long-day condition.)

これらの遺伝子発現の実際のデータは，FIT-DB[7)]，FIT サンプル[8)]から抽出した．Os01g0700100とOs02g0724000の遺伝子発現の時系列データを図７に示す．横軸は１目盛が一回２時間のサンプリング時間で，圃場での実験（採取）周期は４８時間（２日）である．横軸の番号はサンプルに対応し，そのサンプルの採取日を表２に示す．少し複雑になるが対応させて読み取って欲しい．

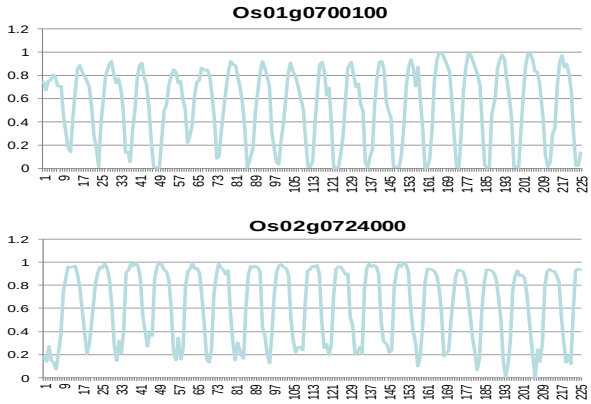


図７　遺伝子発現の時系列データ

表２　サンプリング番号と対応するサンプル日時

| No. | 開始日時 | | 終了日時 |
|---|---|---|---|
| 1 -> 25 | 2008/6/5 7:00 | -> | 2008/6/7 7:00 |
| 26 -> 50 | 2008/6/19 7:00 | -> | 2008/6/21 7:00 |
| 51 -> 75 | 2008/7/3 7:00 | -> | 2008/7/5 7:00 |
| 76 -> 100 | 2008/7/17 7:00 | -> | 2008/7/19 7:00 |
| 101 -> 125 | 2008/8/7 7:00 | -> | 2008/8/9 7:00 |
| 126 -> 150 | 2008/8/14 7:00 | -> | 2008/8/16 7:00 |
| 151 -> 175 | 2008/8/21 7:00 | -> | 2008/8/23 7:00 |
| 176 -> 200 | 2008/8/28 7:00 | -> | 2008/8/30 7:00 |
| 201 ->225 | 2008/9/11 7:00 | -> | 2008/9/13 7:00 |

図７に示すようにOs01g0700100とOs02g0724000の何れの遺伝子発現も概日性の周期性を示しており，気象環境に大きく影響されていることが伺える．また，両遺伝子では周期がずれており，この原因が気象環境の因子として抽出できればとの観点で比較対象遺伝子とした．

遺伝子発現に対応する気象データ（気温，湿度，気圧，風力，日照時間，露天温度，降水量）を気象庁のHP[9)]からダウンロードした．サンプル日時と対応させた気象データを図８に示す．図は，見やすくするために解析とは別に0~1の値をとる単純な正規化法を採用している．

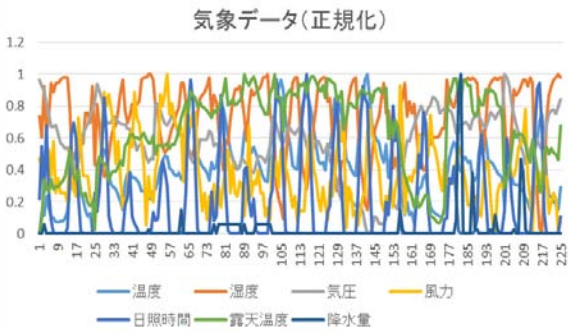


図８　遺伝子発現に対応する気象データ

### 4.1 Lasso（回帰分析）

Os01g0700100とOs02g0724000の遺伝子発現に対してLasso（回帰分析）を実施した．Os01g0700100の結果を図９に，Os02g0724000の結果を図１０に示す．

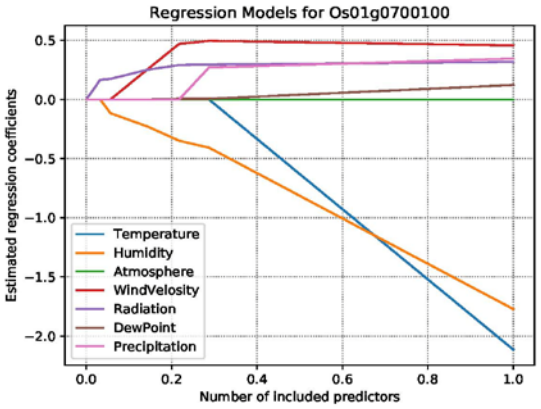


図９　Os01g0700100 に対する Lasso 回帰

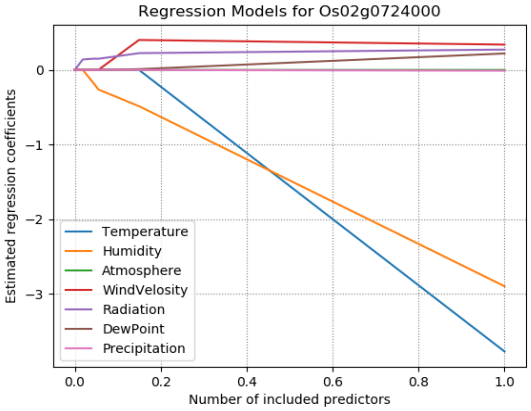


図１０　Os02g0724000 に対する Lasso 回帰

図９と図１０を比較すると多くの因子は同様な変化を示しているが，降水量（Precipitation）はかなり異なった動きをしていることがわかる．Os01g0700100 は糖輸送体関連遺伝子なので雨の直接的な影響を受けている可能性がある．一方，Os02g0724000 は時計関連遺伝子なので降水量の影響を全く受けていない可能性が読み取れる．

### 4.2 ニューラルネットワークモデル

Os01g0700100 と Os02g0724000 の遺伝子発現に対してニューラルネットワークでの解析を実施した．解析は，通常の完全結合型の学習法と周辺のニューロンに対して抑制を掛けながら学習する成長側抑制学習法の２通りである．特に成長側抑制学習法は入力因子の影響度を評価して不要な入力項目の絞り込みをするための有効な手段と考えている．Os01g0700100の通常学習の結果を図１１に，成長側抑制学習の結果を図１２に示す．

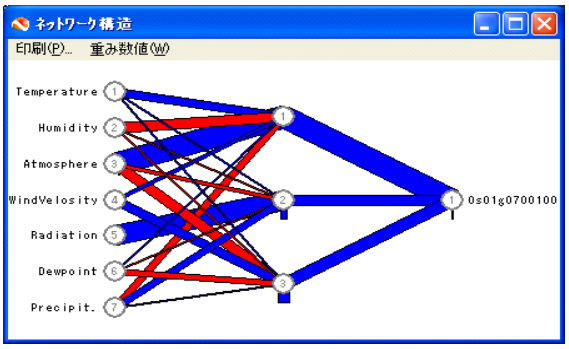


図１１　Os01g0700100に対する通常学習

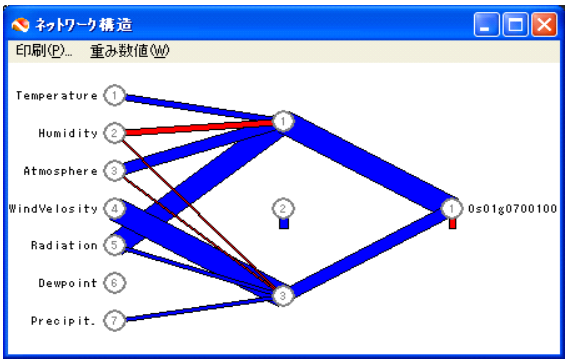


図１２　Os01g0700100 に対する成長側抑制学習

Os01g0700100 での通常学習と成長側抑制学習における影響度を表３に示す．図１２，表３とも成長側抑制学習において日照時間と風力の影響度が強い事が判る．

表３　Os01g0700100 に対する影響度

| | 通常学習 | 成長側抑制学習 |
|---|---|---|
| Temperature | 0.43 | 0.19 |
| Humidity | −0.43 | −0.22 |
| Atmosphere | −0.03 | 0.24 |
| WindVelosity | 0.43 | 0.43 |
| Radiation | 0.79 | 0.62 |
| Dewpoint | −0.26 | −0.0 |
| Precipitation | 0.14 | 0.04 |

次に，Os02g0724000 の通常学習の結果を図１３に，成長側抑制学習の結果を図１４に示す．

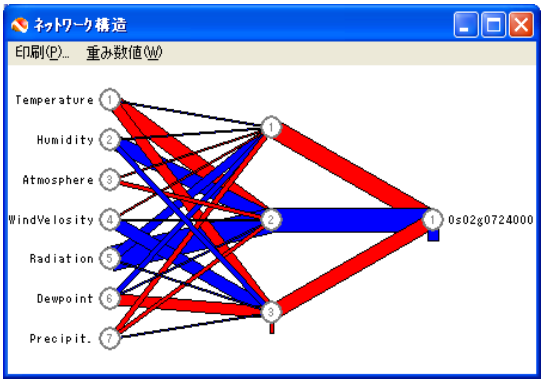


図１３　Os02g0724000 に対する通常学習

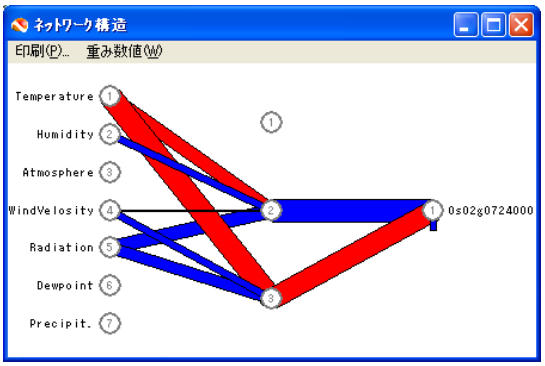


図１４　Os02g0724000 に対する成長側抑制学習

Os02g0724000 での通常学習と成長側抑制学習における影響度を表４に示す．特に成長側抑制学習（図１４）において日照時間と温度の影響度が強い事が判る．

表４　Os02g0724000 に対する影響度

| | 通常学習 | 成長側抑制学習 |
|---|---|---|
| Temperature | −0.58 | −0.32 |
| Humidity | 0.45 | 0.67 |
| Atmosphere | −0.18 | −0.01 |
| WindVelosity | −0.27 | −0.36 |
| Radiation | 0.8 | 0.86 |
| Dewpoint | 0.3 | 0.04 |
| Precipitation | −0.13 | −0.07 |

### 4.3 LSTM モデル

LSTM はリカレント型ニューラルネットワークで時系列の予測問題を得意としている．色々なモデルと実装が提案されているが，ここでは Deep learning では最も実績のある Google が開発しオープンソースで公開されている TensorFlow を使用した．TensorFlow の各パラメータは，look_back = 1，epochs = 3，batch_size = 1 である．また，入力変数は変数選択の結果を参考に温度，風力，日照時間に絞り込んで実験をした．

今回の LSTM では，予測精度の評価を主眼としているので，Os01g0700100 遺伝子の結果だけを記載する．Os02g0724000 の予測もほぼ同じような高い相関値を持つ結果であった．

全データの２/３を学習に使い，残りの１/３で予測実験を行った．結果を図１５に示す．

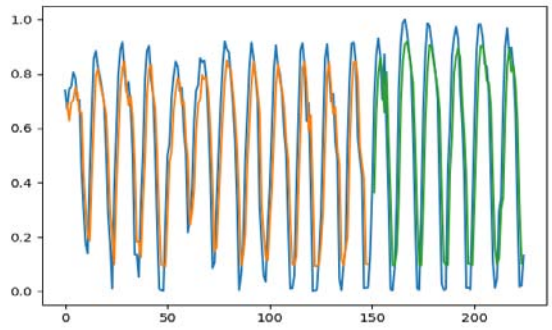


図１５　Os01g0700100 遺伝子の発現予測結果

赤線が学習用データ，緑線が評価テスト用データ，青線が LSTM での予測値である．若干の遅れは見られるものの良く一致する結果を示している．ただし，振幅のピークではかなりアンダーな予測をしており，変化点での予測精度が落ちるのはこの手のツールの特性かと思われる．この結果のテストデータを用いて観測値と予測値との散布図を作成した．結果を図１６に示す．

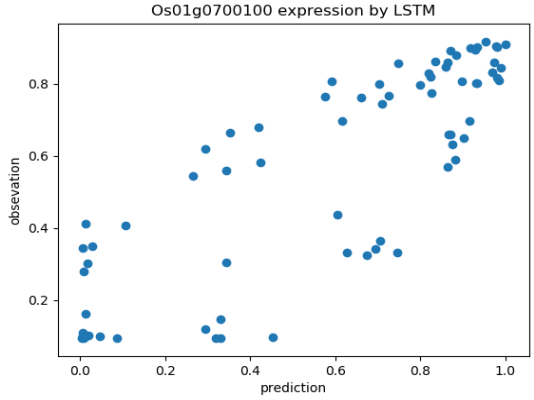


図１６　LSTM での予測精度

図１６において，観測値と予測値との相関係数は 0.84 であり，この手の問題としては高い予測精度を示す結果と言える．また，この散布図での観測値と予測値の対応点が対角線上に乗っておらず，対角線を中心線とする楕円状の形状をしている．この結果は，LSTM モデルの予測において overestimation と underestimation を繰り返すヒステリシス曲線を描いていることが伺える．この違いが若干の時間遅れとピークでの予測精度低下の要因かと推測している．これは，LSTM モデル内でのリカレント情報のメモリー機能に起因する可能性があり，LSTM モデル予測における興味深い結果でもある．

## 5. まとめ

生体重測定に歪センサーを用い，栽培環境データと１対１に対応できるオンライン計測可能なシナリオ栽培システムを開発し，多変量解析による因子分析を実施してきた．しかし，感覚と合うような結果を得ることが出来ていない．今回，もう少し直接的なレスポンスが期待できるレイヤーとして遺伝子レベルでの環境応答を観測することで，影響度因子の絞り込みの可能性を検討した．

具体的にはLasso回帰とニューラルネットワーク，特に成長側抑制学習という一種の競合学習による方法により入力因子を評価した．厳密な評価は難しいが，糖輸送体関連遺伝子は日照時間と風力，時計関連遺伝子は日照時間と温度であった．最も強い因子はいずれも日照時間であるが２番目の因子は風力と温度のような相違があった．風があると蒸散が進み光合成活性が盛んとなり糖輸送が活性化されると考えられる．また，モデル植物であるシロイヌナズナの表皮の時計遺伝子は温度感知をしていることが知られている．今回の供試植物であるライスの時計遺伝子も日照時間（日長）に反応するだけではなく温度感知もしている可能性がある．これらの結果からある程度の関連性の絞り込みの可能性を検証することができた．また，LSTMでの予測は重回帰，ニューラルネットワークに比べて良く一致する結果が得られている．これは，直前の予測した出力結果をフィードバックさせて再入力するモデルで，強い相関を持つデータを入力させるので当然の結果と言える．つまり，気象予報と同じで直前の予報の方が良く当たるのと同じ理屈である．

今後は，正規化法，パラメータチューニング，結果の検定などの課題を克服しながら，多くの実験データを積み重ねてさらなる精度向上に努める必要がある．

**謝辞**：FIT の使い方など，龍谷大学 永野惇先生，滋賀大学 大岩幸治先生から多くのご指導を頂きました．お礼申し上げます．なお，本研究は JST CREST 「野外環境と超並列高度制御環境の統合モデリングによる頑健性限界の解明と応用」の一環として実施しました．

## 参考文献

1) 宇佐見仁英，中井昭，堀口彰文，内堀崇，斎藤和興，″植物工場での遠隔シナリオ栽培に関する一考察″，pp. 76-77(A33), 日本生物環境工学会, 香川大学, 2013
2) Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, Jürgen Schmidhuber . "LSTM: A Search Space Odyssey". IEEE Transactions on Neural Networks and Learning Systems. 28 (10), 2015
3) scikit-learn（糖尿病疾患データベース） http://scikit-learn.org/stable//
4) Nagano AJ, Sato Y, Mihara M, Antonio BA, Motoyama R, Itoh H, Nagamura Y and Izawa T, Deciphering and Prediction of Transcriptome Dynamics under Fluctuating Field Conditions., Cell, 51(6):1358-69.2012
5) 永野惇，″フィールド・トランスクリプトミクスから 30 年後の生物学を考える″，光合成研究 23 (3) , pp129-135，2013
6) 永野惇，工藤洋，″屋外の環境における生物の環境応答の理解に向けて：トランスクリプトームデータと気象データの統合，領域融合レビュー″， 3, e009 ， DOI: 10.7875/leading.author.3.e009，2014
7) FIT-DB（遺伝子発現データ情報） http://fitdb.dna.affrc.go.jp/
8) FIT サンプル（遺伝子発現データ） https://cran.r－project.org/web/packages/FIT/
9) 気象庁 HP（気象データ） https://www.data.jma.go.jp/obd/stats/etrn/index.php